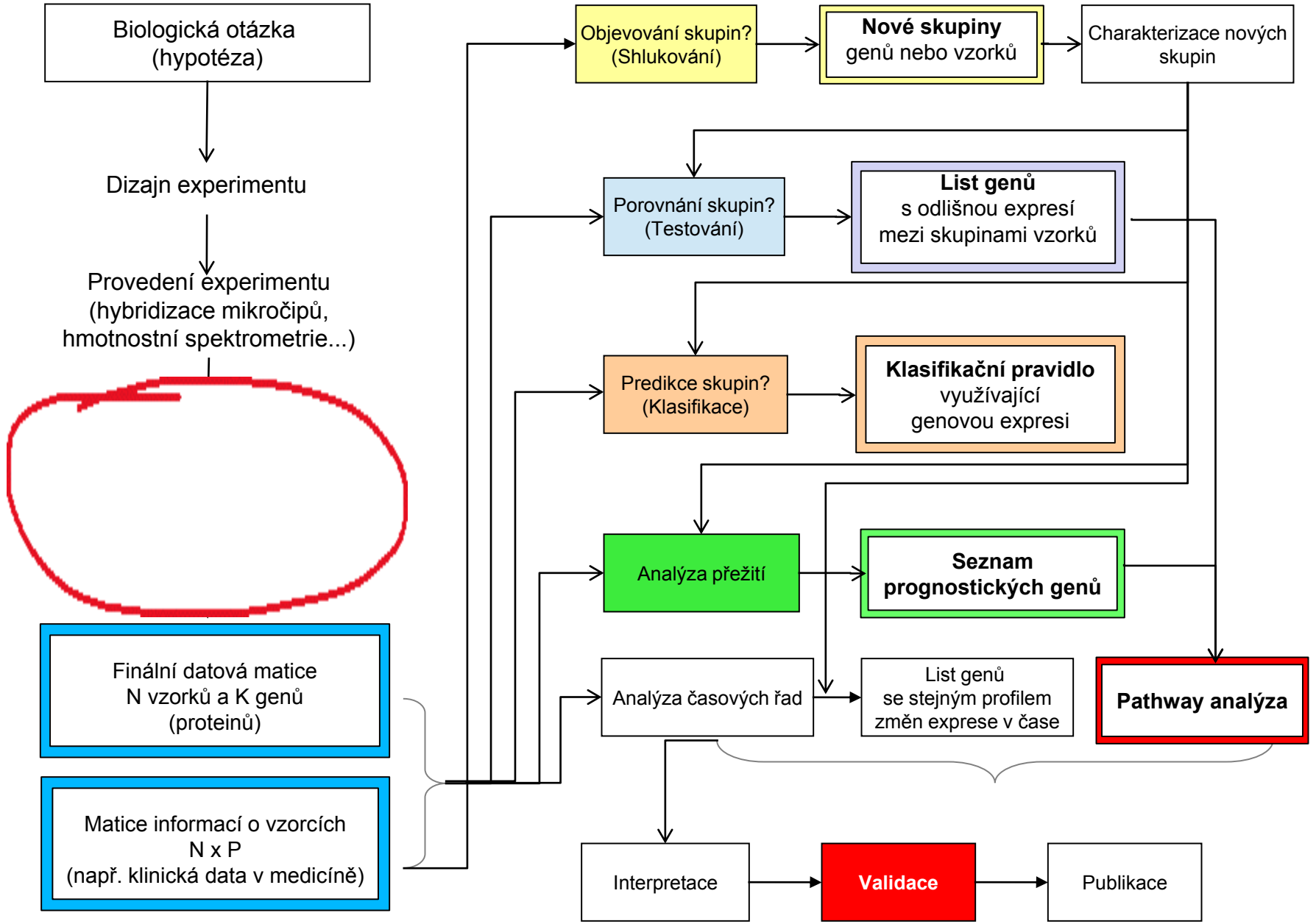
# Detekce biomarkerů z omics experimentů


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

- Podzim 2023

# Jak se hledá potenciální biomarker v omics datech



# Jak asi probíhá předzpracování dat?

Vzorky a jejich příprava

Mikročip

Sekvenace

Hmotnostní spektrometrie

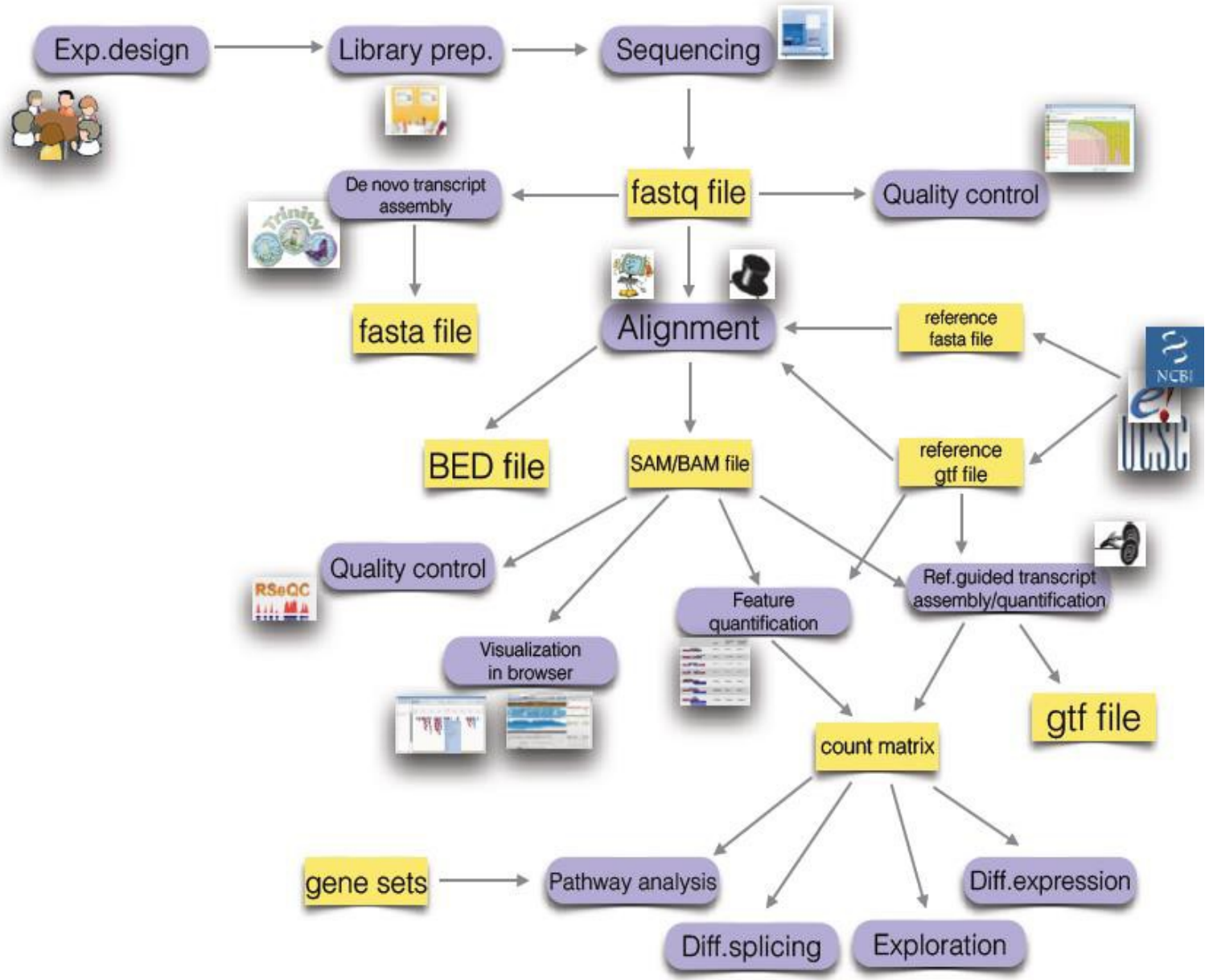
data

STATISTICKÁ ANALÝZA A DATA MINING

# Jak vypadají data z přístroje - cvičení

Interaktivní osnova (muni.cz)

# Příklad postupu bioinformatickéh o zpracování dat ze sekvenování nové generace



# Co obsahuje finální datový soubor s molekulárními daty?

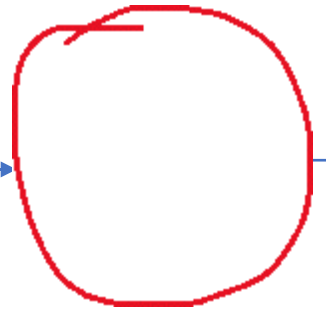
Vzorky a jejich příprava

Mikročip

Sekvenace

Hmotnostní spektrometrie

BIOINFORMATICKÉ ZPRACOVÁNÍ

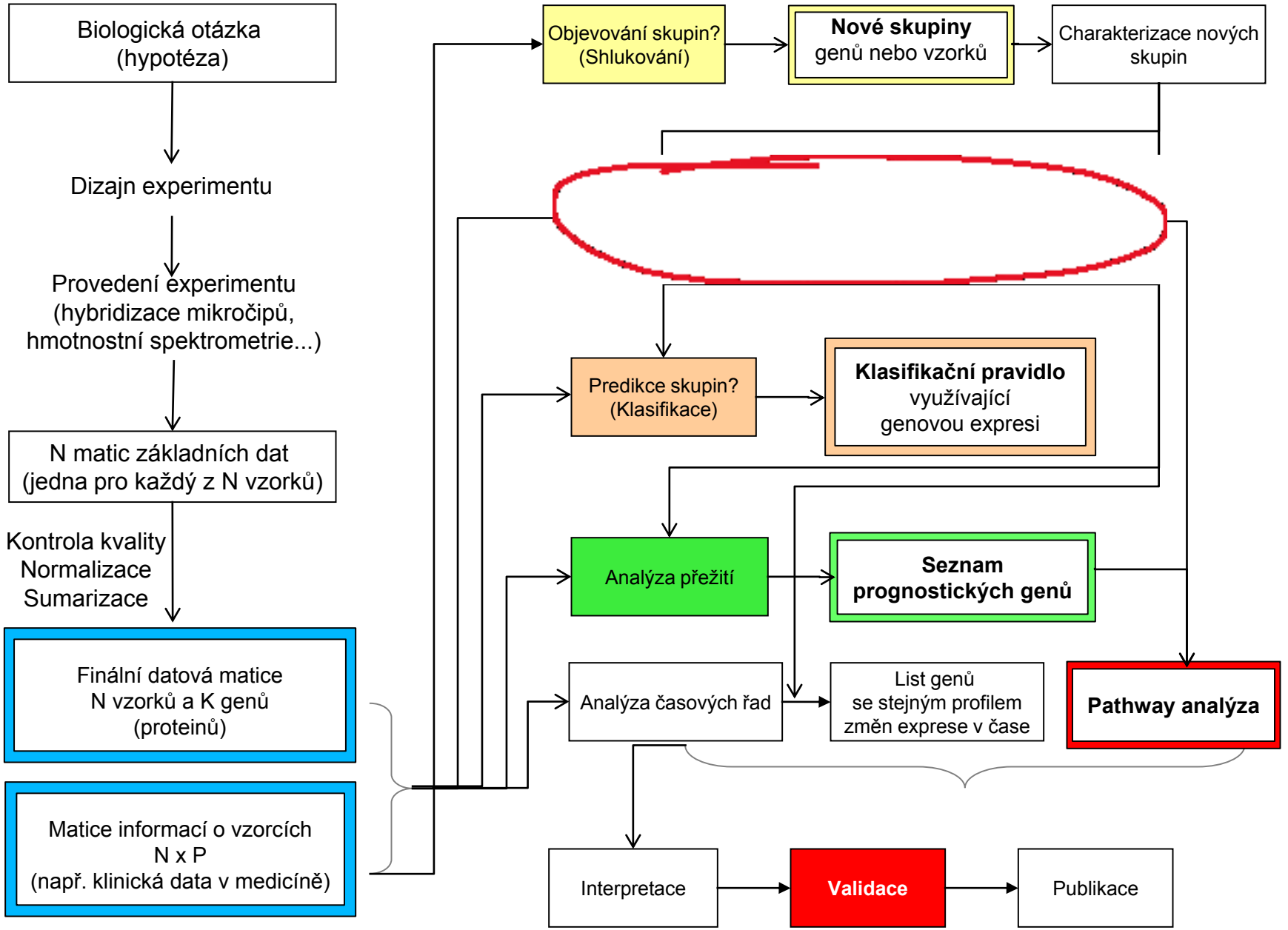


STATISTICKÁ ANALÝZA A DATA MINING

# Co obsahuje finální datový soubor s molekulárními daty? - cvičení

Interaktivní osnova (muni.cz)

# Jak se hledá potenciální biomarker v omics datech



# Hledání rozdílů mezi skupinami

Odpovídáme na otázku:

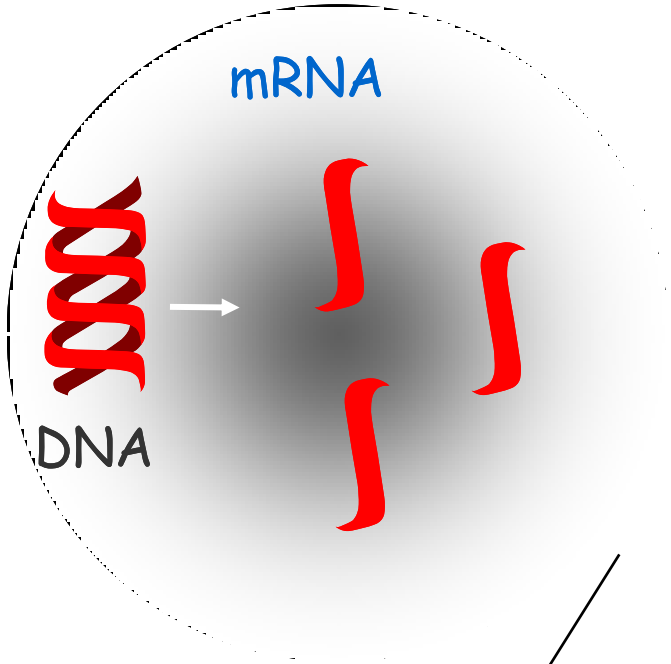
jaký je rozdíl v přítomných genech/metabolitech/proteinech mezi dvěma nebo více skupinami

# Příklady porovnáván í skupin

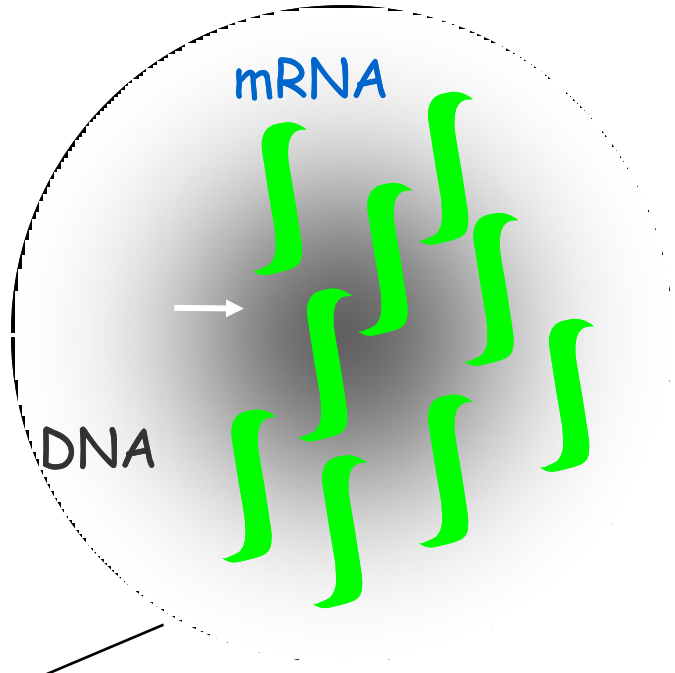
- nemocní vs. zdraví pacienti
- pacienti před vs. po terapii
- pacienti v čase diagnózy a v čase relapsu
- bakterie v aerobním vs. anaerobním prostředí
- druh 1 vs. druh 2
- porovnáváme podtypy onemocnění

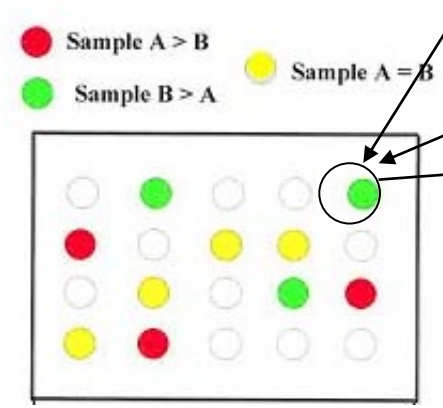
# Příklad na mikročipu

**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

**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**

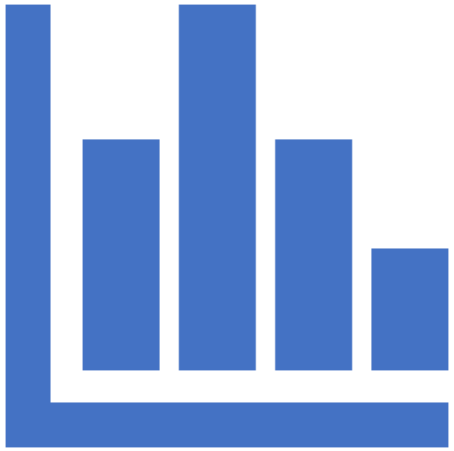
# Metoda dělící hranice velikosti efektu / změny

## 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!**

# Testování hypotéz

# 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...)*

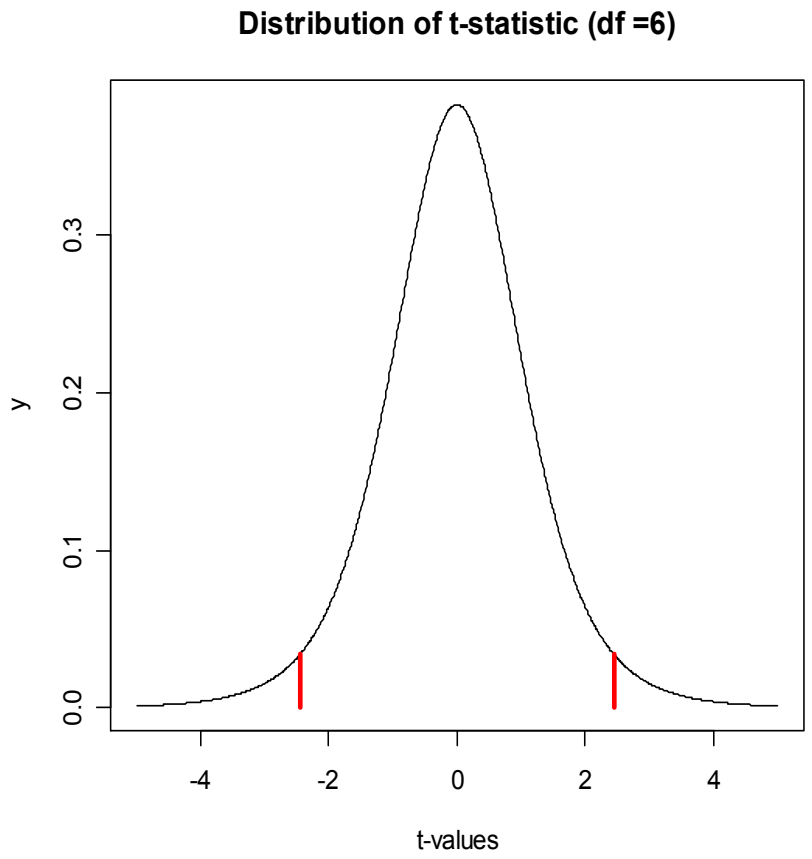
- Statistiky jsou definovány různě a mají různé předpoklady.
- Například T-statistika porovnává signál se šumem a předpokládá normalitu dat.

# 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*

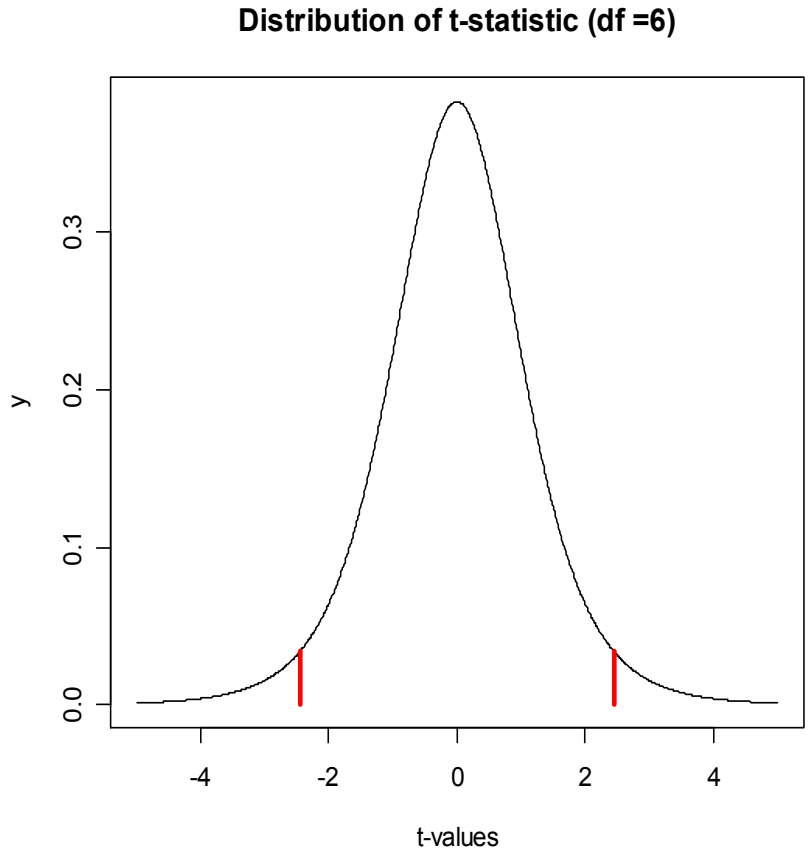
# T-test a T-statistika




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

Variabilita (vyjádřená jako směrodatná odchylka)

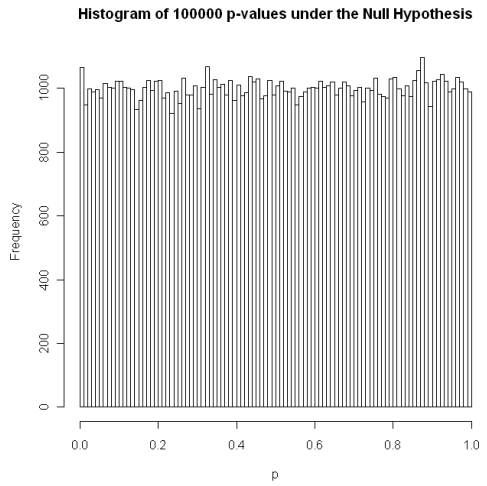
# T-test a T-statistika

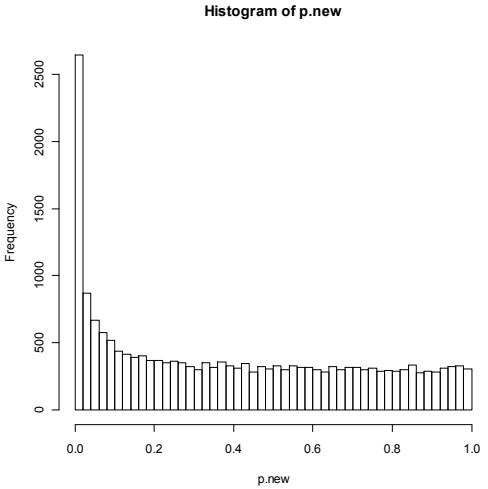



- Pokud data mají **normální rozložení a neexistuje rozdíl mezi skupinami**, tak T-statistiky pocházejí z **T-rozložení**.
- **p-hodnota** = pravděpodobnost že dostaneme danou hodnotu T-statistiky nebo hodnotu větší, v případě, že neexistuje rozdíl mezi skupinami

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

- 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

- **Chyby 1. druhu:**

- **Family–wise error rate (FWER)**: Pravděpodobnost alespoň jedné chyby prvního druhu (falešné pozitivity): FWER = Pr(FP > 0)

- **False discovery rate (FDR)**(Benjamini & Hochberg,1995):
- Očakávaný podíl falešně pozitivních výsledků mezi zamítnutými hypotézami
- FDR= E[FP/Z]

| | # 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) |

# Korekce p-hodnot při mnohonásobném testování

## Kontrolujeme FWER

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

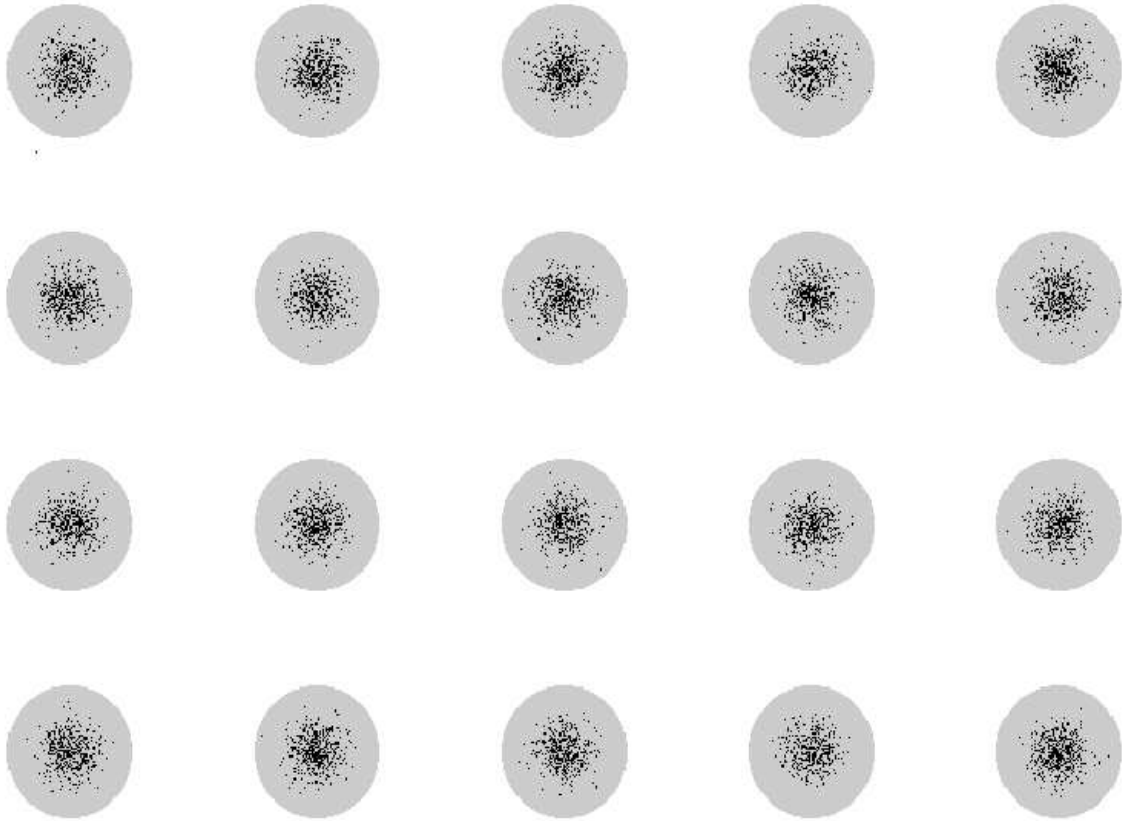
## Kontrolujeme FDR

- 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!

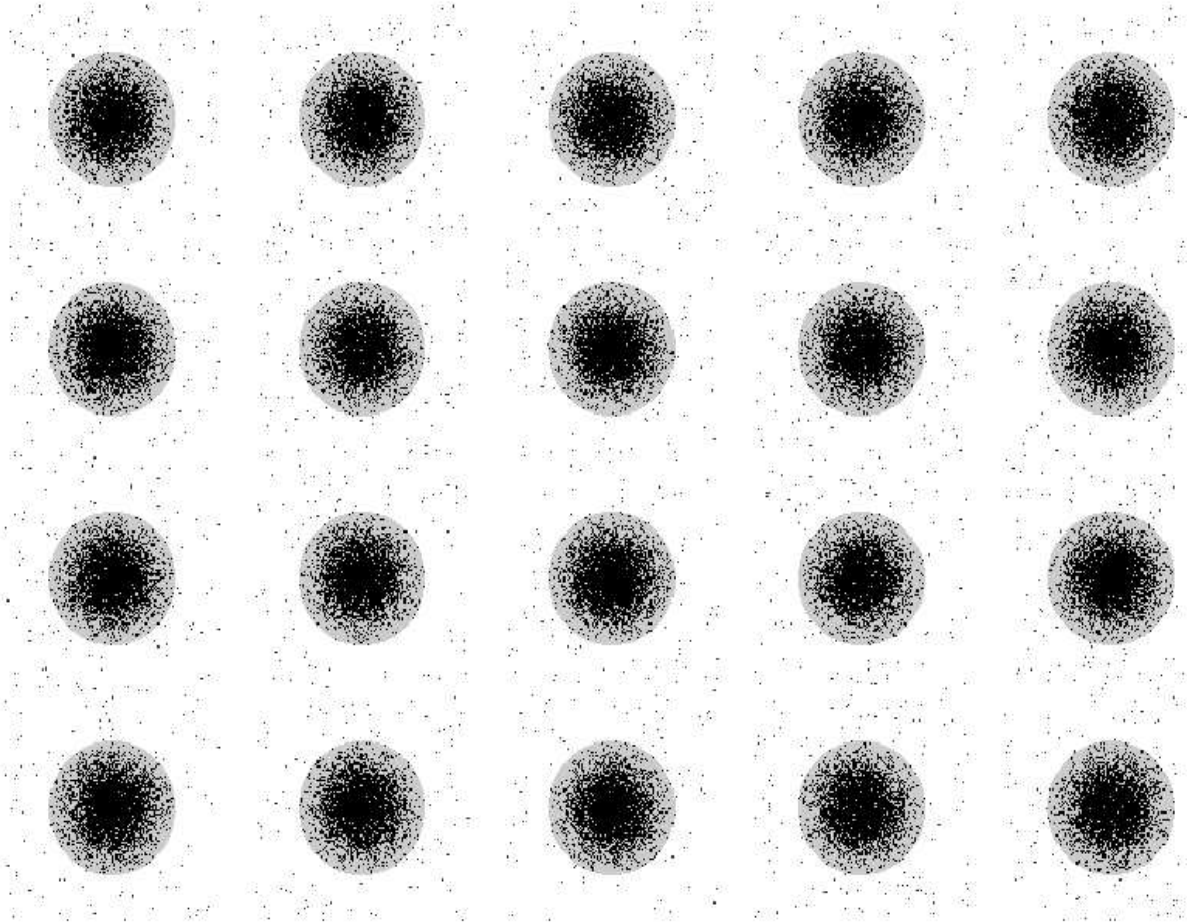
# Který typ korekce použít?

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

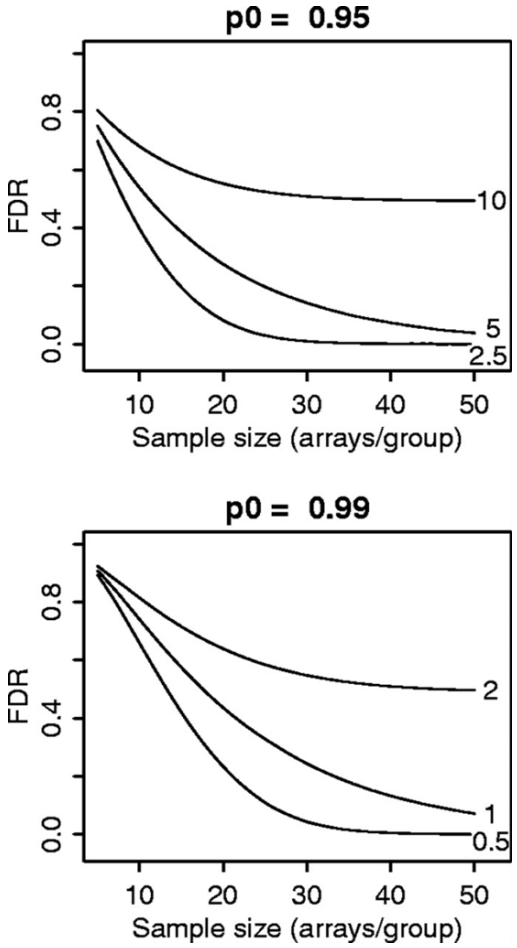


# Který typ korekce použít?

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



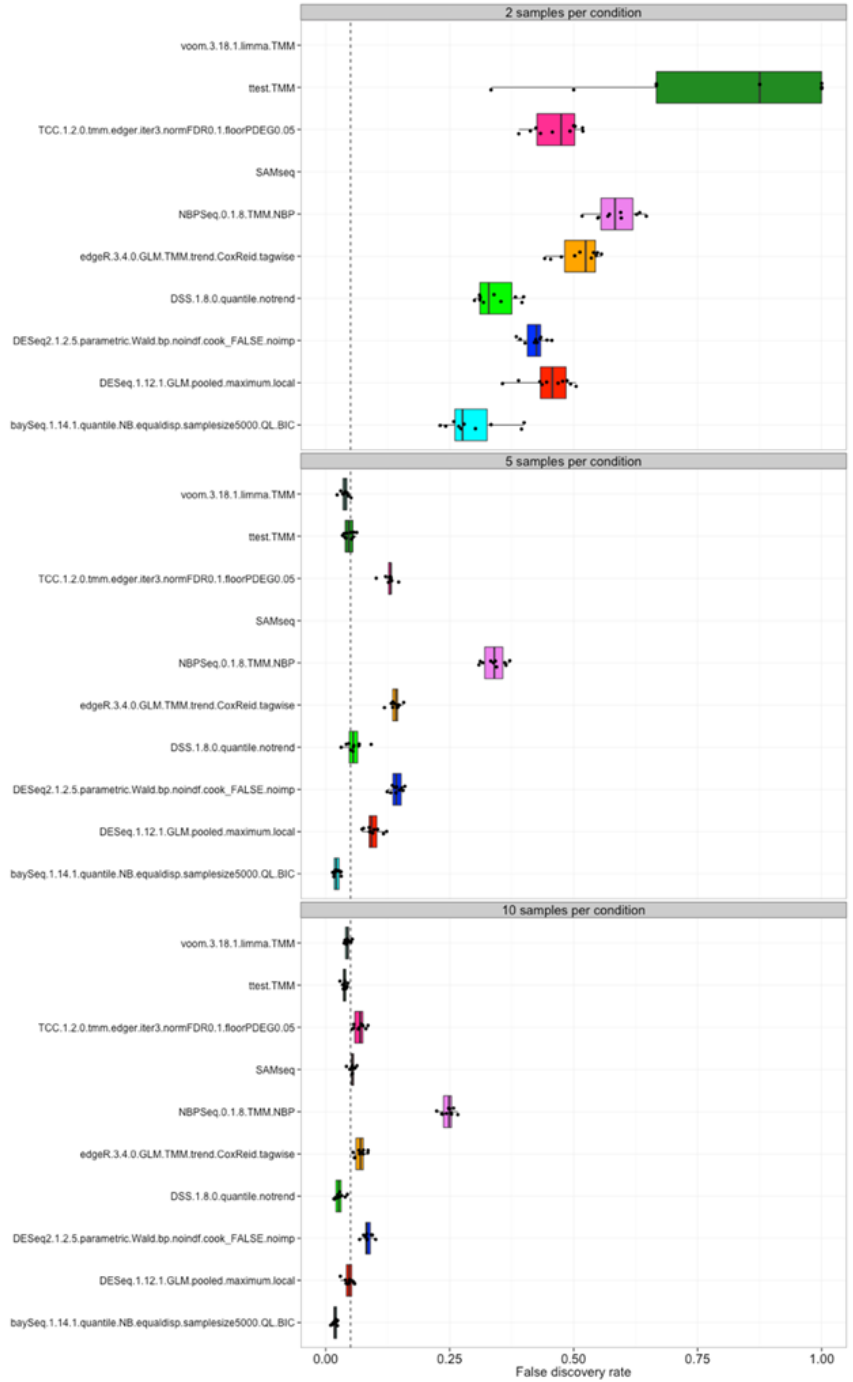
# Vliv počtu vzorků na falešně pozitivní výsledky




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

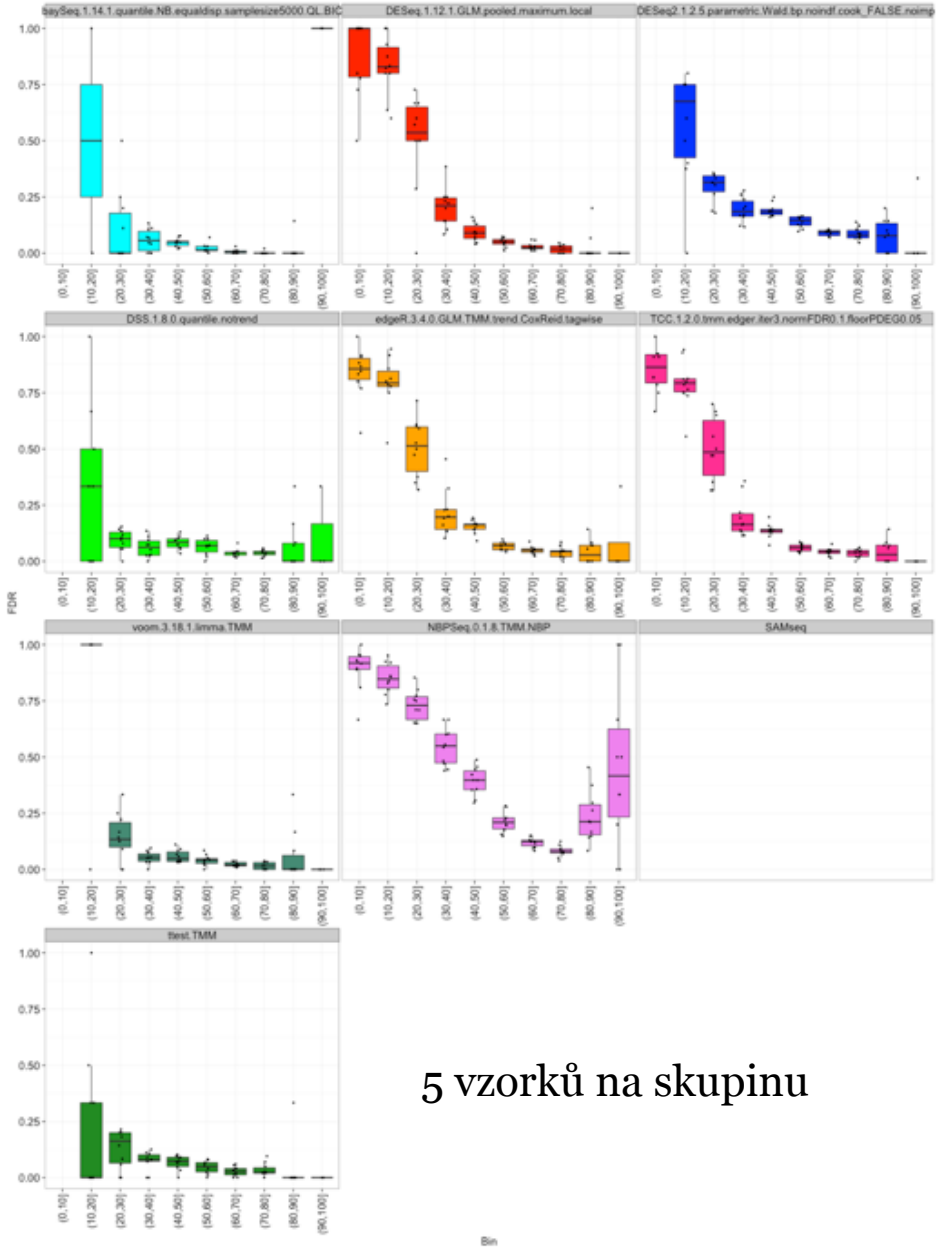
FDR (false discovery rate) jako funkce velikosti vzorku a percenta významných výsledků.
Každá křivka představuje fixní percento genů označených jako významných.

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

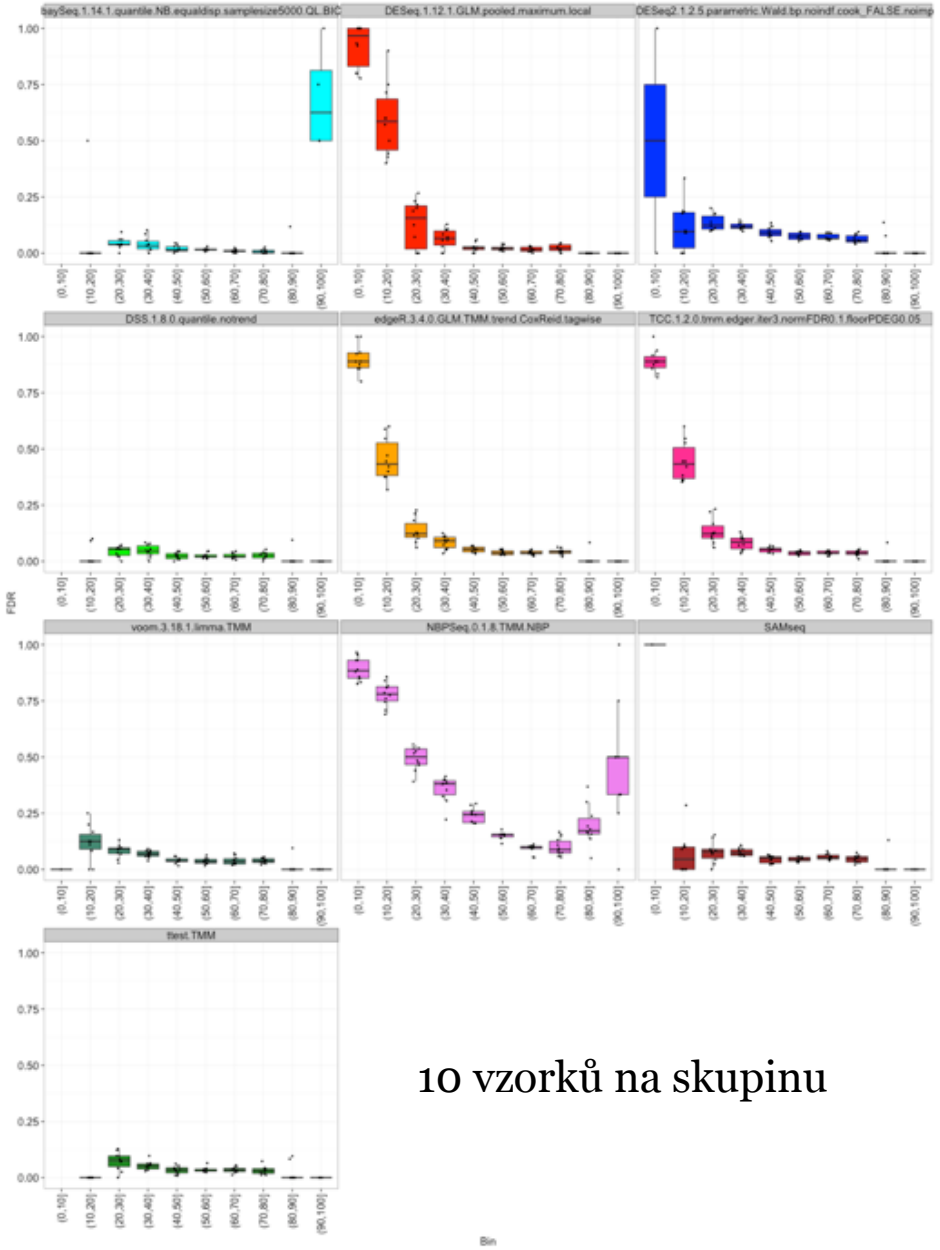



FDR (False discovery rate) jako funkce počtu vzorků na skupinu a metody použité pro normalizaci sekvenačních dat a testování hypotéz

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

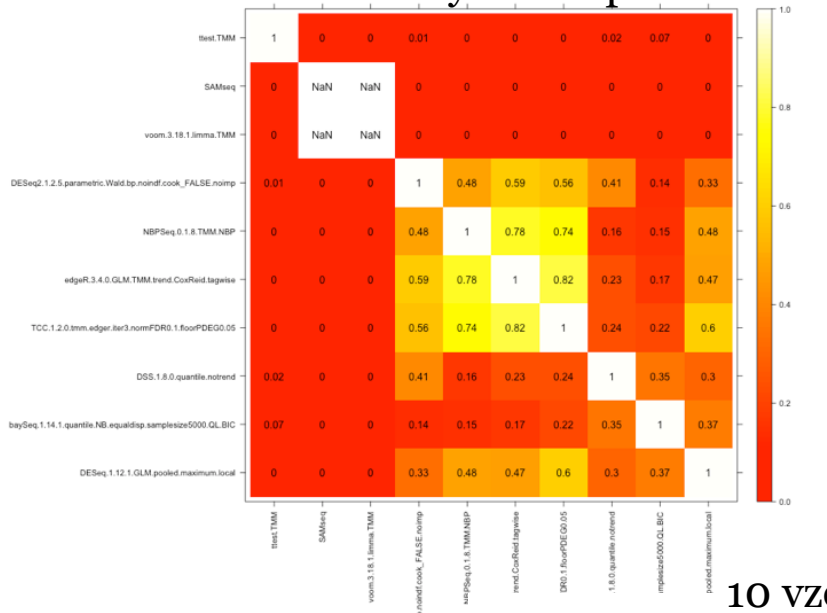



5 vzorků na skupinu

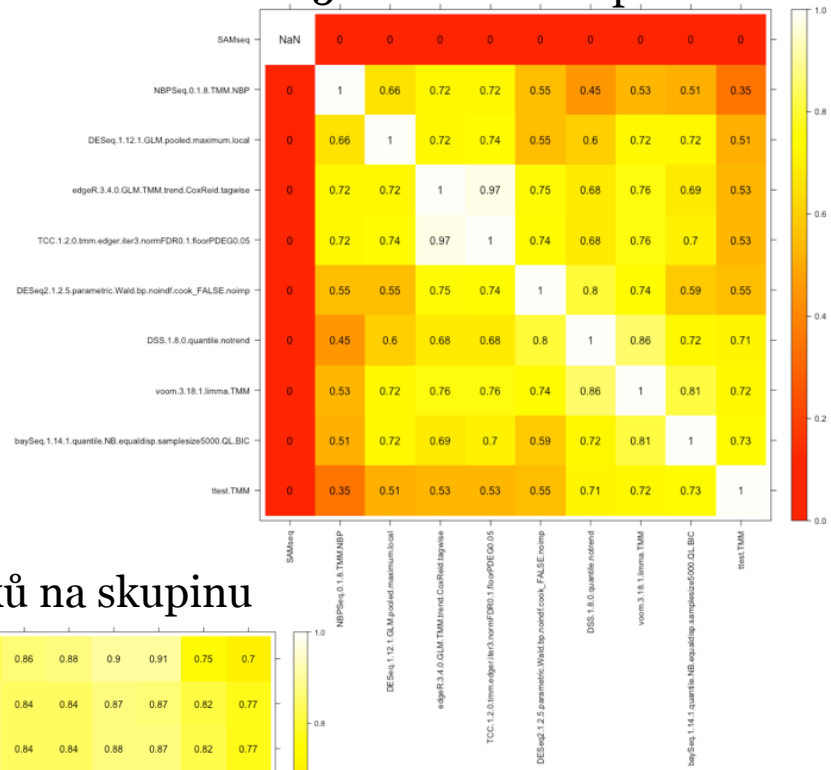



10 vzorků na skupinu

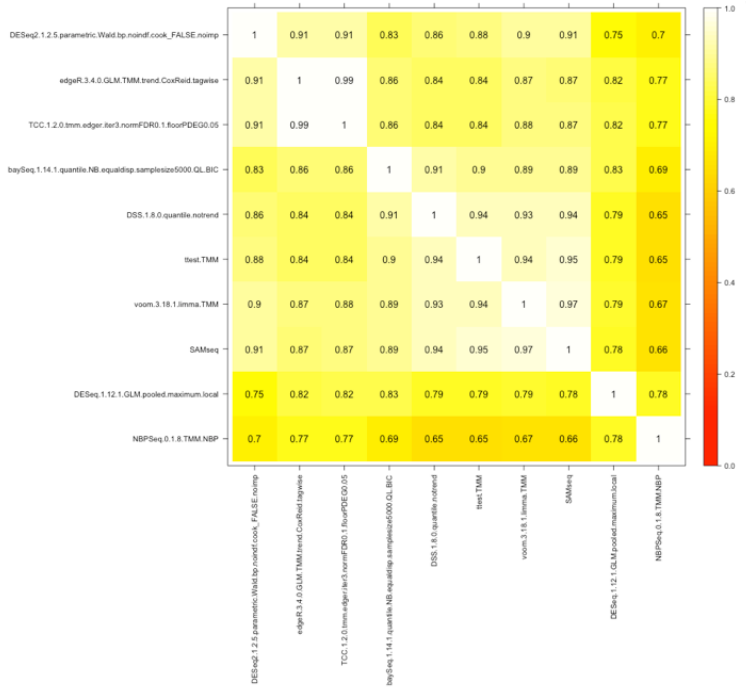
2 vzorky na skupinu




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

# 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*

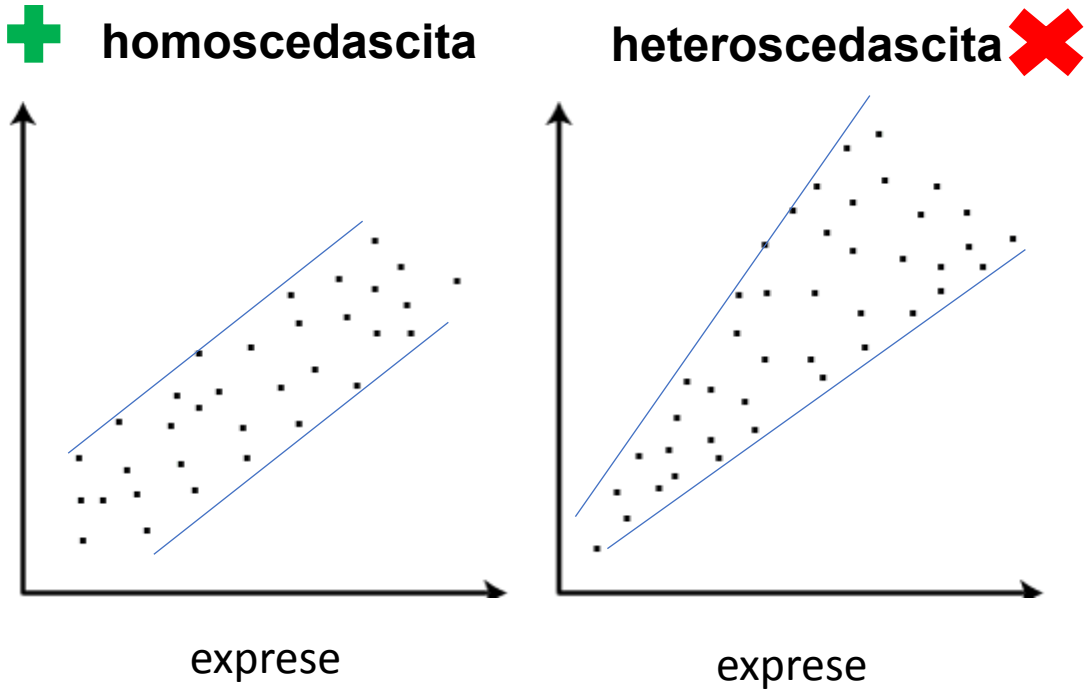
- Chceme najít pravděpodobnost, že vzorek patří do určité skupiny na základě expresní hodnoty daného genu
- *Logistická regrese*

# Můžeme používat klasické statistiky u omicsových dat?

# Problém omicsových dat

Příliš malé hodnoty exprese (blízké šumu) vykazují malou variabilitu a naopak (heteroscedascita)

=> vysoké T-statistiky u biologicky nerelevantních genů!



# Moderovaná T-statistika

Aby se daly statistiky porovnat, je potřeba nějako sjednotit variabilitu

**Buď se přidá konstanta korigující variabilitu**

**Alternativně se data znormalizují tak aby variabilita byla stejná**

# Significance analysis of microarrays (SAM)

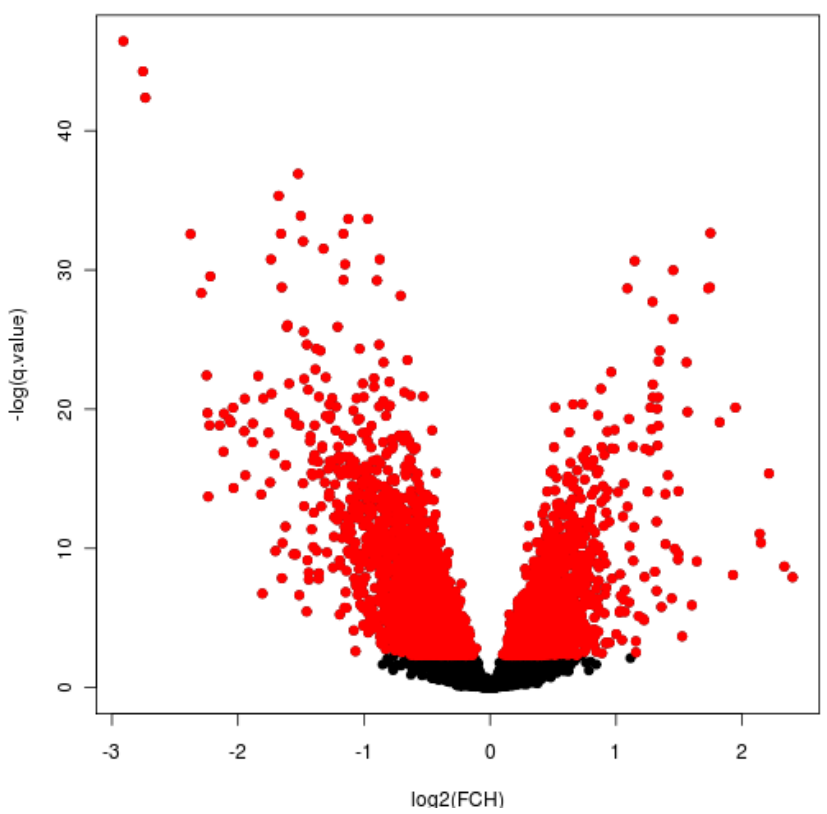
- Tusher, Tibshirani a Chu (2001)
- Založená na moderované *t*-statistice ($d_g$), počítá FDR
- Statistická významnost $d_g$ je následně stanovená permutacemi původních dat a kalkulací očekávaného skóre v případě, že platí nulová hypotéza ($d_e$)
- Gen je statisticky významný, pokud splňuje podmínku $|d_g - d_e| > \Delta$.
- Výhody: jednoduché
- Nevýhody: výpočtově náročné (permutace)
- Výstup: *q*-hodnoty
- `biocLite("samr")`
- `library(samr)`

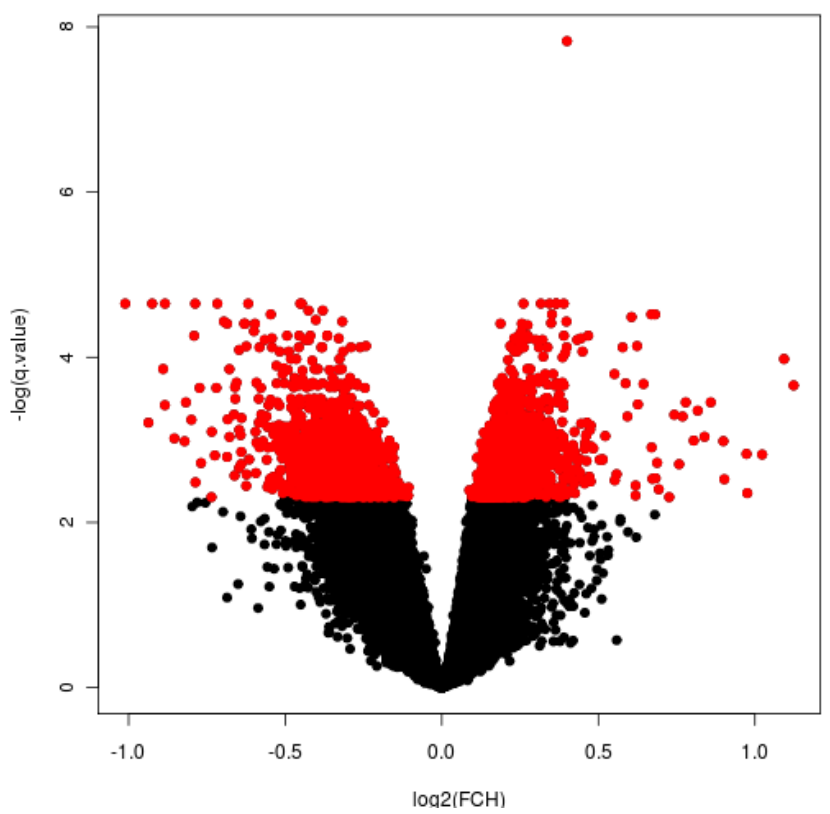
# Limma

- 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

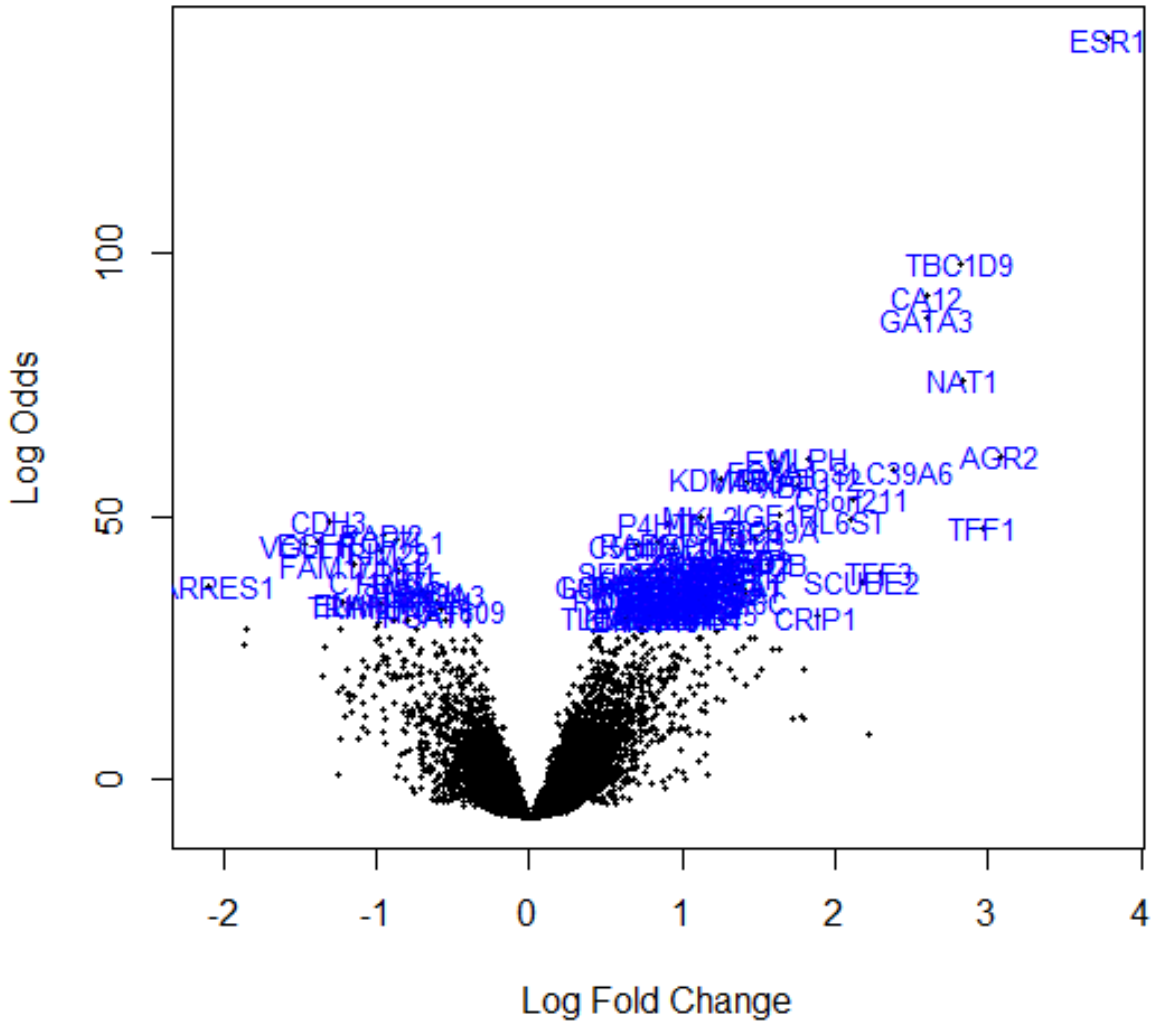
# Typické zobrazení významnosti genů
# Volcano plot

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



# Volcano plot



# Cvičení

- Připravili jsme pro Vás cvičení na genových expresních datech

- Zadejte do prohlížeče tuto adresu:

Pokud jste na MU, nebo připojeni přes VPN: 10.16.117.29

Pokud jste mimo MU: 147.251.21.95