# Affymetrix cdf files

**Popis formátu cdf file:**
http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat545/Notes/AffxFileFormats/cdf.html

**Stiahnutie cdf ku konkrétnej platforme (hg-u133-plus)**
http://www.affymetrix.com/support/technical/byproduct.affx?product=hg-u133-plus

BiocInstaller::biocLite("hgu133a2cdf") – instalacia platformy uz existujucej
BiocInstaller::biocLite("makecdfenv") – nastroj na vytvorenie prostredia k akejkolvek platforme (nutno mat cdf file)

MU
IBA

# Krátky úvod metacentrum

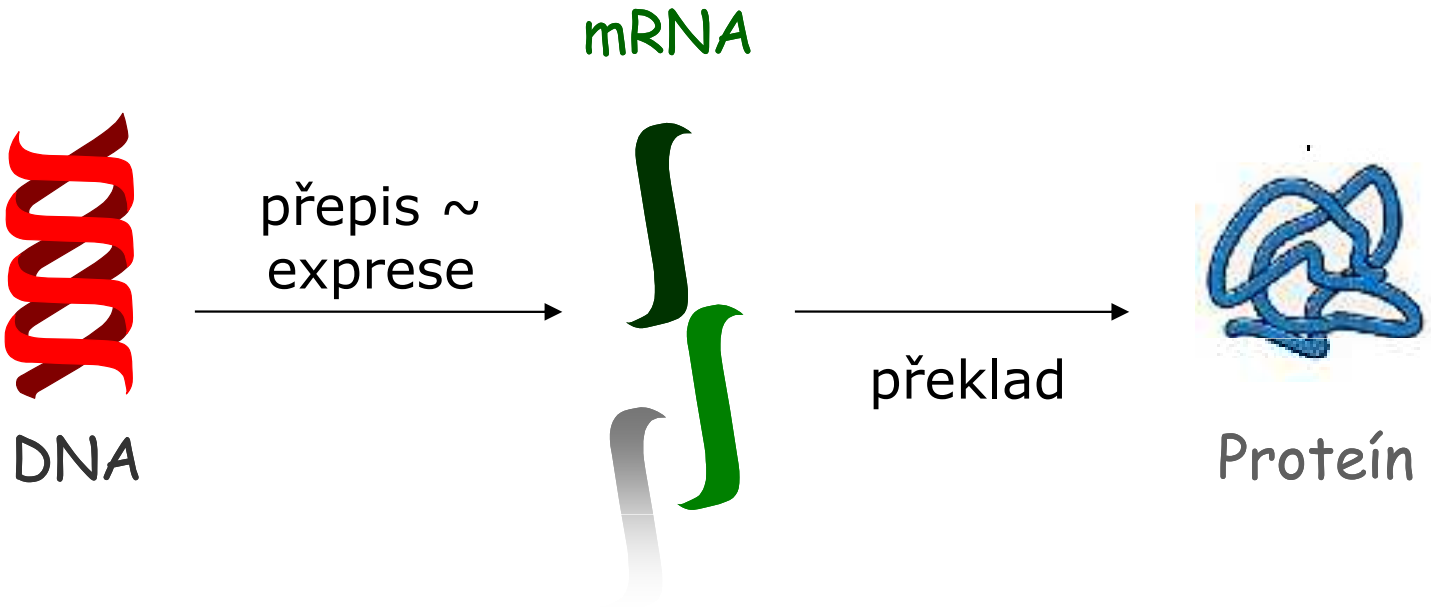
https://wiki.metacentrum.cz/wiki/How_to_compute/Accessing_machines/From_Window

MU IBA

# Kapitola III.

# Společné principy analýzy genomických a proteomických dat

# Genová exprese




- Gen je exprimovaný, pokud se *přepisuje* do mRNA
- Pokud se gen přepisuje, znamená to, že je aktivní
- Aktivitu genu můžeme měřit měřením množství příslušné mRNA v buňce

MU
IBA

# Tradiční schémata analýzy I.

- Každý experiment má odlišné cíle, v závislosti od typu dat a zájmů výzkumníků, ale existují tradiční schémata které se opakují:
- ***Učení s učitelem (supervised learning)***

**Známa struktura dat musí byt zevšeobecněná na nové data**

- Porovnávání skupin (class comparison)
  - hledáme rozdíly v expresi, v počtě kopií či struktůře genů/proteinů mezi už *definovanými skupinami*
- Předpovídání skupin (class prediction)
  - *na známých skupinách* se snažíme vytvořit klasifikátor, který by dokázal *zařadit nového pacienta* do jedné ze skupin

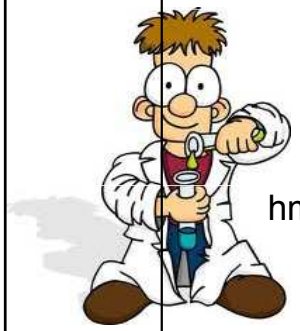
MU IBA

# Tradiční schémata analýzy II.

- ***Učení bez učitele (unsupervised learning)***
  - Objevování skupin (class discovery)
    - *Struktura v datach není známa, je potřebné ji vytvořit, objevit!*
    - Na základě informací o genech/proteinech *hledáme nové skupiny*
    - Příklady:
      - Existují nějaké soubory genů které se exprimují stejně ve všech podmínkách?
      - Onemocnění X je velmi heterogenní. Můžeme identifikovat specifičtější podtypy, které by mohli být cílem cílené 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 genových sad / genových sítí**

Interpretace

**Validace**

Publikace

# Kapitola V.1.
# Porovnávání skupin

# Příklady porovnávání skupin

- Pokud chceme zjistit
  - jaké geny jsou aktivní/neaktivní
  - jaký je rozdíl v přítomných proteinech mezi dvěma nebo více skupinami:
    - 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:

- Metody studující velikost efektu změny mezi skupinami
- Testování hypotéz
- Regresní strategie

# Základní metody pro porovnávání

Můžeme rozdělit do tří hlavních skupin:

- Metody studující velikost efektu změny mezi skupinami
- Testování hypotéz
- Regresní strategie

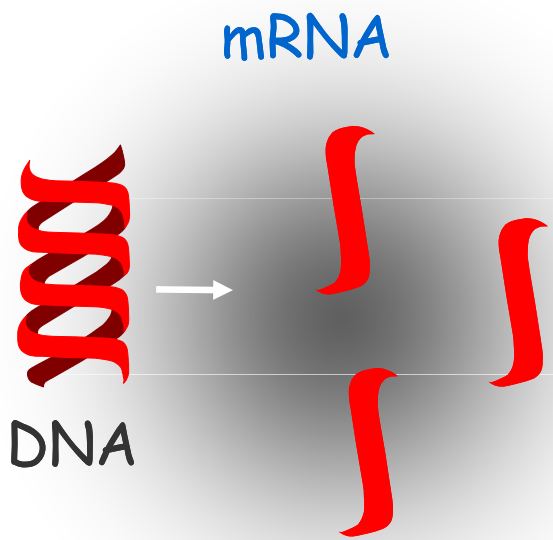
# Velikost efektu / změny II.

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

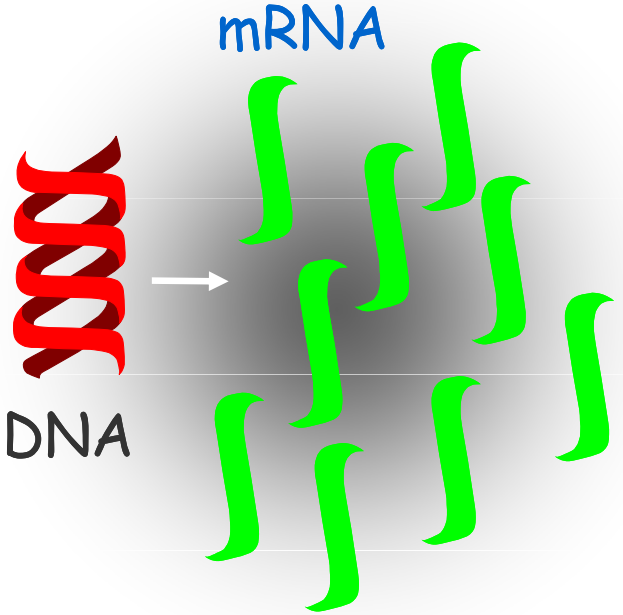
- Příklad: genová exprese, mean(X)/mean(Y), kde X a Y jsou genové exprese ve skupinách. Použitá hranice: 2!
- Výhody:
  - jednoduché

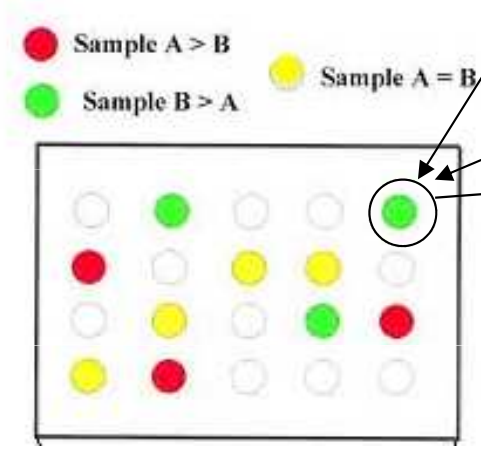
# Velikost efektu / změny III.

Skupina A. Zdravá tkáň




Skupina B. Nádor




9/3 = 3

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

# Velikost efektu / změny IV.

- **Nevýhody:**
  - I menší změny mohou být biologicky významné

(malý efekt genu/proteinu může být znásobený kooperací více genů v dráze)

  - Data jsou ovlyvněné technickou a biologickou variabilitou:
    - Co pokud máme 1.9?
    - Poměry mohou být vychýlené směrem k nule (například u nádorů s příměsí normálních buněk ve vzorce)
    - Neberou do úvahy variabilitu!

**Testování hypotéz**

# Základní metody pro porovnávání

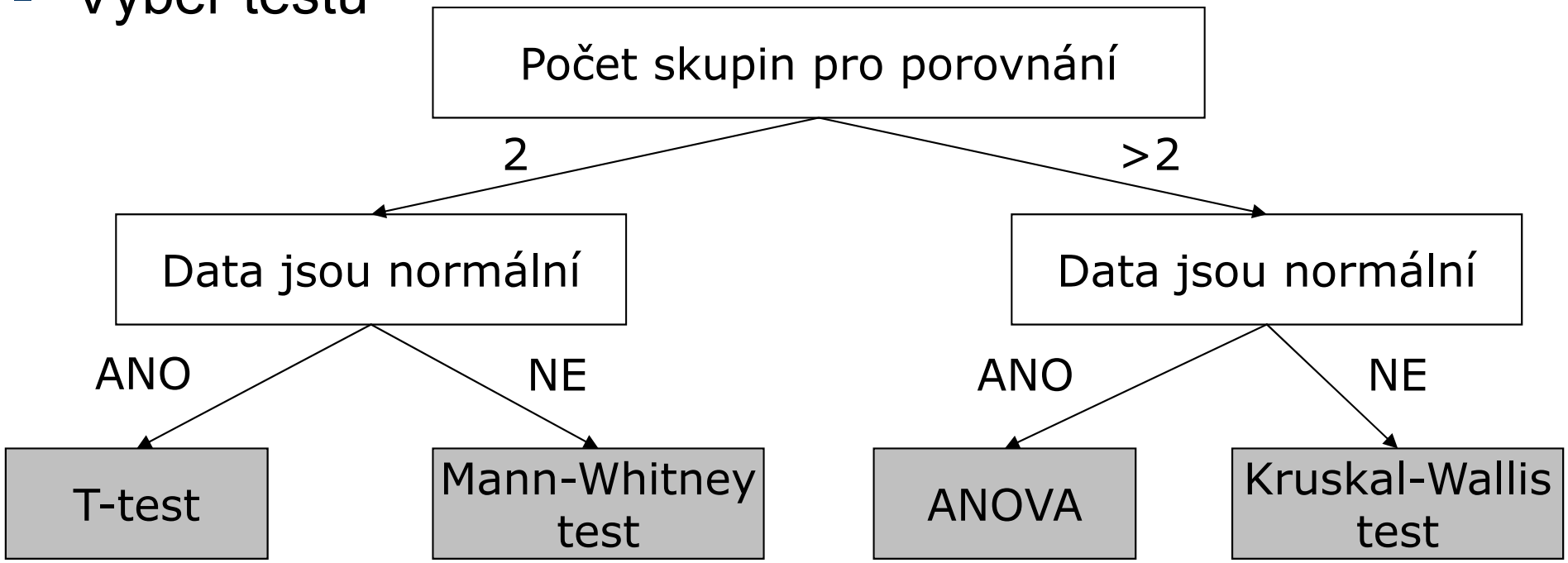
Můžeme rozdělit do tří hlavních skupin:

- Metody studující velikost efektu změny mezi skupinami
- Testování hypotéz
- Regresní strategie

# Testování hypotéz

- *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 aplikujeme statistický test, kterým získáme $T_g$ statistiku a příslušné *p-hodnoty*
- Výběr testu



# Testování hypotéz II.

Testuje se

- *Nulová hypotéza ($H_0$)*:

Gen / protein není odlišně exprimovaný mezi skupinami

versus

- *Alternativní hypotéza ($H_1$)*:

Gen je odlišně exprimovaný mezi skupinami

➔Na základě našich dat musíme rozhodnout, co je pravda

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

# T-statistika I.

- 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...)*

- T-statistika porovnává signál se šumem
  - Signál = rozdíl průměrů ve skupinách (u microarray dat se jedná o log(skupina 1)-log(skupina 2) = log(skupina1/skupina2))
  - Šum = směrodatná odchylka rozdílu (SD)

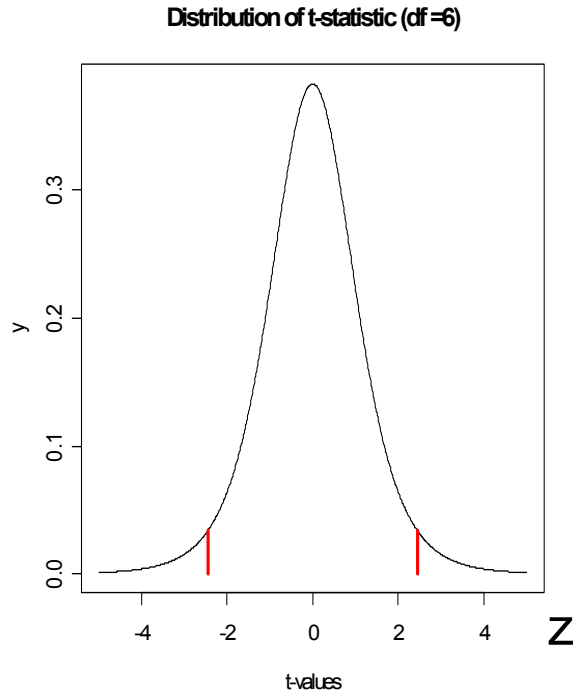
- T = log(skupina 1/skupina 2)/SD
- T hodnoty daleko od nuly indikují  snížení a nebo zvýšení exprese v jedné ze skupin

# T-statistika II.

- 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




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

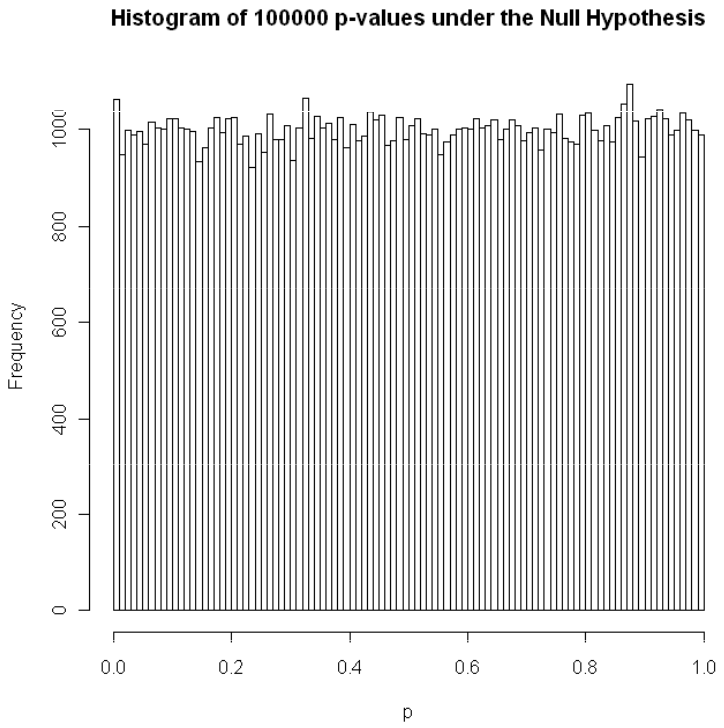
  $p_g = \Pr(T_g \leq T)$
- Dostatečně malá p-hodnota = významný rozdíl (silná evidence)

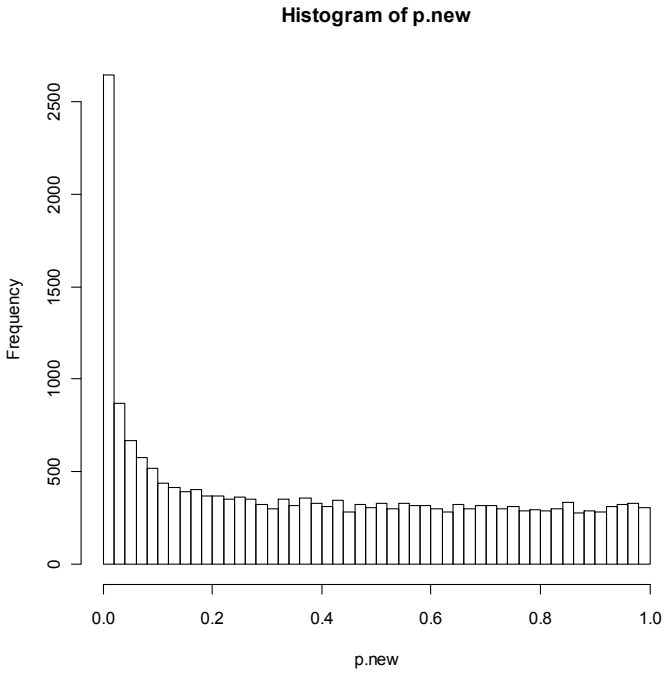
# Testování hypotéz III.

| | H0 nezamítneme | H0 zamítneme |
|---|---|---|
| H0 je pravdivá (gen není odlišně exprimovaný) | Pravdivá negativita (PN) | Falešná pozitivita (FP) Chyba I. druhu |
| H0 není pravdivá (gen je odlišně exprimovaný) | Falešná negativita (FN) Chyba II. druhu | Pravdivá pozitivita (PP) |

# Testování hypotéz IV.

- 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 rovnoměrně rozložené (vlevo). V případě, že je značná část genů odlišně exprimovaná, rozložení p-hodnot už není uniformní (vpravo).



# 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**

IBA

# Korekce problému mnohonásobného porovnávání

| | # nezamítnuté (NZ) | # zamítnuté (Z) |
|---|---|---|
| #bez rozdílu | Pravdivá negativita (PN) | Falešná pozitivita (FP)<br>Chyba I. druhu |
| # odlišné geny/proteiny | Falešná negativita (FN)<br>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

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

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

- Kontrolujeme FDR
  - Benjamini/Hochberg procedura

    FDR = 10% (ze 100 zamítnutých hypotéz očekáváme 10 falešně pozitivních)

# Který typ korekce použít?

- FWER pokud chceme aby VŠECHNY vybrané geny/proteiny byly opravdu významné. Na druhou stranu, nevybereme tak všechny významné geny!
- FDR pokud preferujeme vybrat většinu významných genů/proteinů, a nevadí nám nějaké falešně pozitivní
- q-hodnota je nejmenší FDR při které daný gen ještě zůstává na listu pozitivních

# Moderovaná T-statistika

- Problém ve statistickém testovaní mikročipových dat:

Příliš malé hodnoty exprese (blízké šumu) vykazují malou variabilitu => vysoké T-statistiky u biologicky nerelevantních genů!

Příklad:

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

$$\mu_{g1} = 2, \mu_{g2} = 2.5,$$
$$s_g = 0.02$$
$$\Rightarrow T_g = -25$$

- Aby se daly statistiky porovnat, je potřeba sjednotit variabilitu:
- Moderovaná T-statistika:

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

**Konstanta korigující variabilitu**

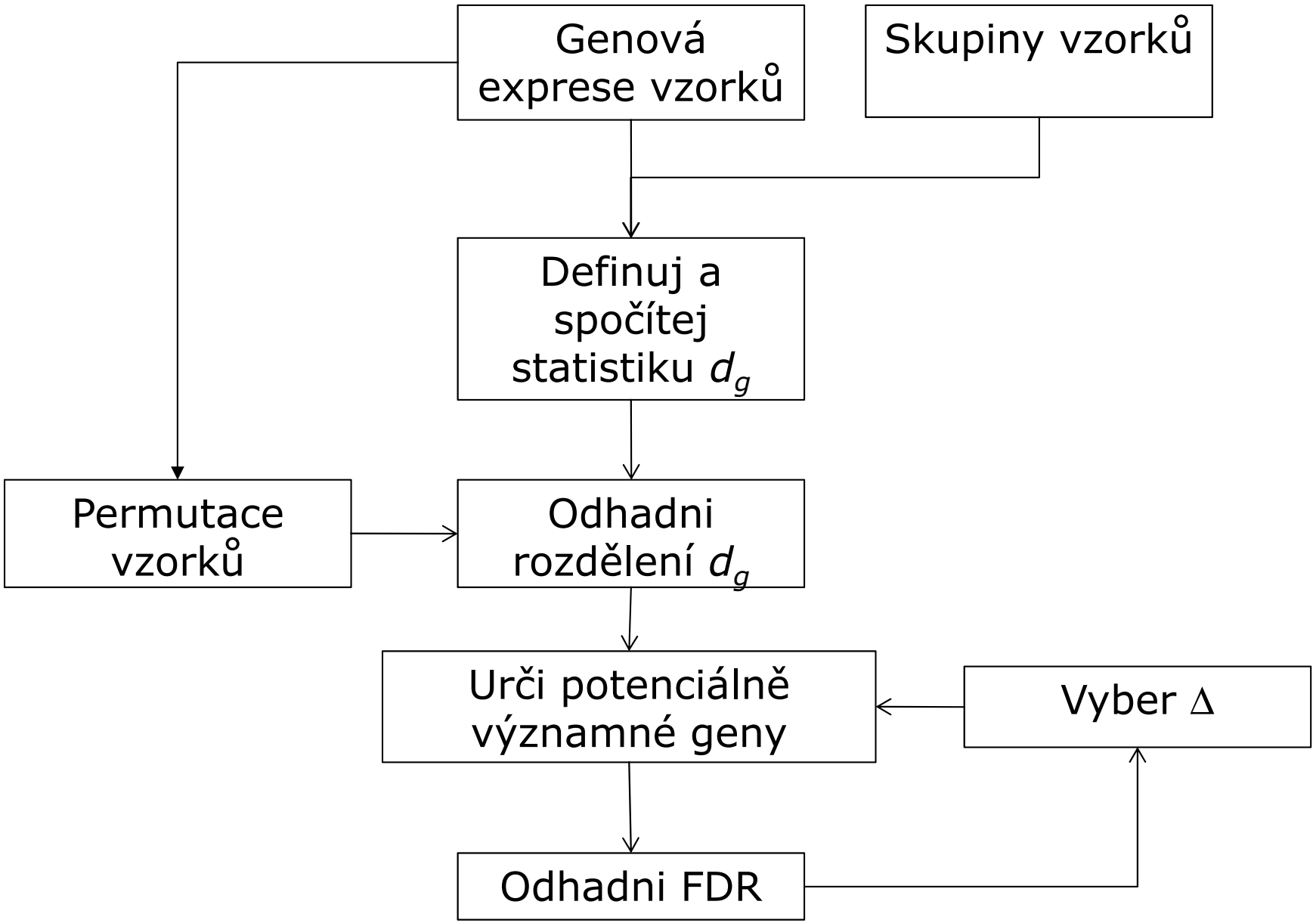
# Significance analysis of microarrays (SAM)

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

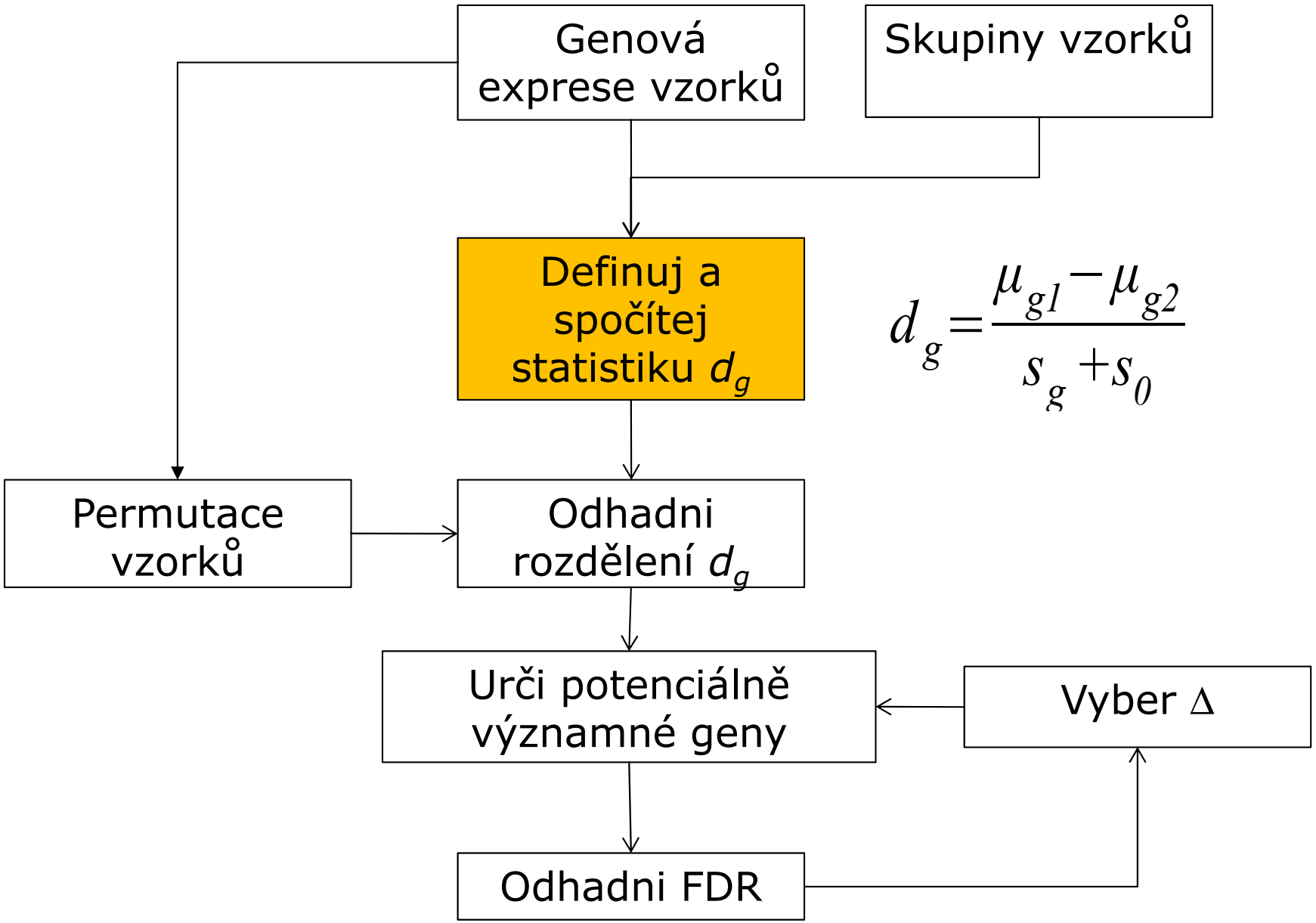
$$d_g = \frac{\mu_{g1} - \mu_{g2}}{s_g + s_0}$$

- 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)`

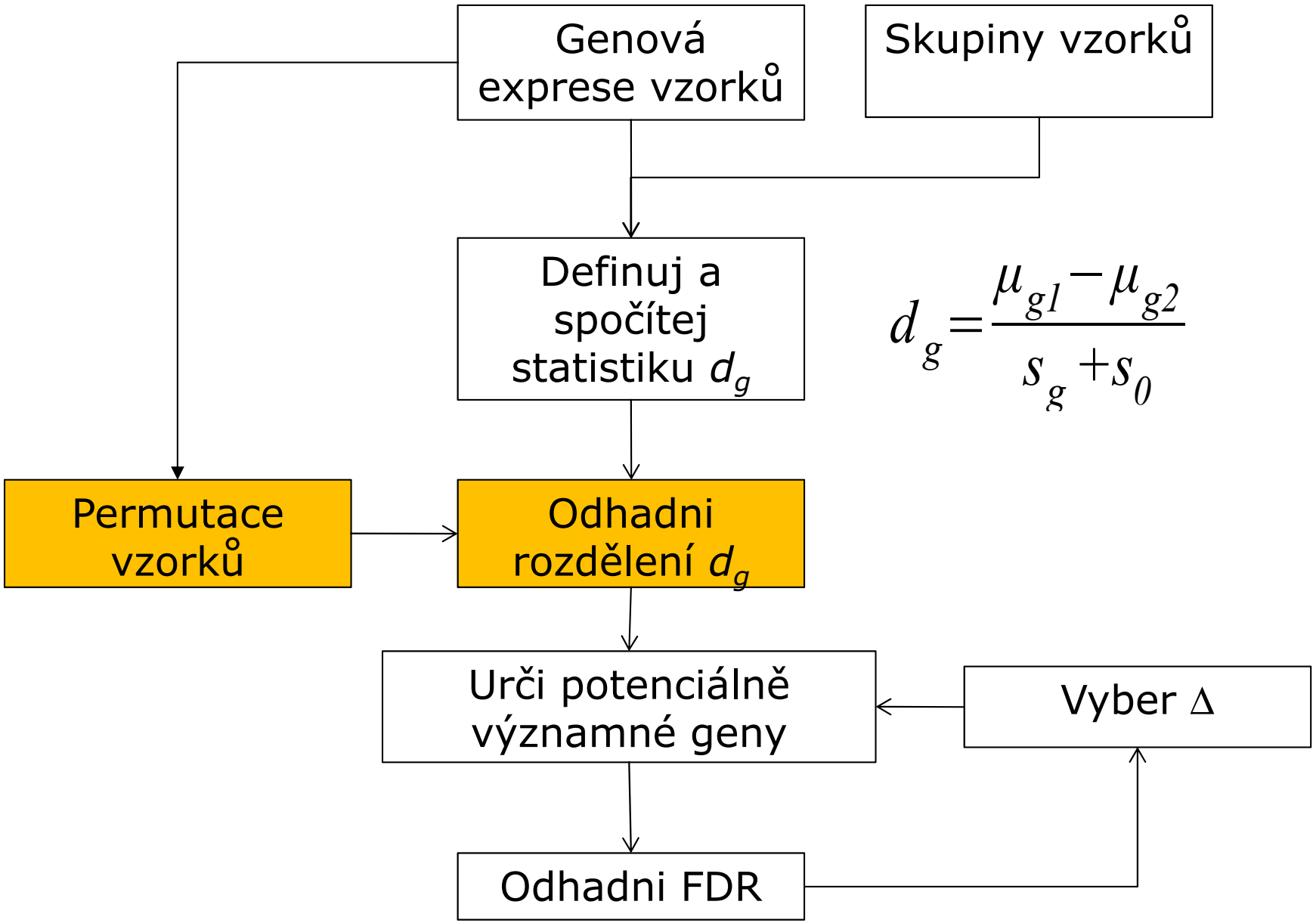
# SAM - algoritmus



# SAM - algoritmus



# SAM - algoritmus



# SAM - výpočet očekávaných hodnot

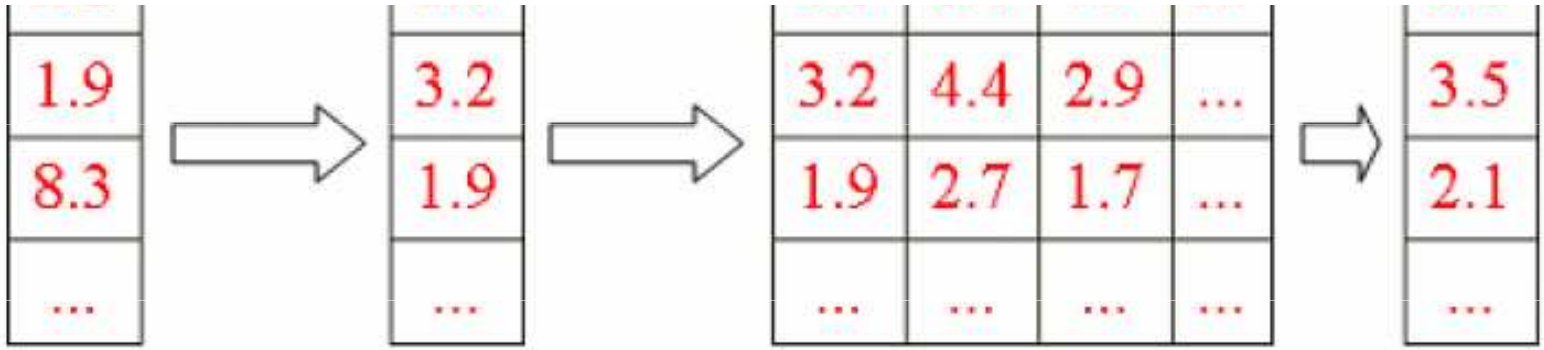
- Pro každou permutaci p spočítej $d_{gp}$

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

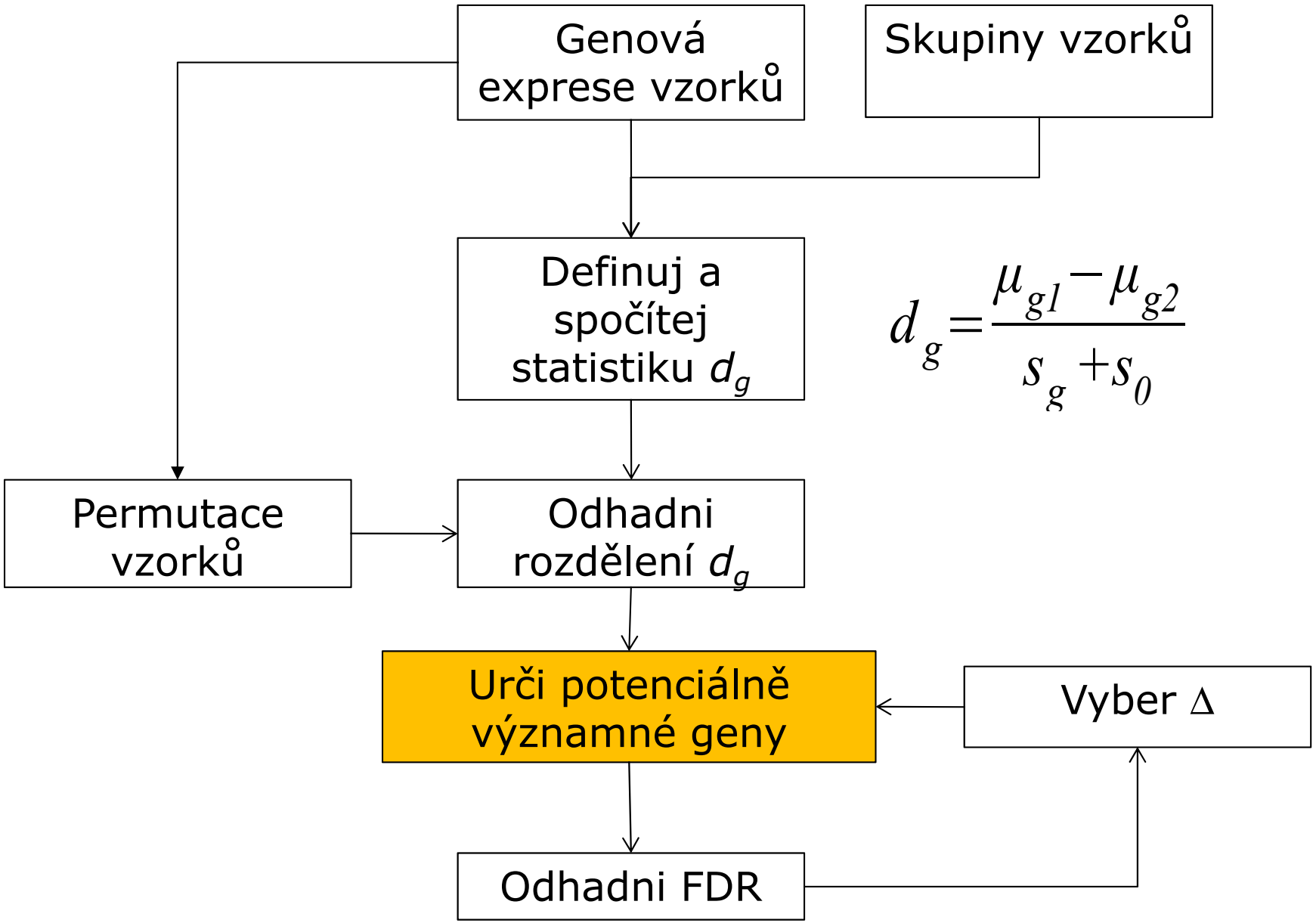
- Seřaď statistiky podle velikosti
- Definuj g-tou očekávanou hodnotu na základě N permutací

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

*$d_p$* *Seřaď* *Zkombinuj permutace* *$d_e$*

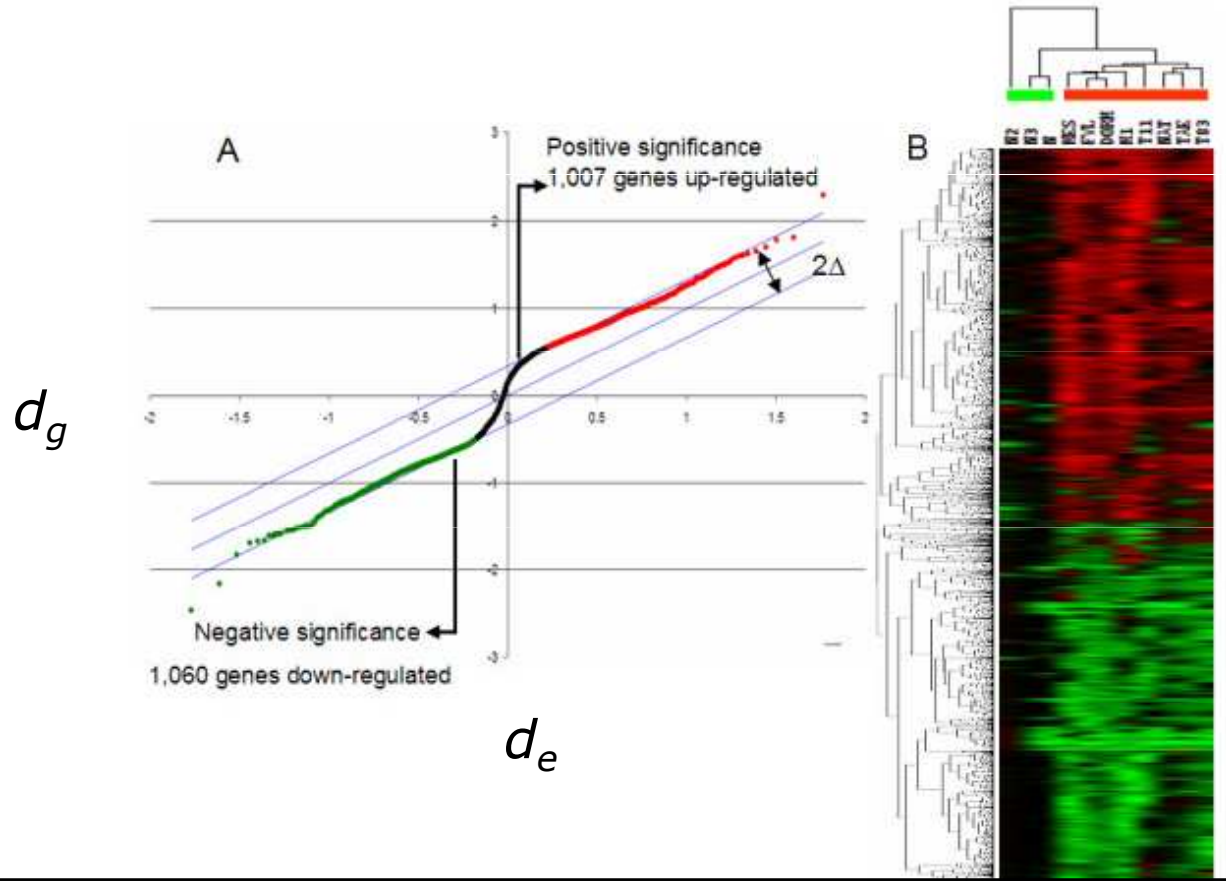


# SAM - algoritmus

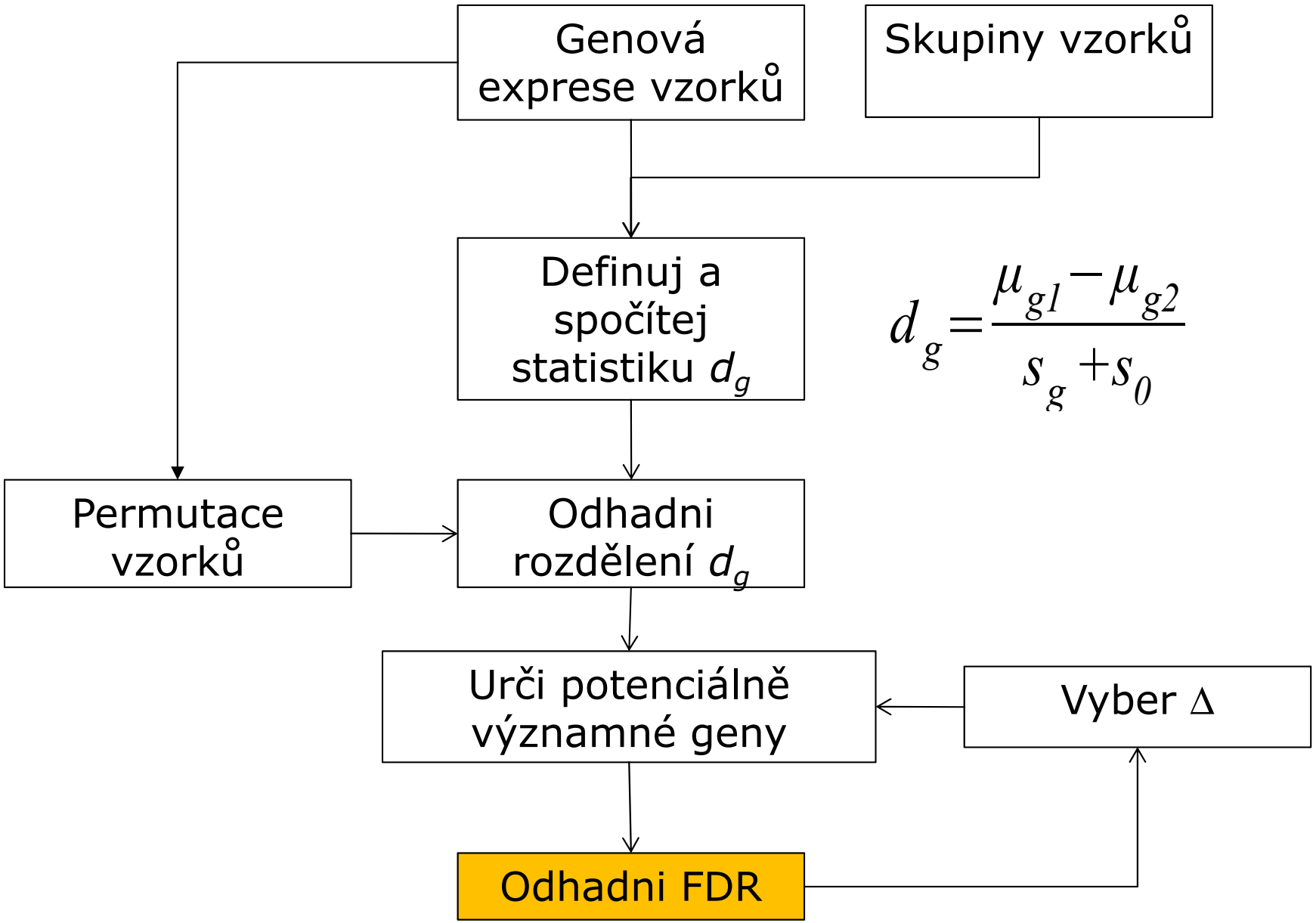


# SAM – určení významných genů I

- 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:

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

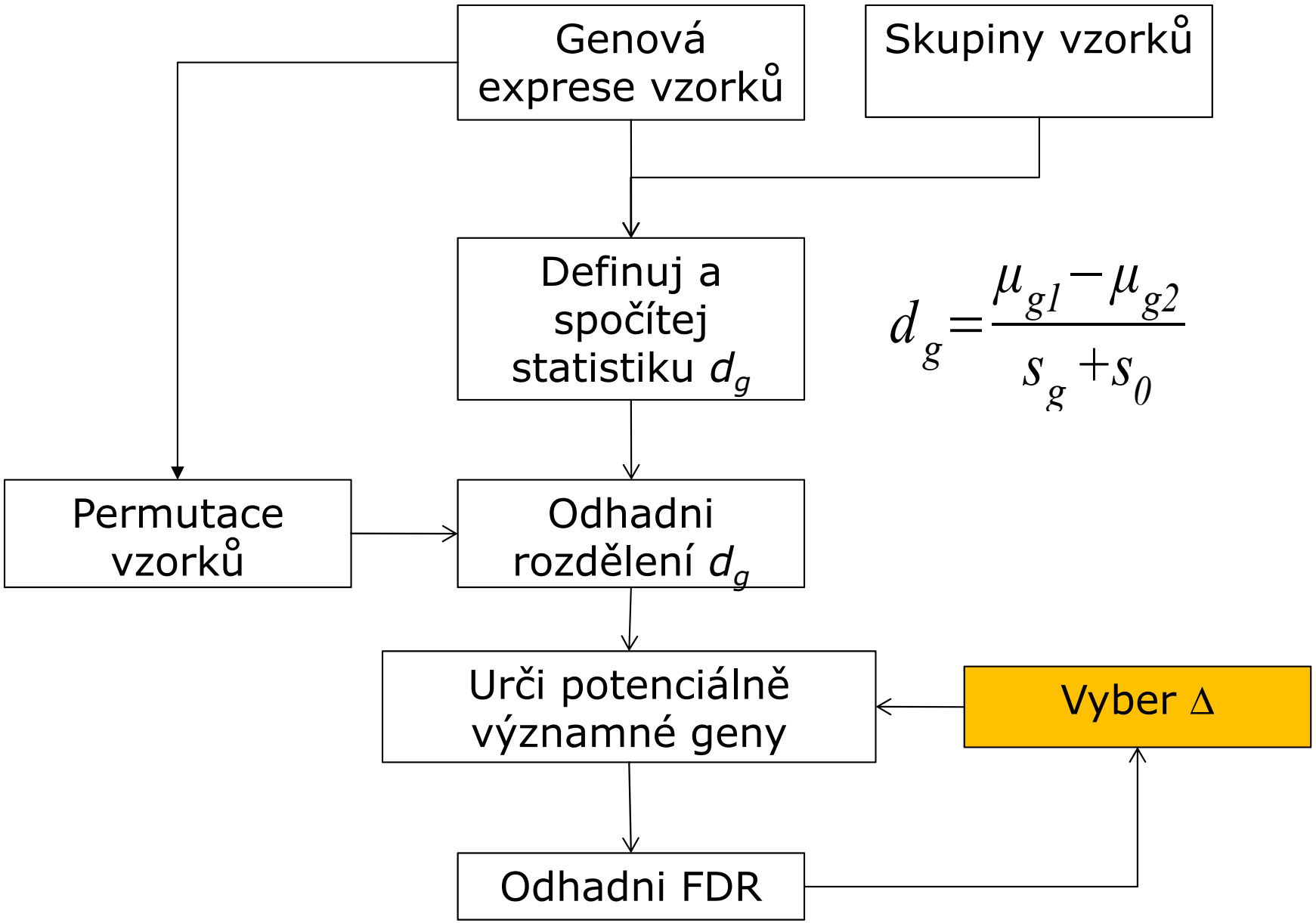
# SAM – výpočet FDR, příklad

| | $d_g$ |
|---|---|
| | 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

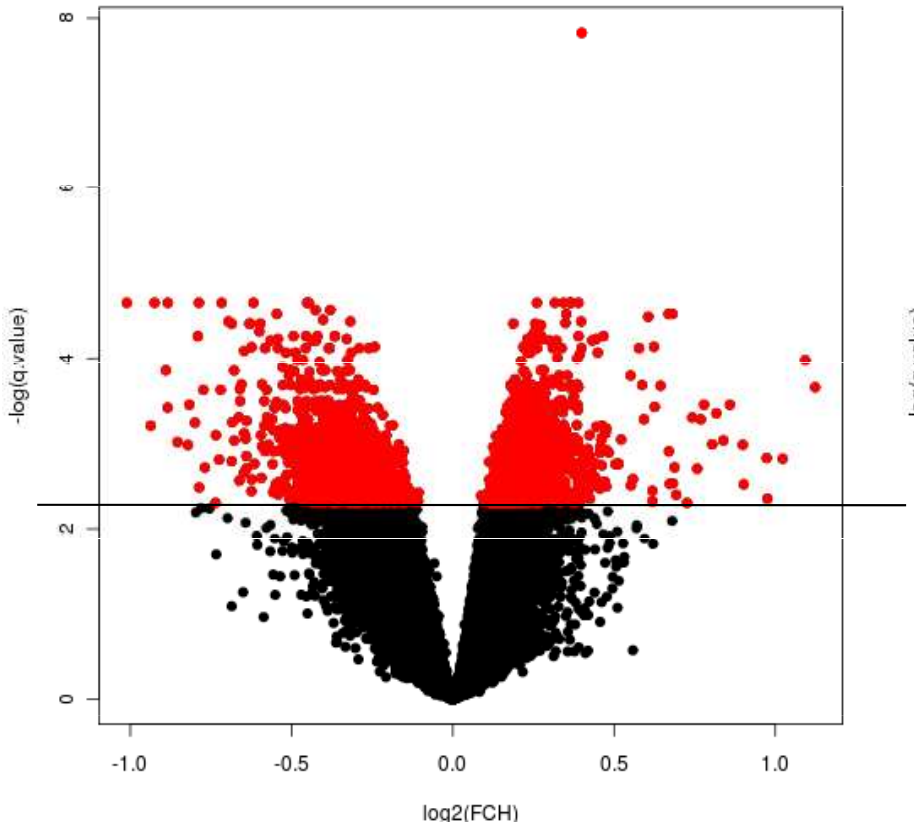
| Parametr | Počet falešně pozitivních (z permutací) | Počet označených za významné (v orig.) | FDR |
|---|---|---|---|
| SAM | | | |
| $\Delta = 0.4$ | 134.9 | 288 | 47% |
| $\Delta = 0.5$ | 78.1 | 192 | 41% |
| $\Delta = 0.6$ | 56.1 | 162 | 35% |
| $\Delta = 0.9$ | 19.1 | 80 | 24% |
| $\Delta = 1.2$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

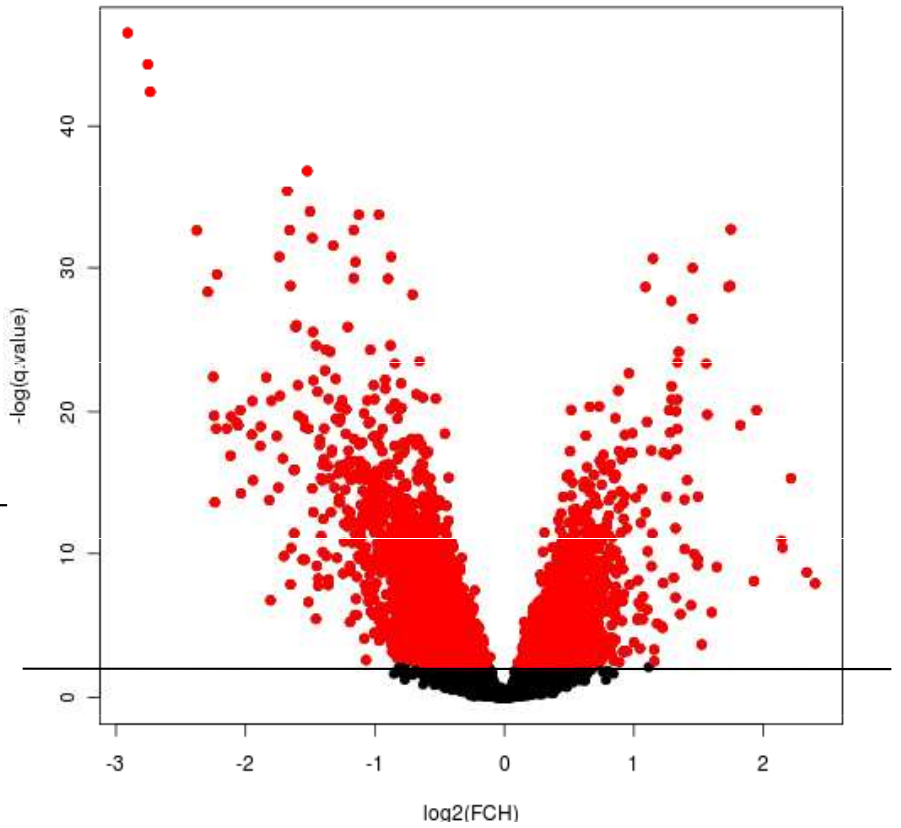
# Limma

- 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. http://www.bepress.com/sagmbvol3/iss1/art3
- **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
- `biocLite("limma")`
- `library(limma)`

# Volcano plots I.

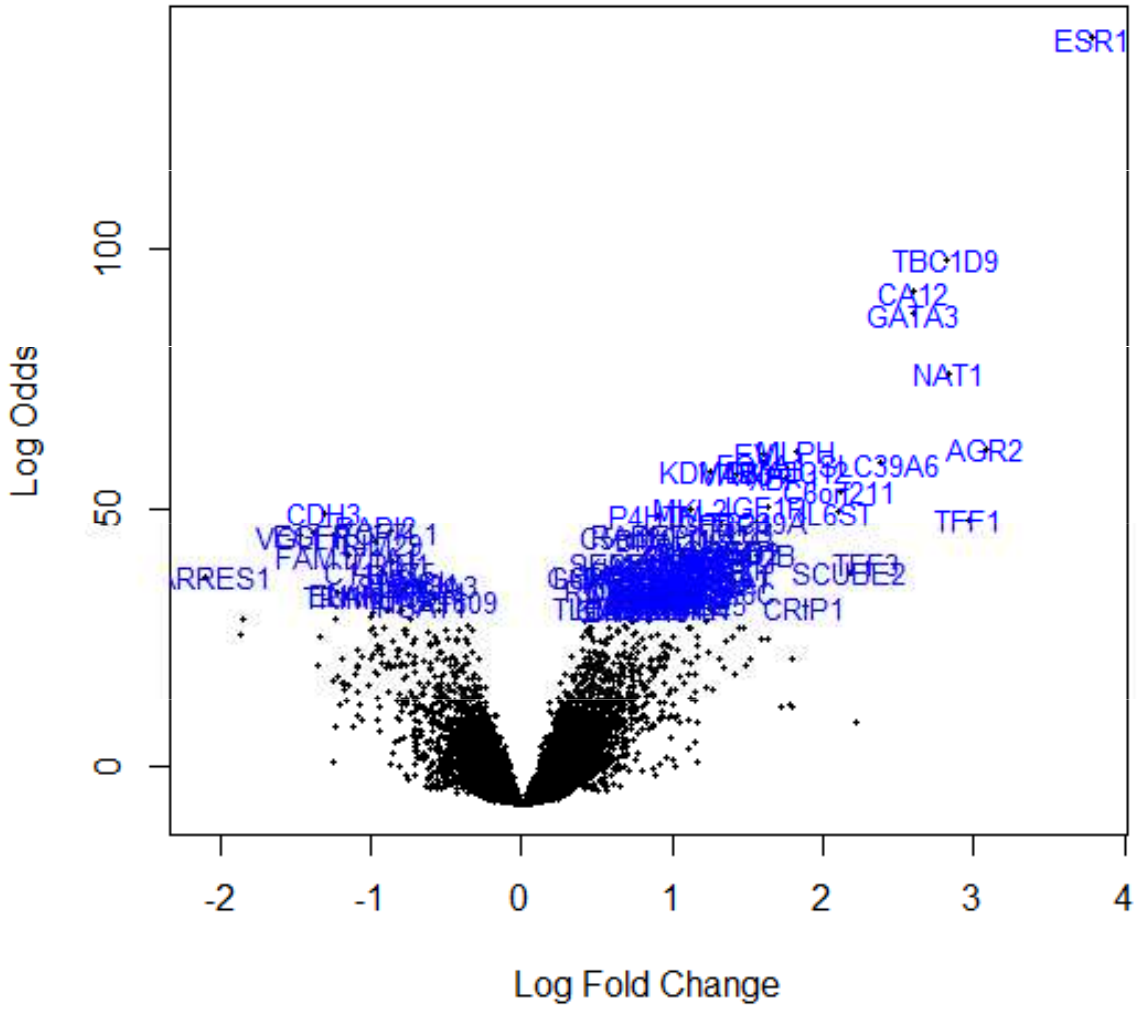
- log10(q-value) ~ -log10(0.1)=2.3



# Volcano plots II.

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



# Základní metody pro porovnávání

Můžeme rozdělit do tří hlavních skupin:

- Metody studující velikost 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*

- 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, Coxův model proporcionálních rizik*

- Chceme najít pravděpodobnost, že vzorek patří do určité skupiny na základě expresní hodnoty daného genu

  *Logistická regrese*

# Porovnání skupin