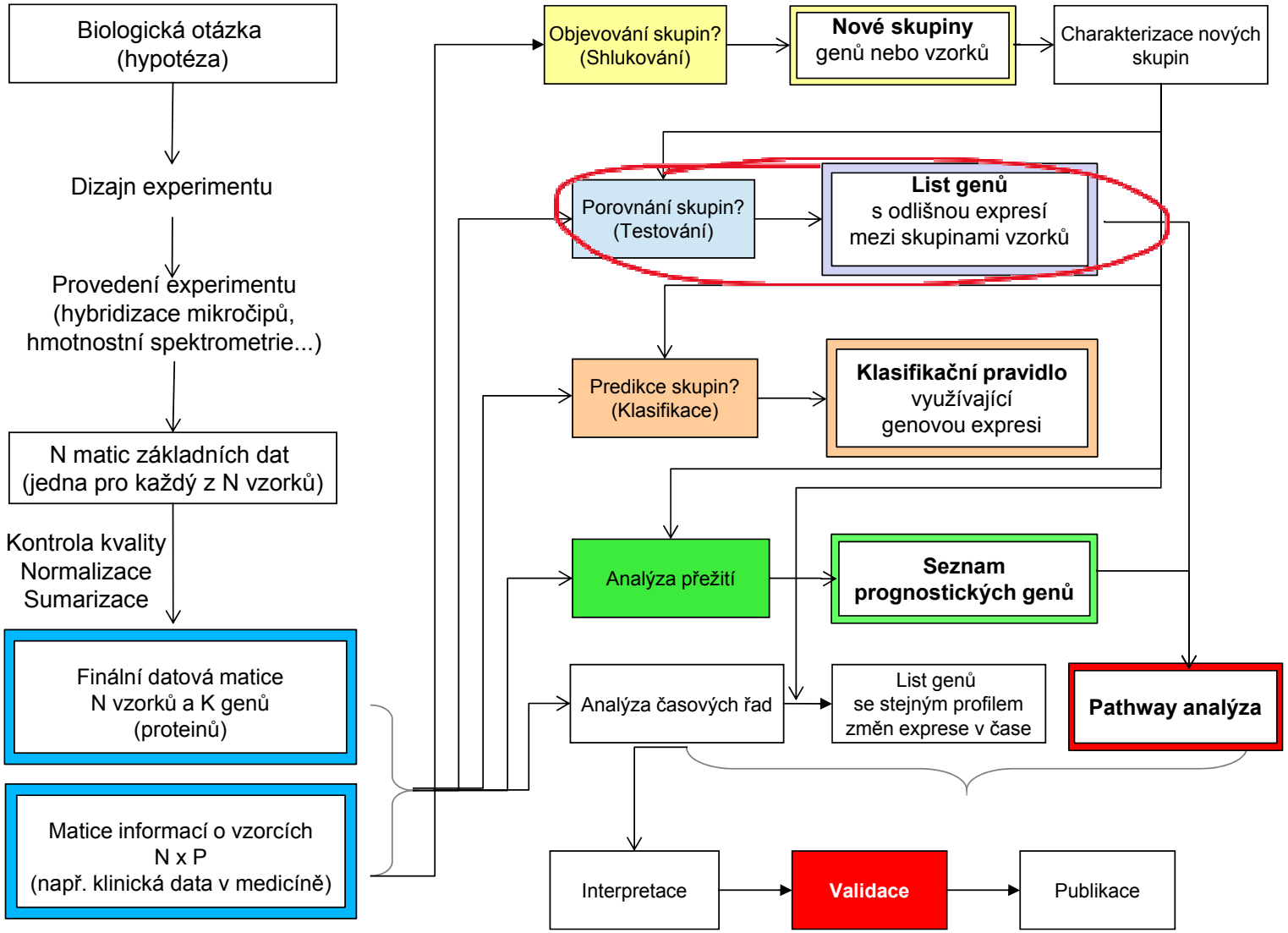
Detekce biomarkerů z omics experimentů

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

- Podzim 2024

Hledání rozdílů mezi skupinami

Jak se hledá potenciální biomarker v omics datech



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

Přístupy

Jednoduchá 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

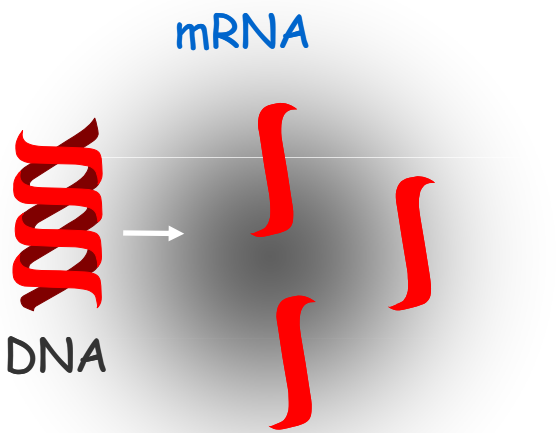
Příklad:

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

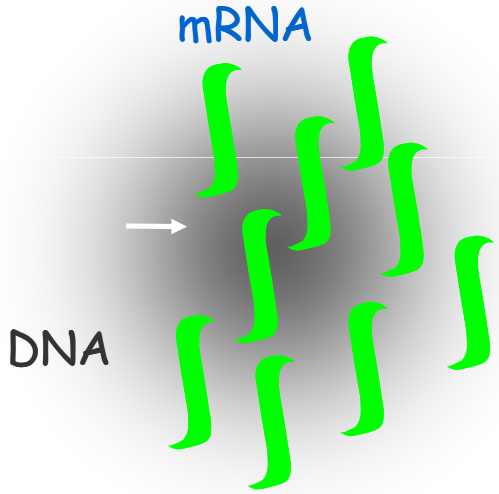
Výhoda: jednoduché

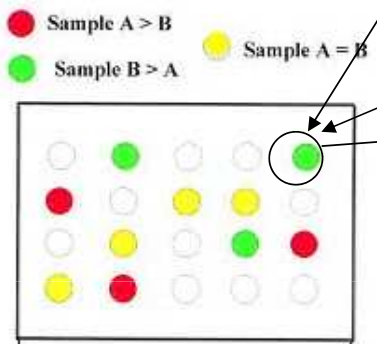
Metoda dělící hranice velikosti efektu / změny

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

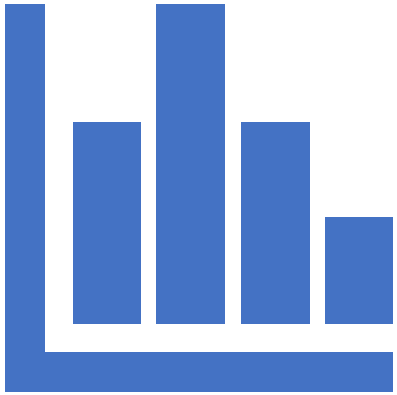
Testování hypotéz

- 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...)*

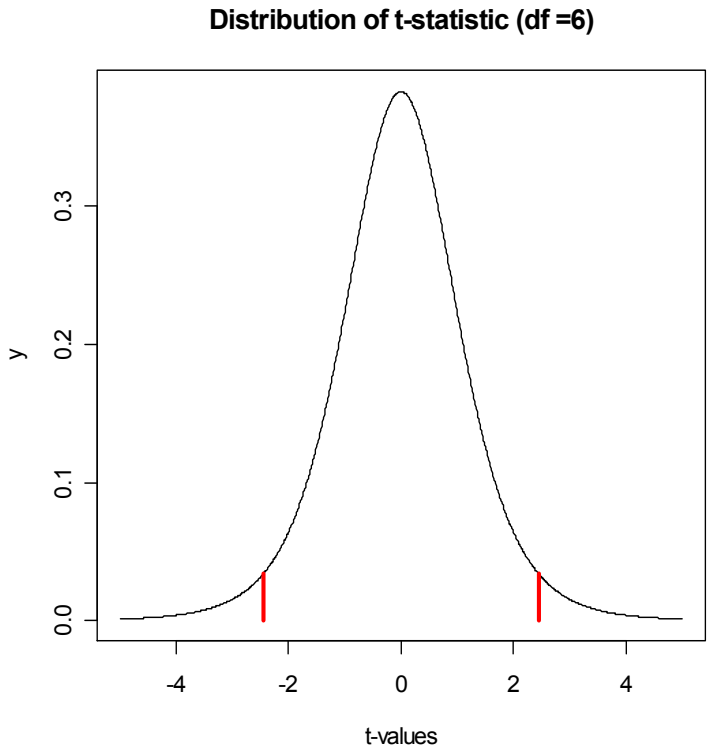
- Statistiky jsou definovány různě a mají různé předpoklady.
- Například T-statistika porovnává signál se šumem a předpokládá normalitu dat.

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*

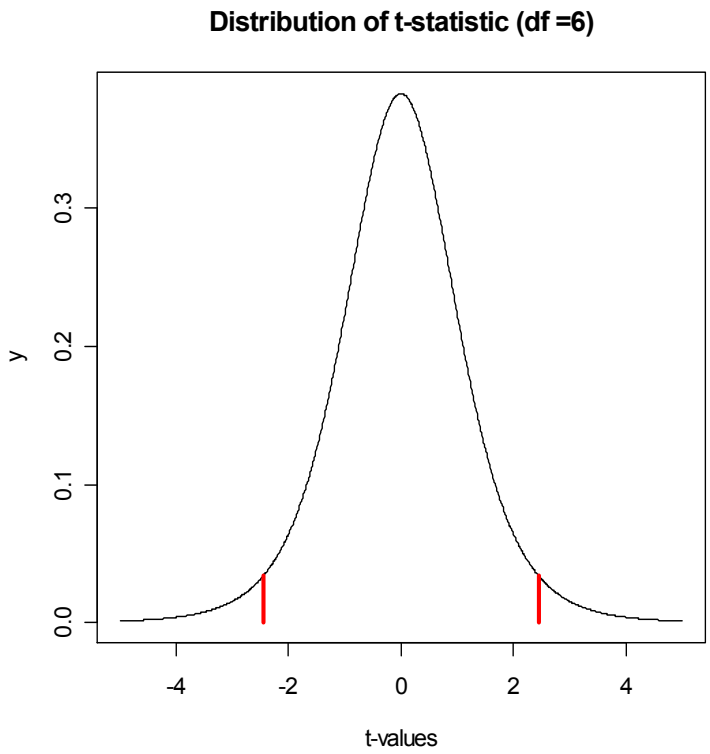
T-test a T-statistika



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

Variabilita (vyjádřená jako směrodatná odchylka)

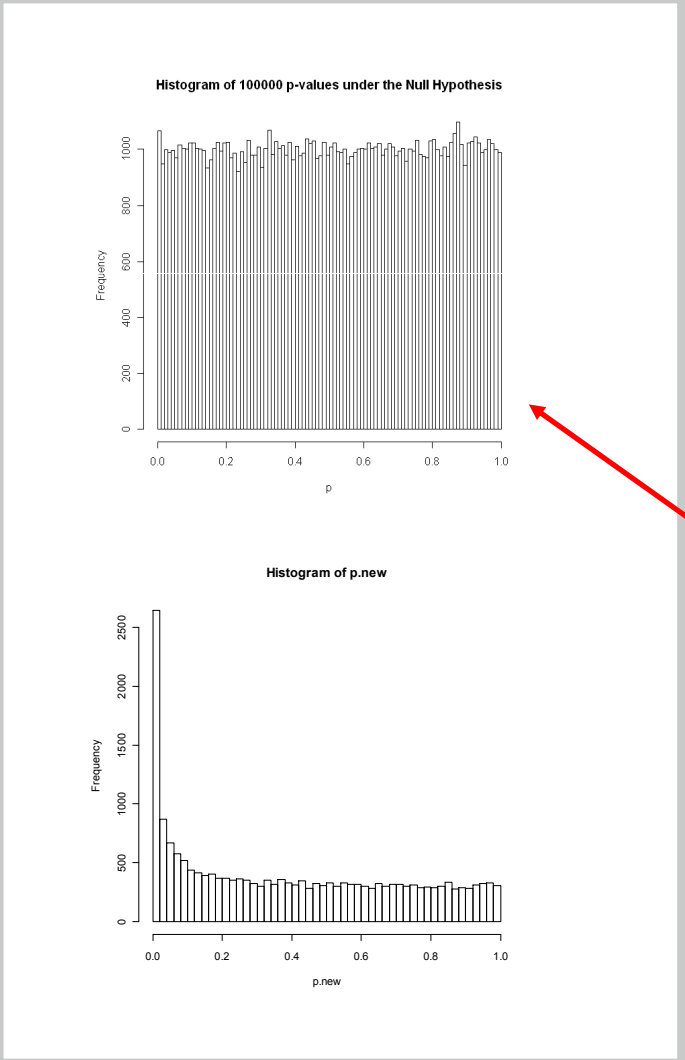
T-test a T-statistika



- 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)

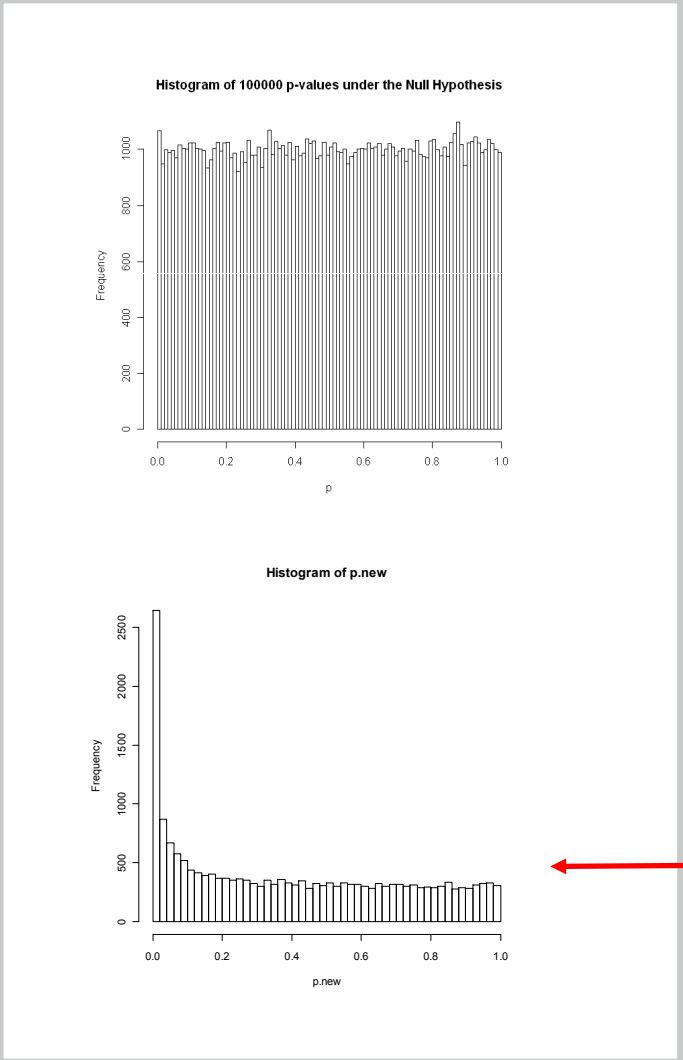


Testování hypotéz

- 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**.
-



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. $p < 0.05/10\,000$)

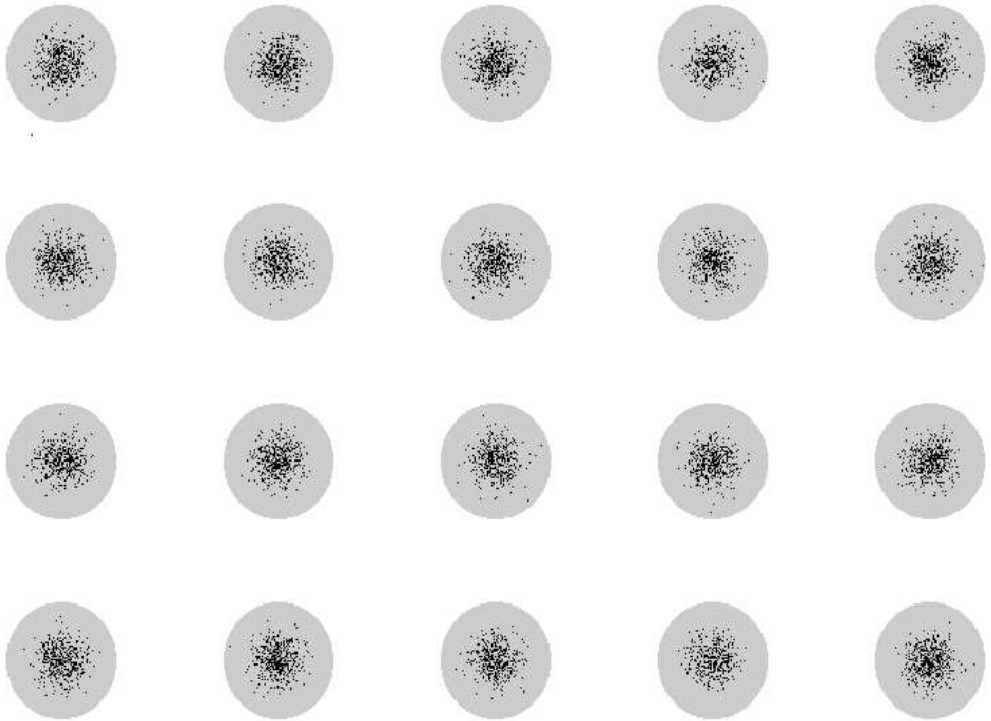
Kontrolujeme FDR

- 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!

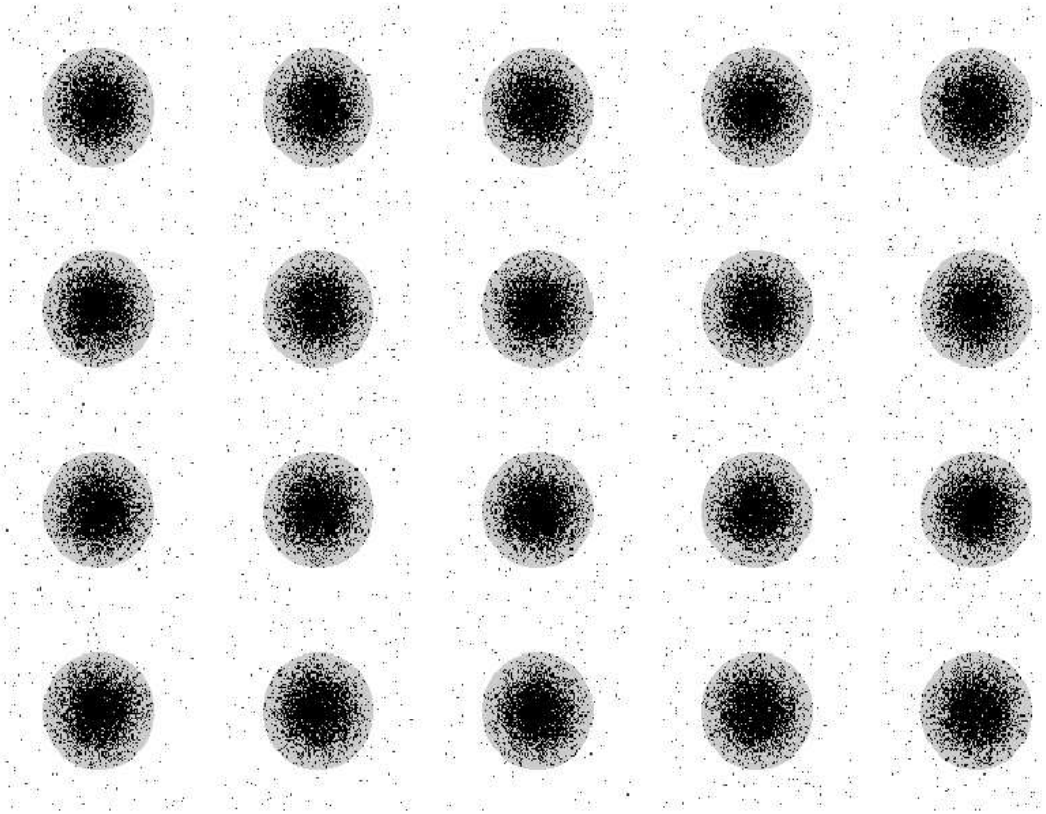
Který typ korekce použít?

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

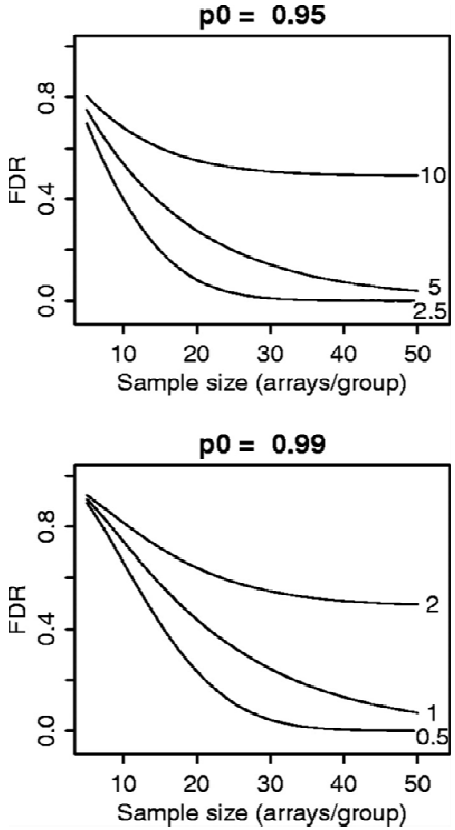


Který typ korekce použít?

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



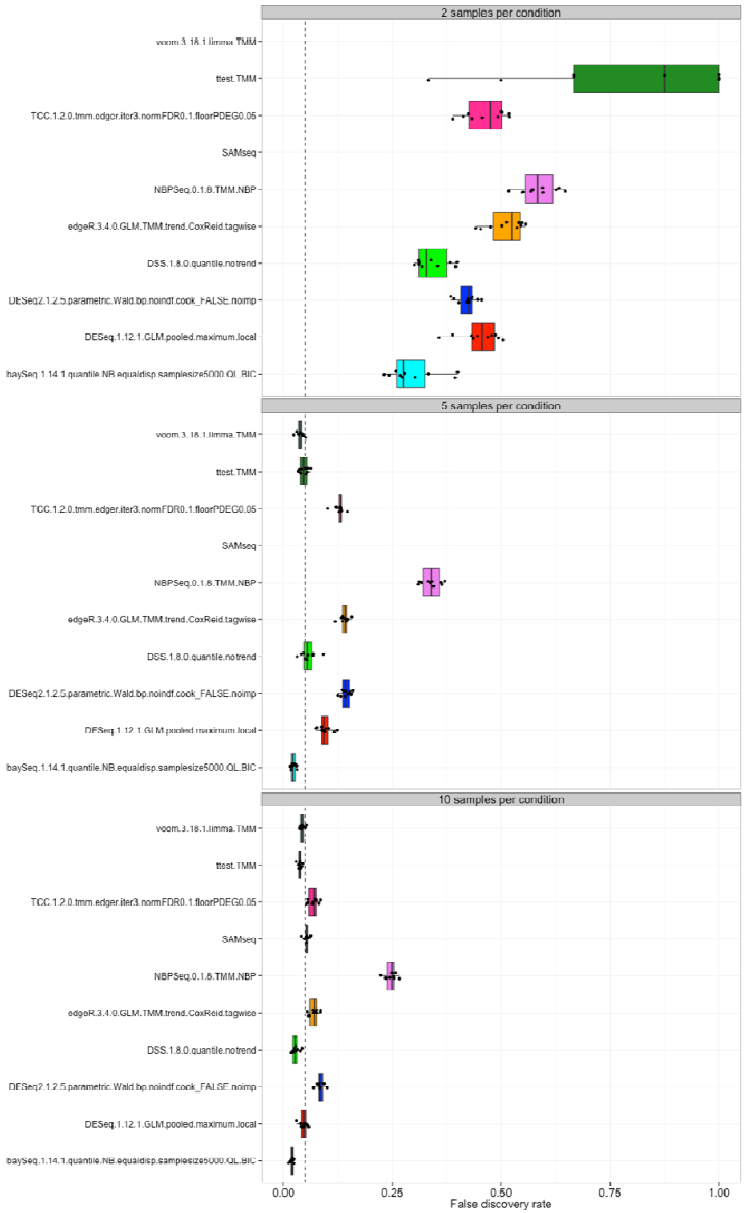
Vliv počtu vzorků na falešně pozitivní výsledky



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

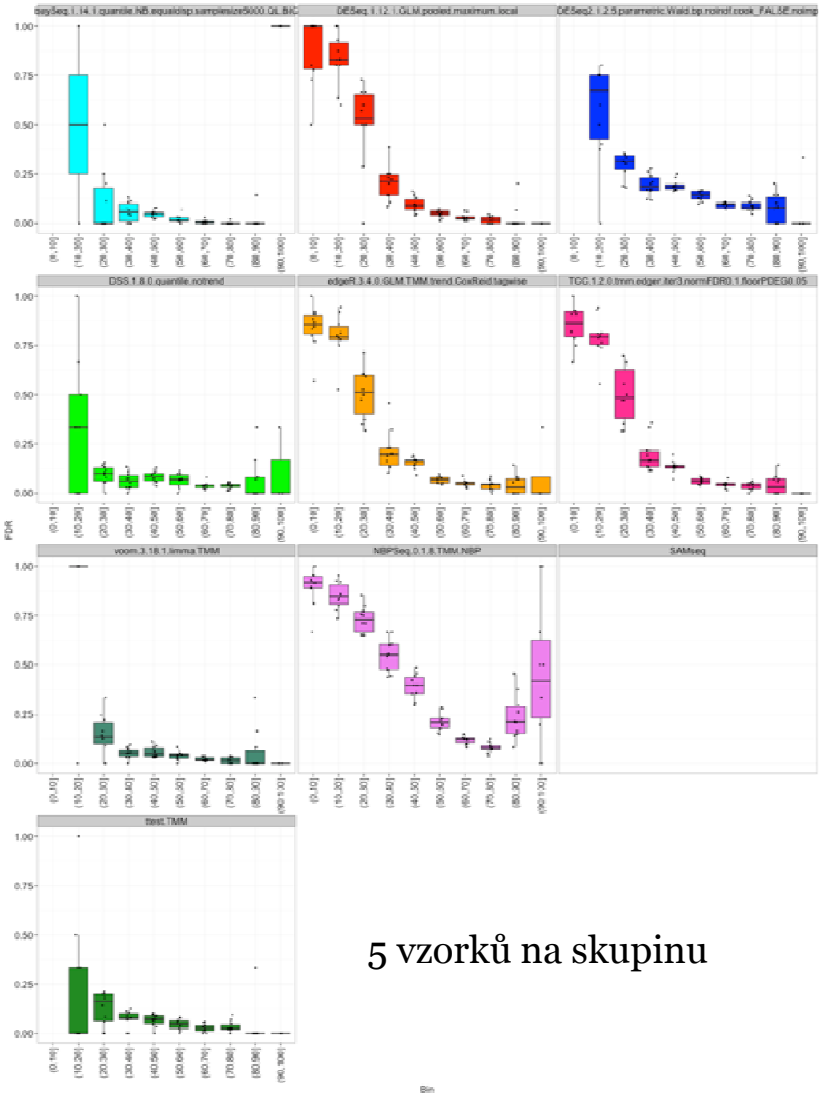
FDR (false discovery rate) jako funkce velikosti vzorku a procenta významných výsledků.
Každá křivka představuje fixní procento genů označených jako významných.

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

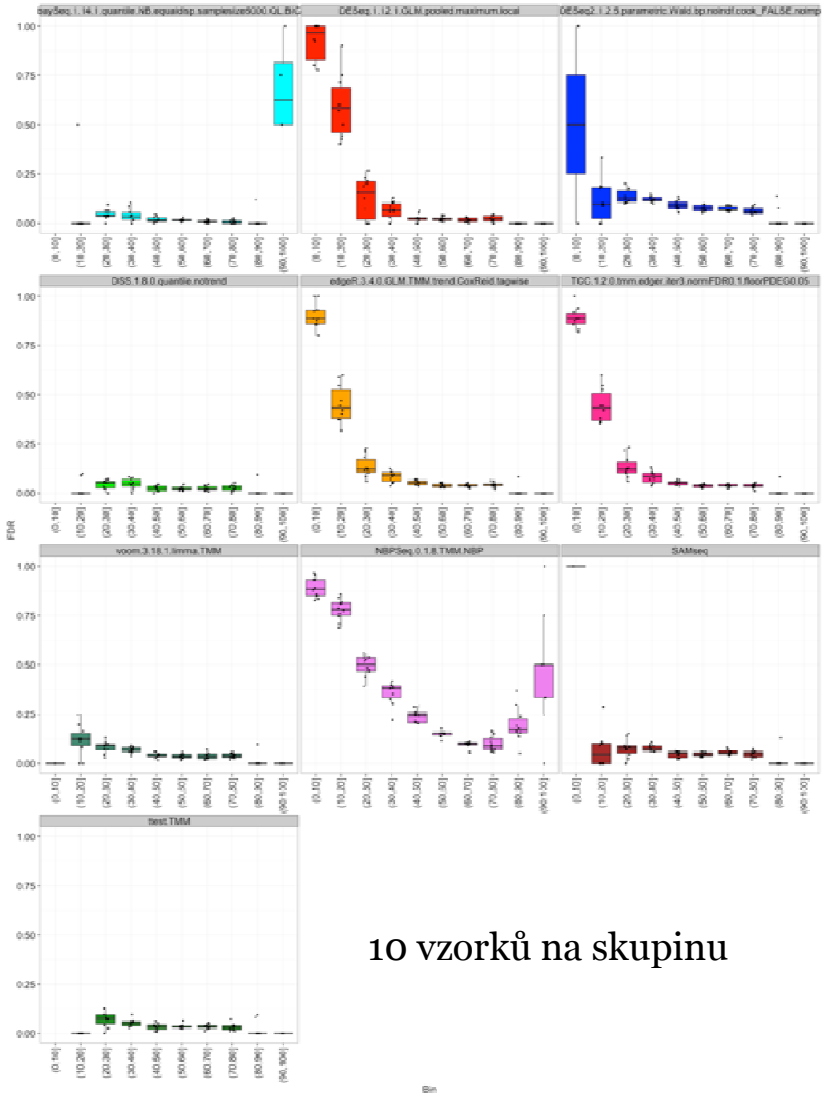


FDR (False discovery rate) jako funkce počtu vzorků na skupinu a metody použité pro normalizaci sekvenačních dat a testování hypotéz

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

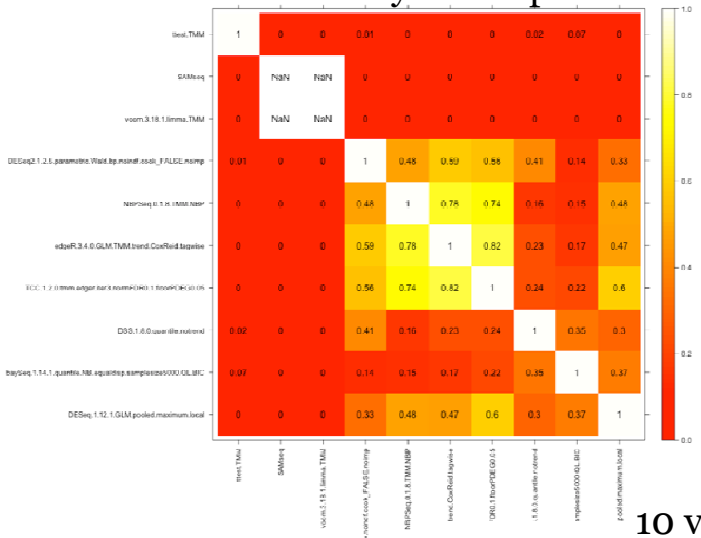


5 vzorků na skupinu

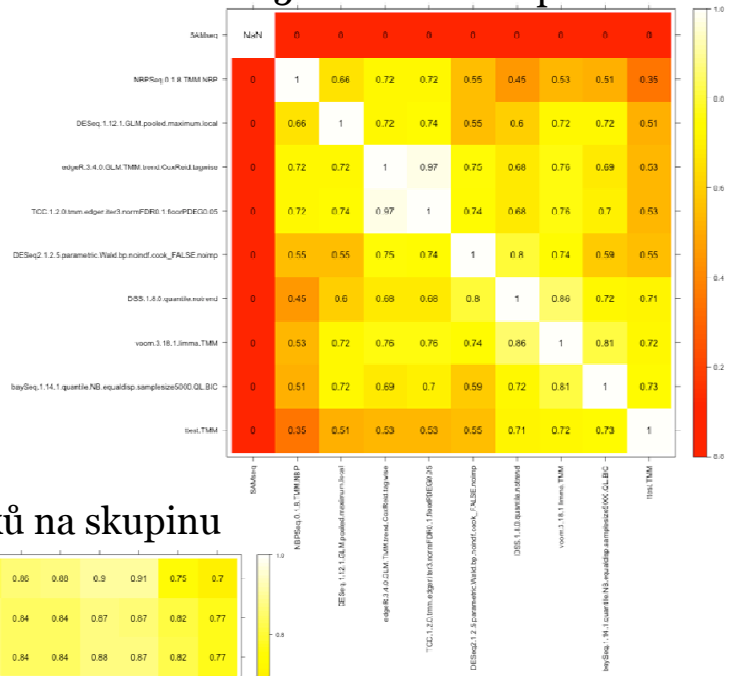


10 vzorků na skupinu

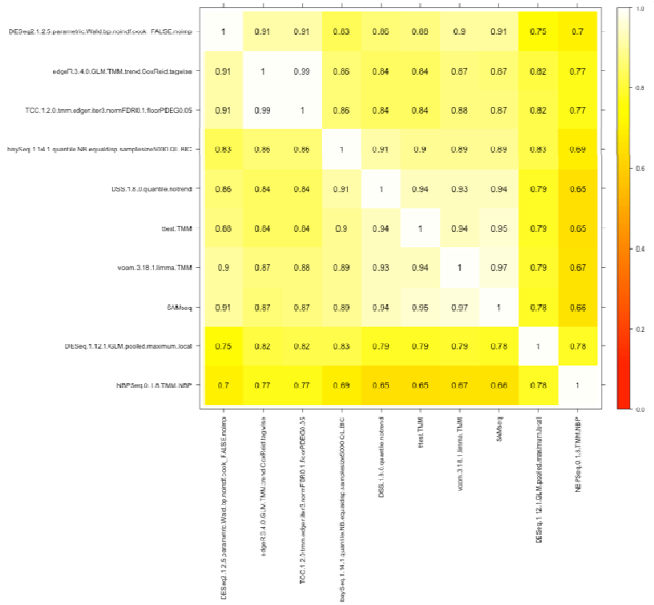
2 vzorky na skupinu



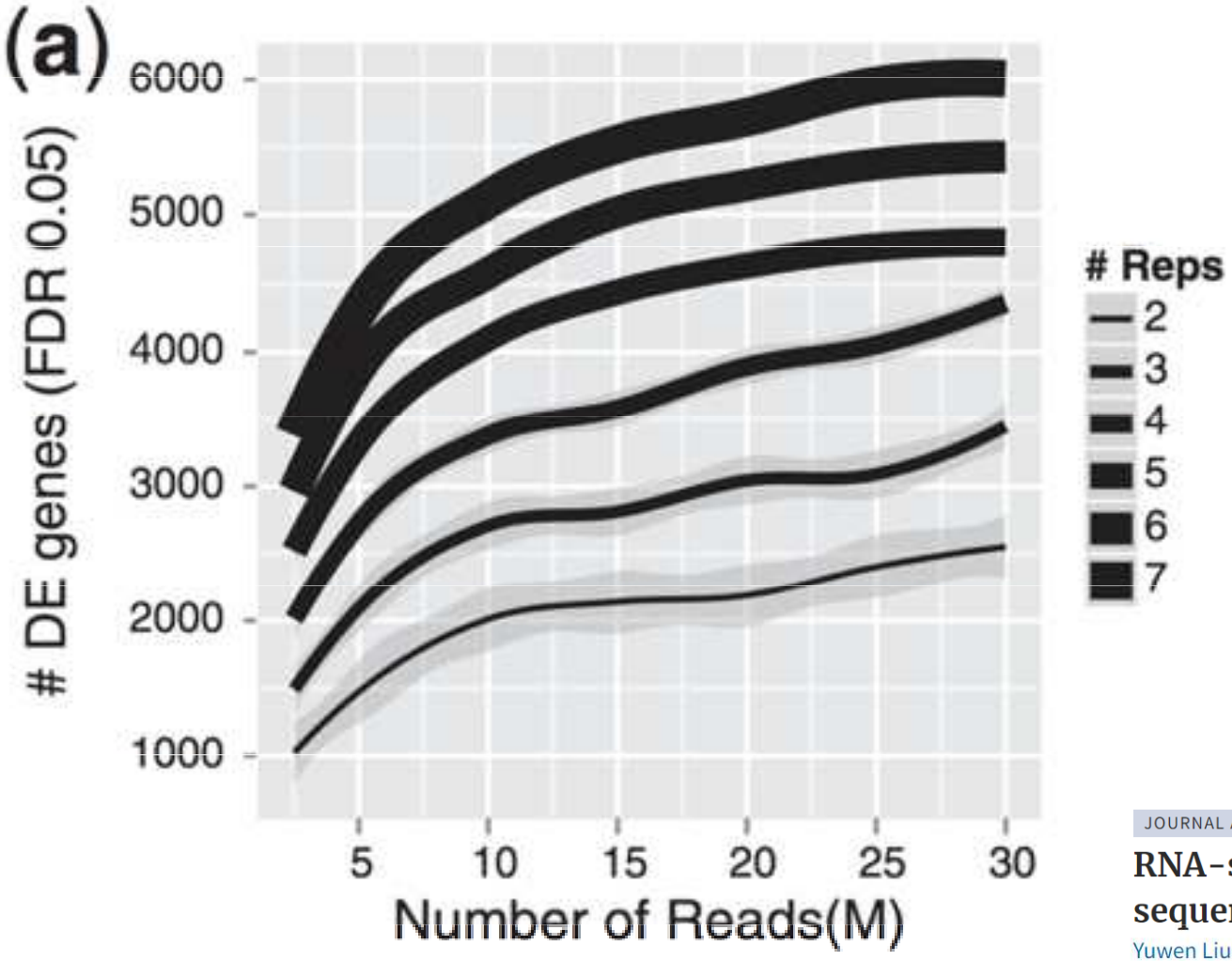
5 vzorků na skupinu



10 vzorků na skupinu



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



Je lepší mít víc vzorků osekvenovaných méně hluboko než málo hluboce osekvenovaných vzorků.

[RNA-seq differential expression studies: more sequence or more replication? | Bioinformatics | Oxford Academic]

JOURNAL ARTICLE

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu, Jie Zhou, Kevin P. White ✉ Author Notes

Bioinformatics, Volume 30, Issue 3, February 2014, Pages 301–304,
https://doi.org/10.1093/bioinformatics/btt688
Published: 04 December 2013 **Article history** ▾

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

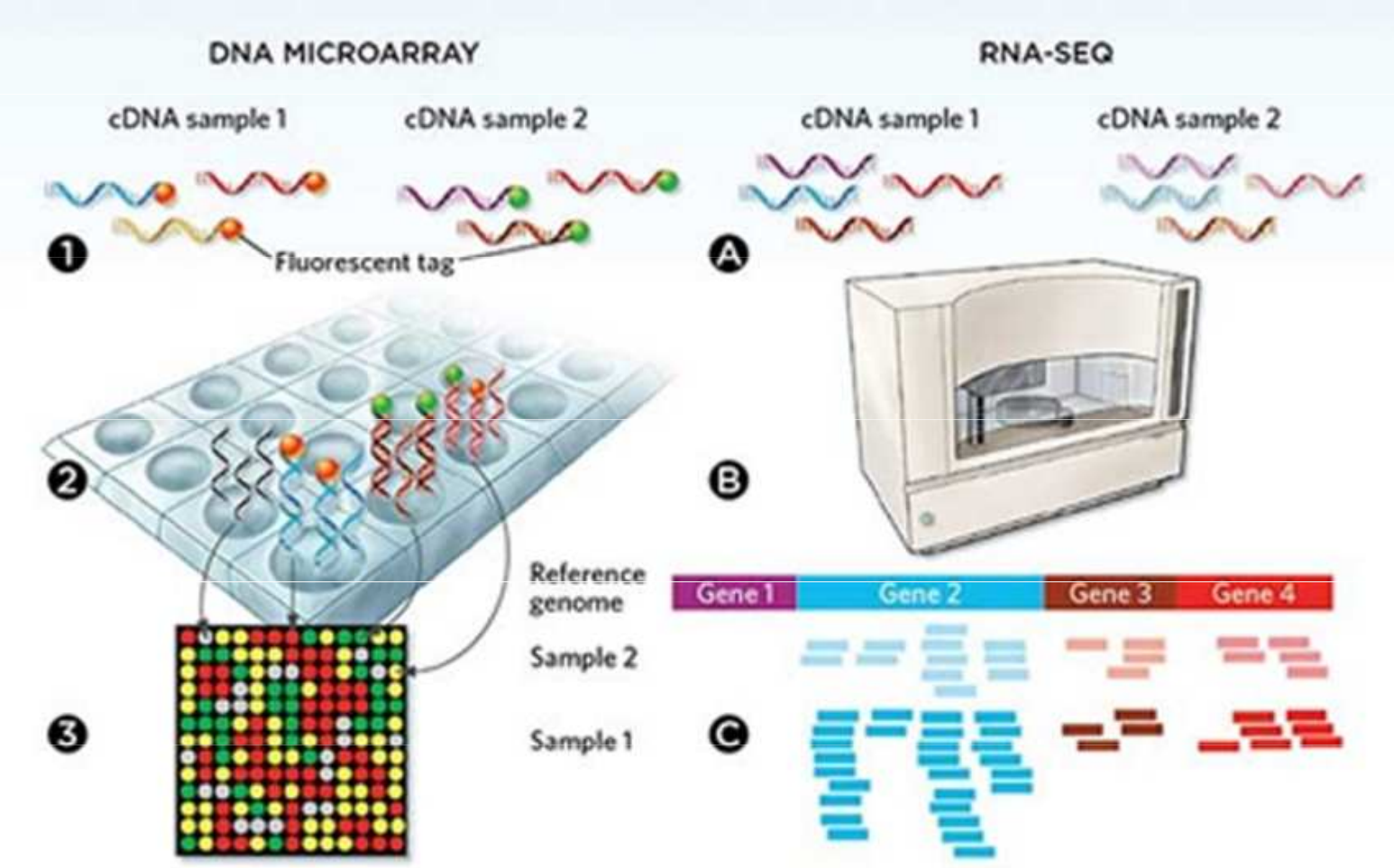
- Pokud máme víc jak 1 proměnnou, která může ovlivnit genovou/proteinovou expresi
  - genová exprese ~ skupina + pohlaví
  - *Lineární modelování (limma)*

- 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*

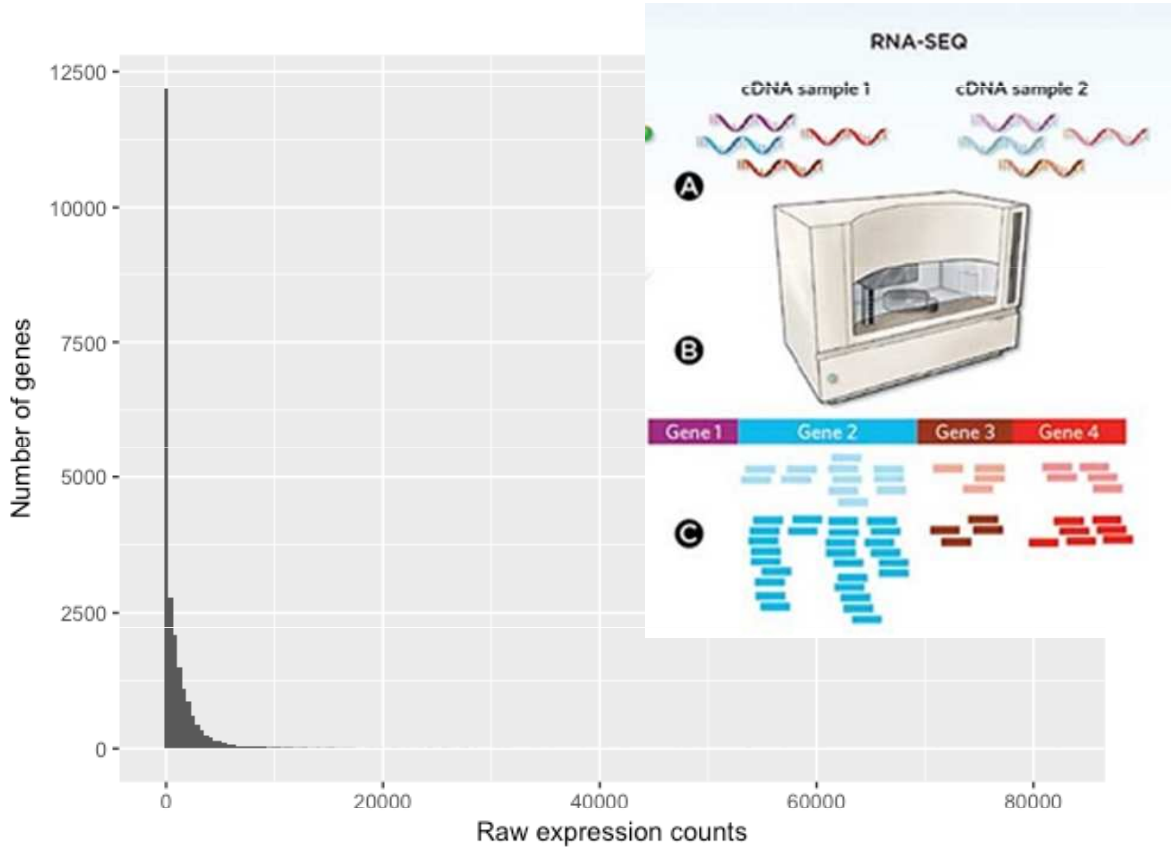
Můžeme používat běžné statistické testy u omicsových dat?

Není měření jako měření



Příklad dat z RNAseq experimentu

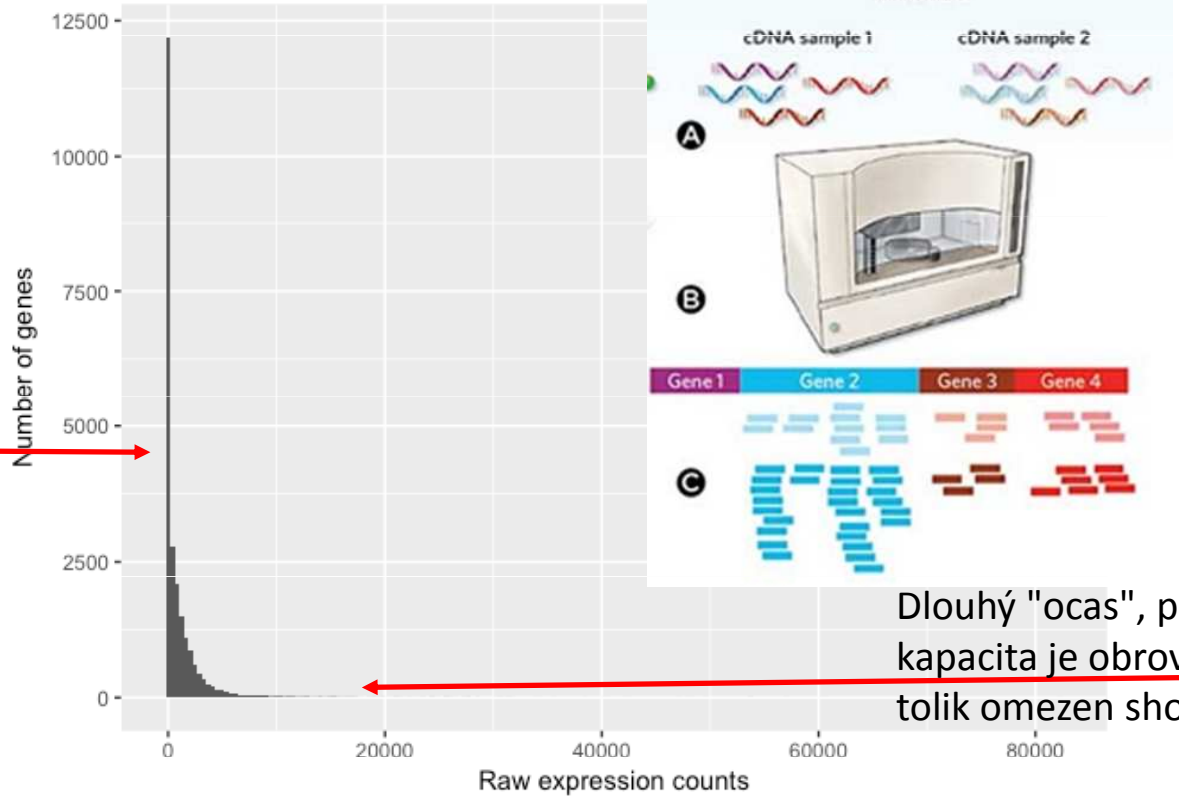
Data exprese genu jsou vyjádřené jako **POČTY ČTENÍ** (od 0 **do maximální kapacity přístroje - sdílí s jinými geny**)



Příklad dat z RNAseq experimentu

Data exprese genu jsou vyjádřené jako **POČTY ČTENÍ** (od 0 **do maximální kapacity přístroje - sdílí s jinými geny**)

Většina genů má velice nízkou expresi (počet čtení 0-100)



Dlouhý "ocas", protože kapacita je obrovská, gen není tolik omezen shora

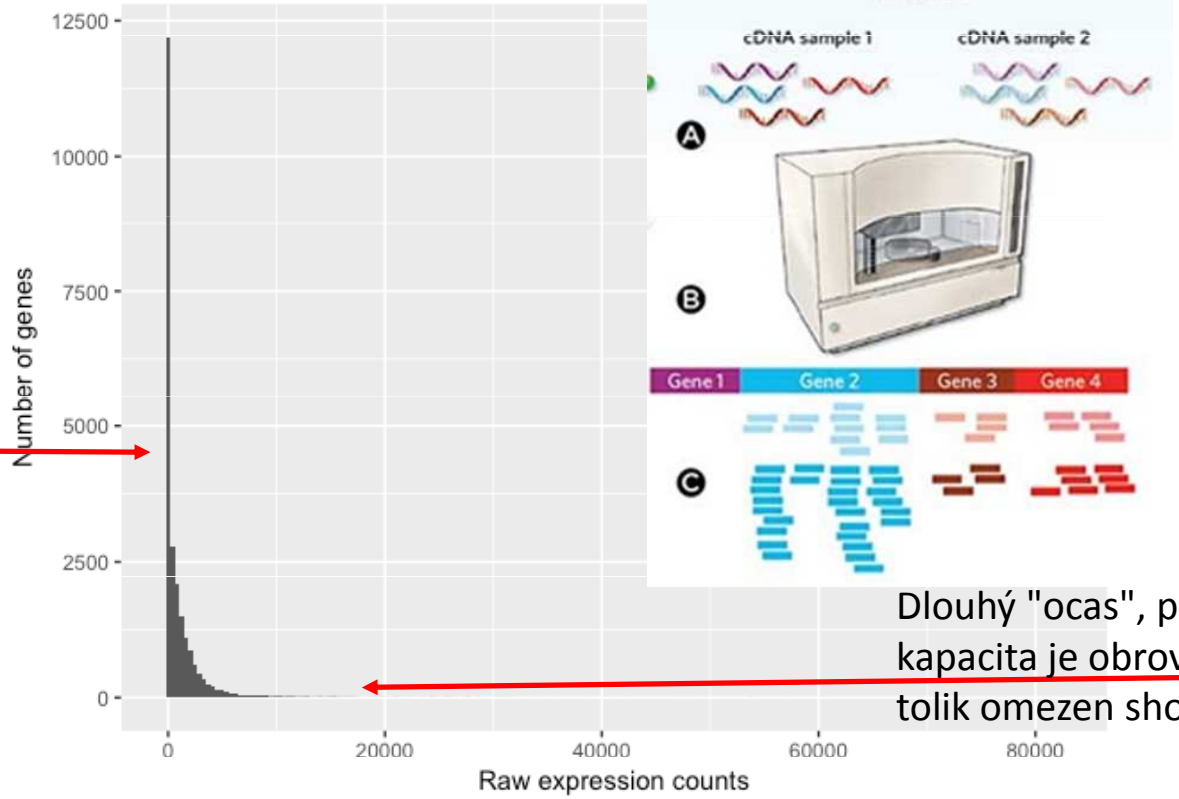
Příklad dat z RNAseq experimentu

Data exprese genu jsou vyjádřené jako **POČTY ČTENÍ** (od 0 **do maximální kapacity přístroje - sdílí s jinými geny**)

Většina genů má velice nízkou expresi (počet čtení 0-100)

PROČ:

- Silně exprimované geny **"vyžerou"** kapacitu sekvenátora a nezůstane na ty ostatní, málo exprimované (i když se silným efektem)



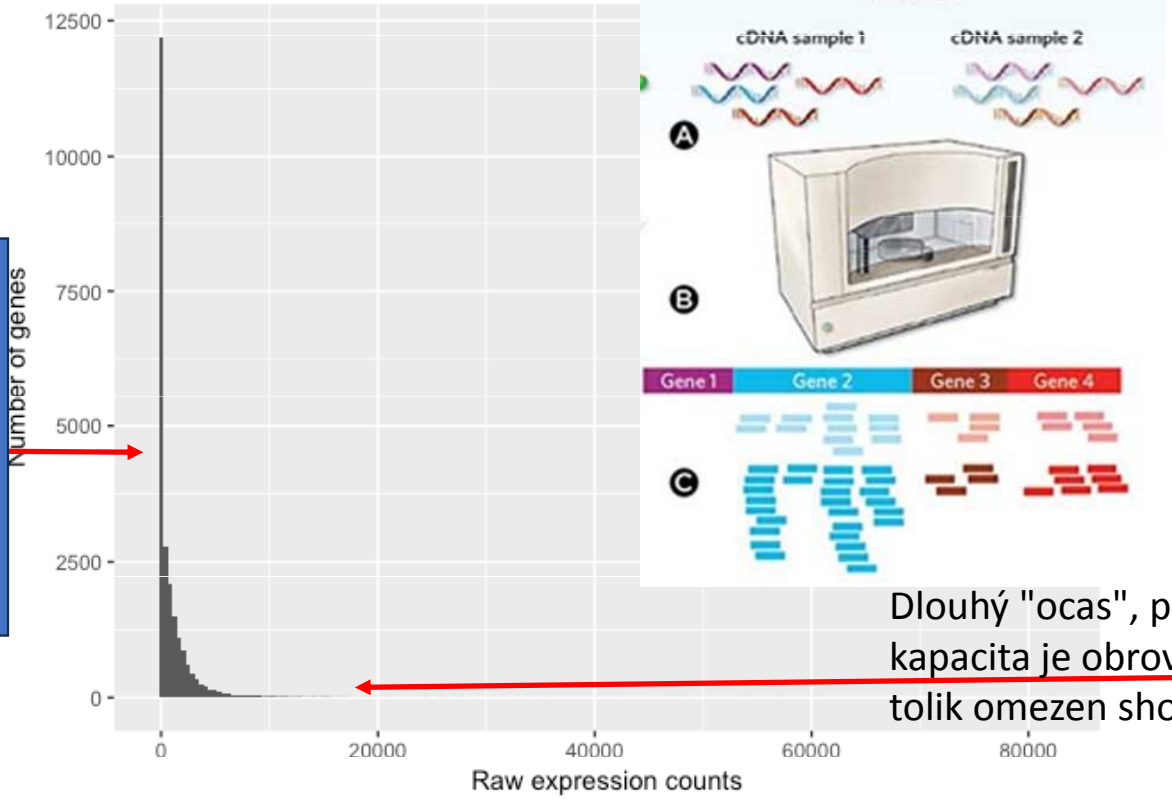
Dlouhý "ocas", protože kapacita je obrovská, gen není tolik omezen shora

Příklad dat z RNAseq experimentu

Data exprese genu jsou vyjádřené jako **POČTY ČTENÍ** (od 0 **do maximální kapacity přístroje - sdílí s jinými geny**)

Data nemají **Normální**, ale **Poisonnovo** rozložení...?

"vyžerou" kapacitu sekvenátora a nezůstane na ty ostatní, málo exprimované (i když se silným efektem)

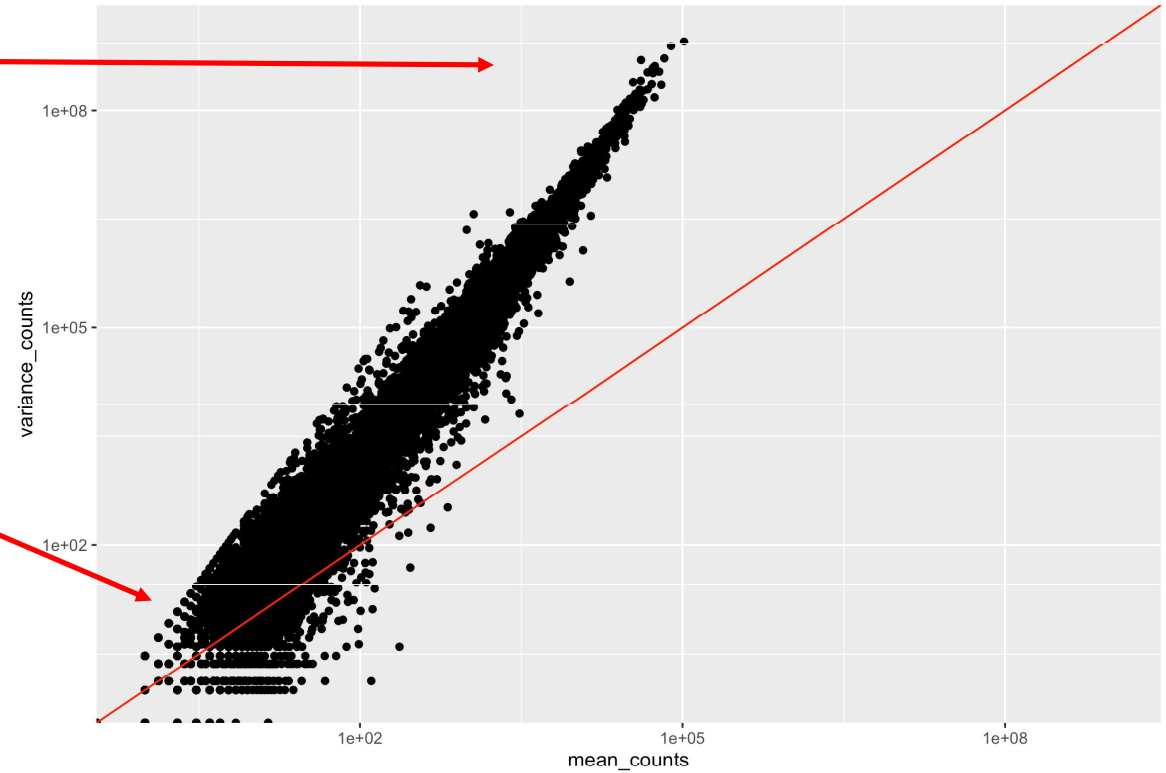


Dlouhý "ocas", protože kapacita je obrovská, gen není tolik omezen shora

Heteroskedasticita RNAseq dat

- Geny s vyšší expresí mají mnohem vyšší variabilitu

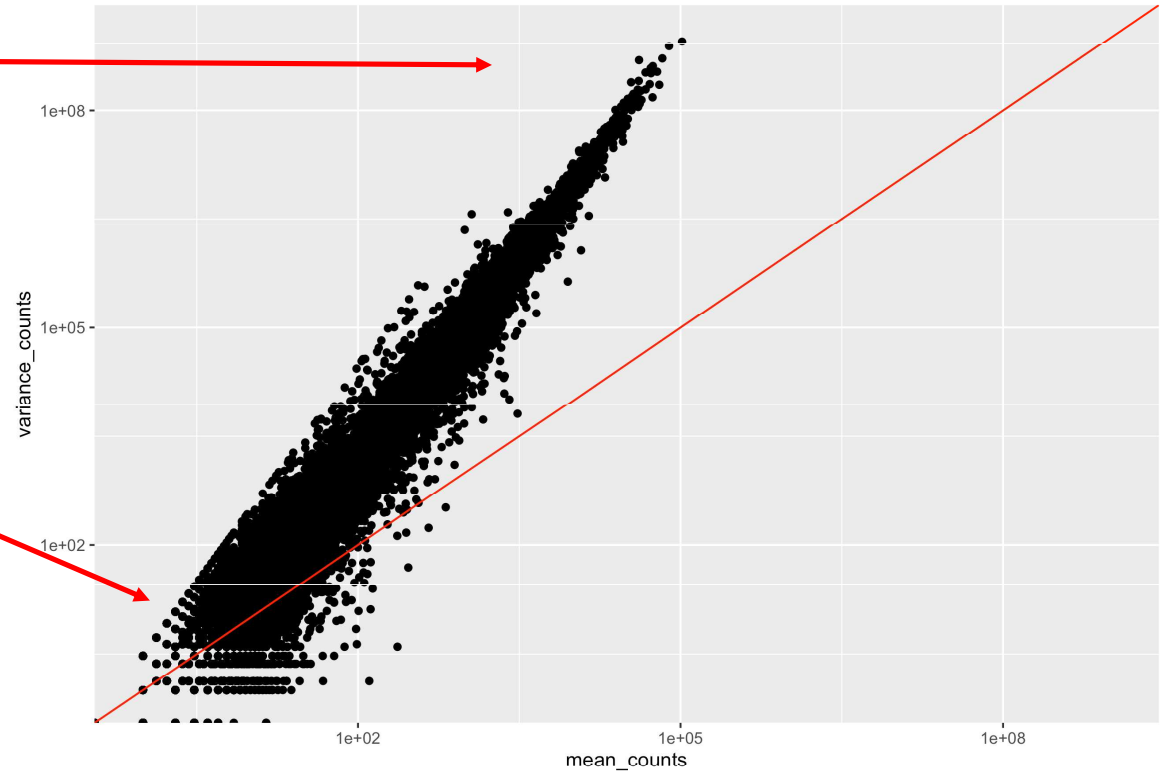
- Variabilita je zároveň více variabilní u nižších hodnot



Heteroskedasticita RNAseq dat

- Geny s vyšší expresí mají mnohem vyšší variabilitu

- Variabilita je zároveň více variabilní u nižších hodnot

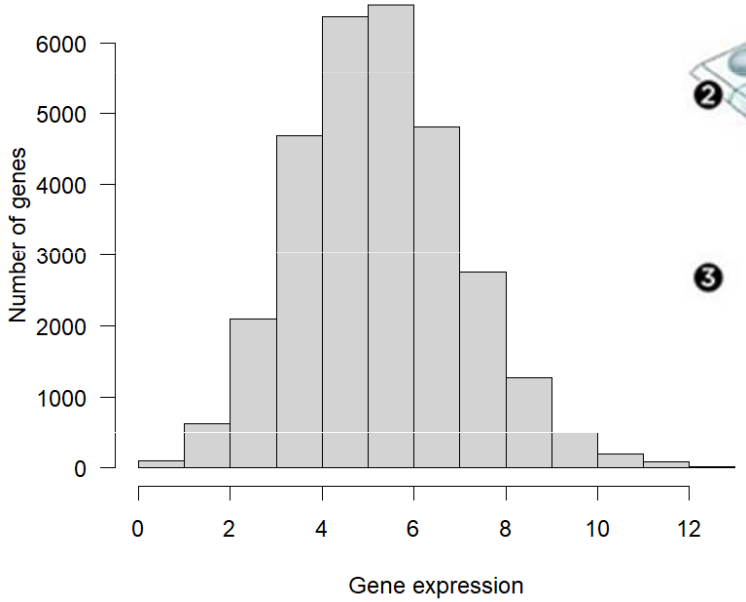


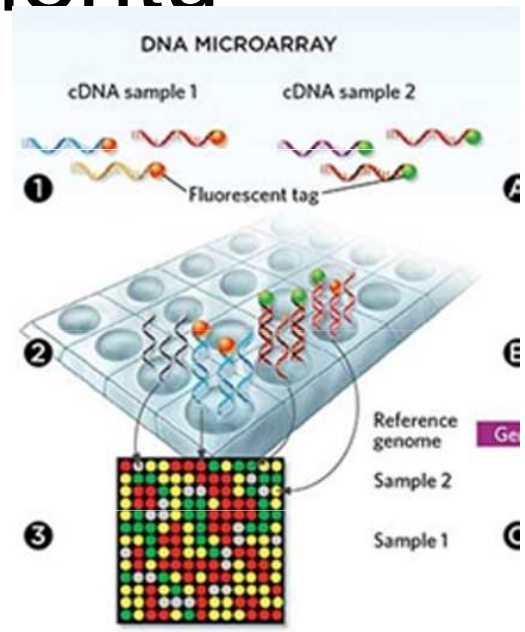
Pokud by bylo rozložení Poissonovo, tak by platilo, že variabilita se rovná průměru (diagonála)….

S malým N to neplatí …. je to teda spíše **Negativní binomiální** rozložení

Příklad dat z microarray experimentu

Data exprese genu jsou vyjádřené jako intensity pixelů od 0 do maxima **65,535 – toto maximum nesdílí s jinými geny**.

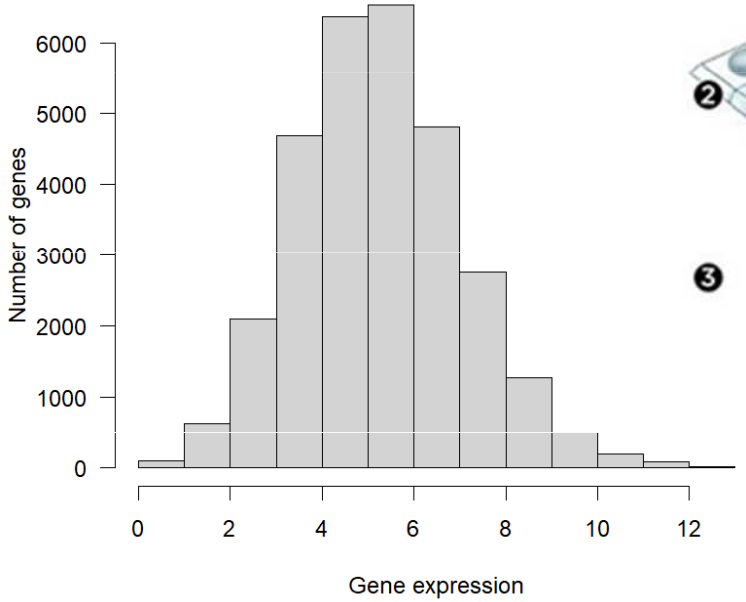


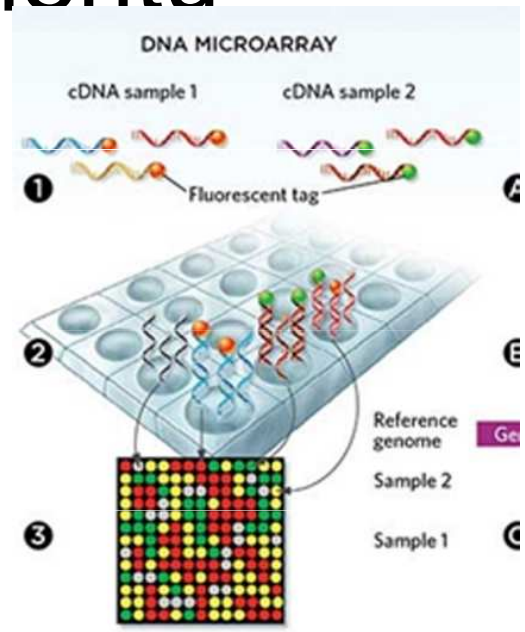


Příklad dat z microarray experimentu

Data exprese genu jsou vyjádřené jako intensity pixelů od 0 do maxima **65,535 – toto maximum nesdílí s jinými geny**.

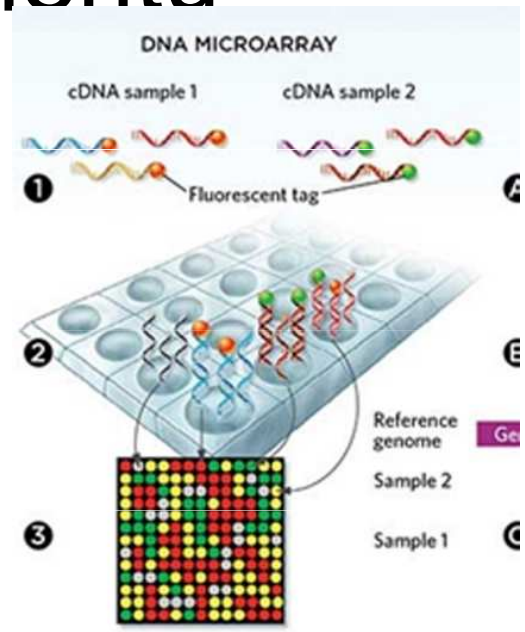
Není zde tolik nízce exprimovaných genů, distribuce je mírné posunutá



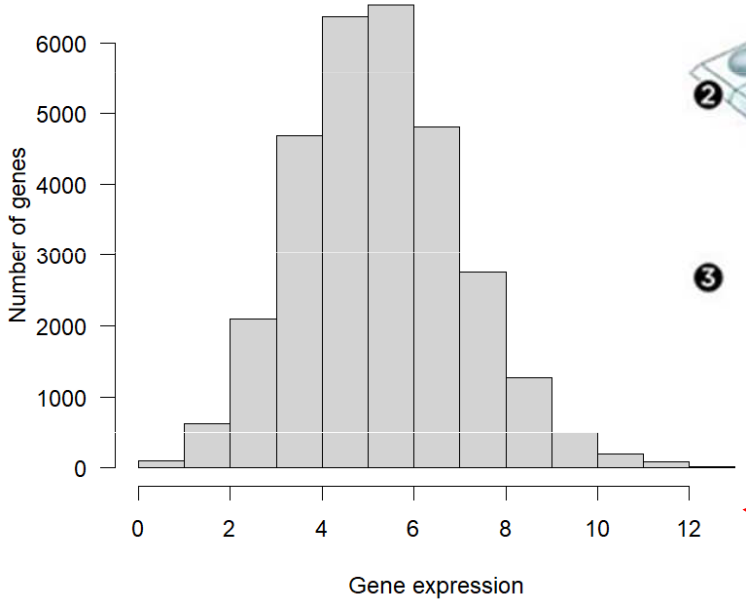


Příklad dat z microarray experimentu

Data exprese genu jsou vyjádřené jako intensity pixelů od 0 do maxima **65,535 – toto maximum nesdílí s jinými geny**.



Není zde tolik nízce exprimovaných genů, distribuce je mírné posunutá



PROČ:

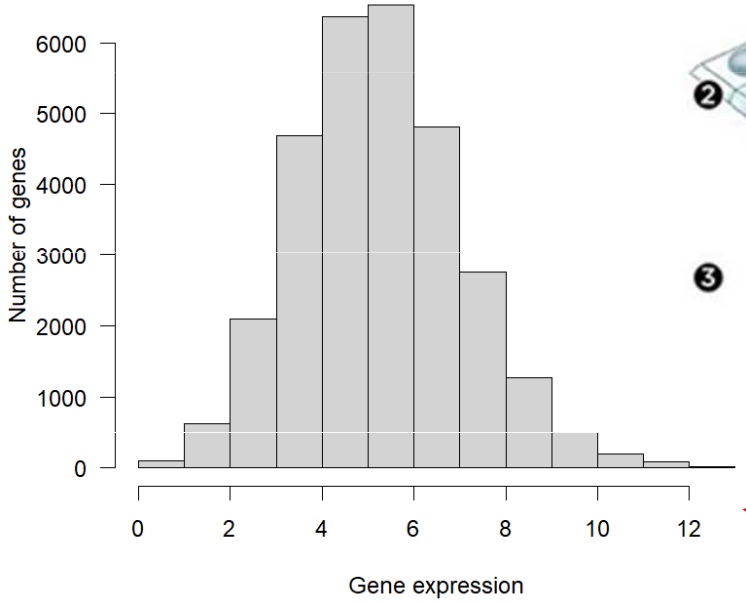
- Měření jsou mezi sebou nezávislá, I málo exprimované geny mají šanci, protože mají sondu, která je "vychytá"

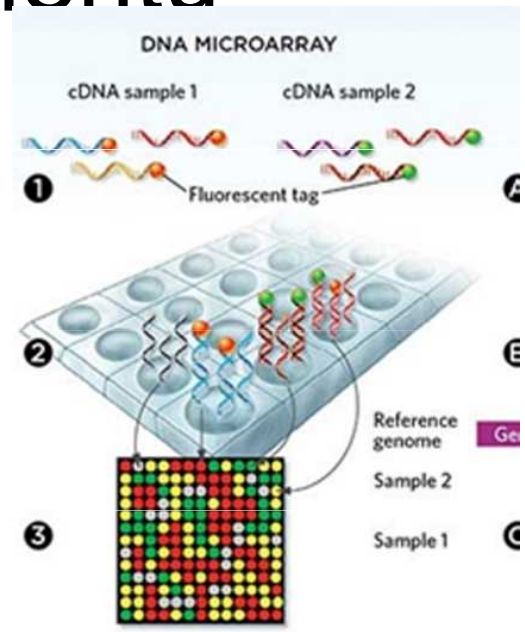
Díky maximu nemá tak dlouhý ocas

Příklad dat z microarray experimentu

Data exprese genu jsou vyjádřené jako intensity pixelů od 0 do maxima **65,535 – toto maximum nesdílí s jinými geny**.

Data mají spíše **normální** rozložení



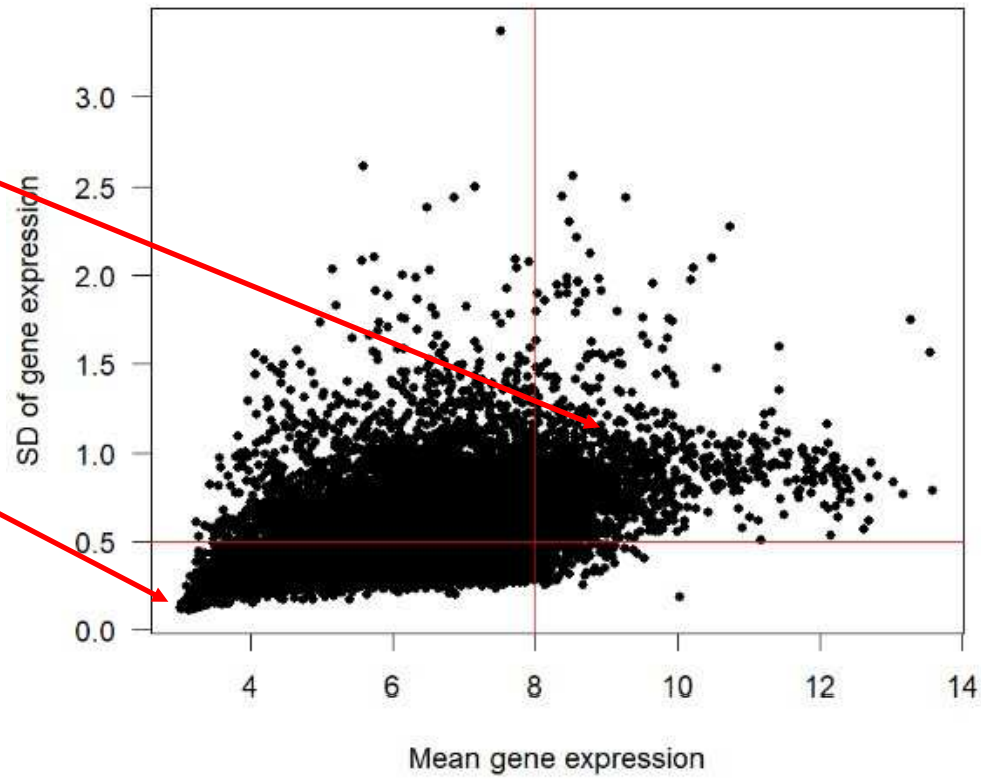


- Měření jsou mezi sebou nezávislá, I málo exprimované geny mají šanci, protože mají sondu, která je "vychytá"

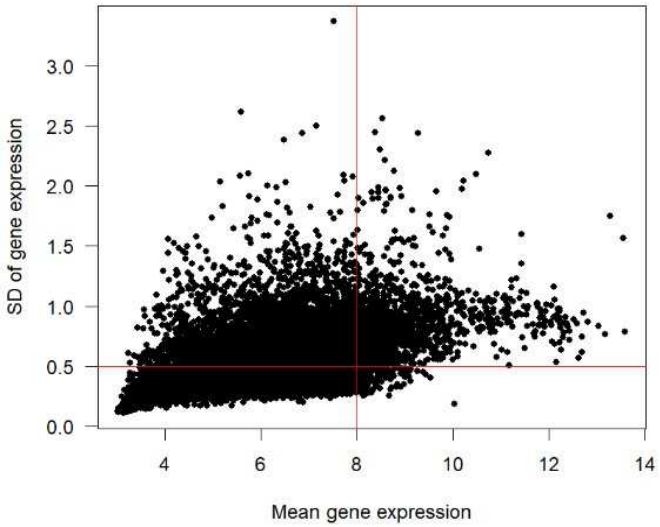
Díky maximu nemá tak dlouhý ocas

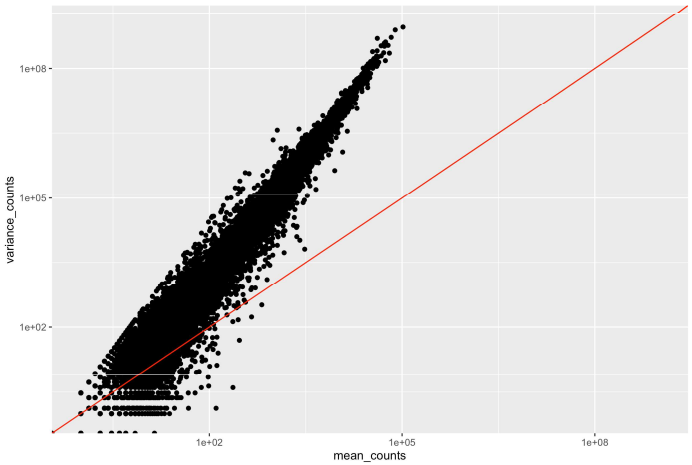
Heteroskedasticita microarray dat

- Geny s vyšší expresí mají mnohem vyšší variabilitu
- Variabilita je zároveň méně variabilní u nižších hodnot



Heteroskedasticita - důsledky





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

Aby se daly statistiky porovnat, je potřeba sjednotit variabilitu a hlavně správně modelovat data.

Co s tím?

- Znormalizujme variabilitu před testováním - například s pomocí kvantilové normalizace
- Upravíme samotnou statistiku
- Nebo obojí

Jednoduchá korekce konstantou

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 nějak sjednotit variabilitu:

Konstanta korigující variabilitu (zvyšuje variabilitu pokud je nízká, u vysoké dohromady nic neudělá)

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

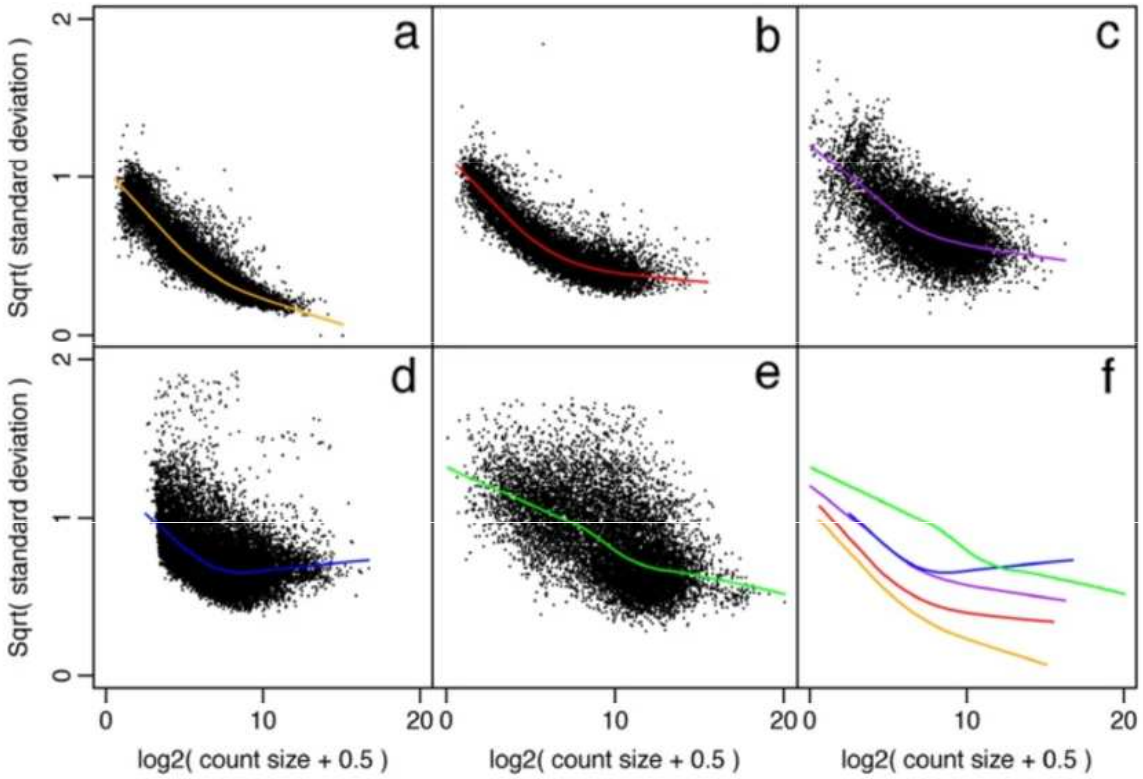
  Nevýhody: výpočetně náročné (permutace)

  Výstup: *q*-hodnoty

```
library(samr)
```

Odhad konstanty pro korekci pro každý gen zvlášť

Figure 1



[voom: precision weights unlock linear model analysis tools for RNA-seq read counts | Genome Biology | Full Text]

Limma (+ voom)

- 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
- Voom se používá u RNAseq dat – je to krok korigující variabilitu s pomocí loess, u microarray jsou data takto již upravena

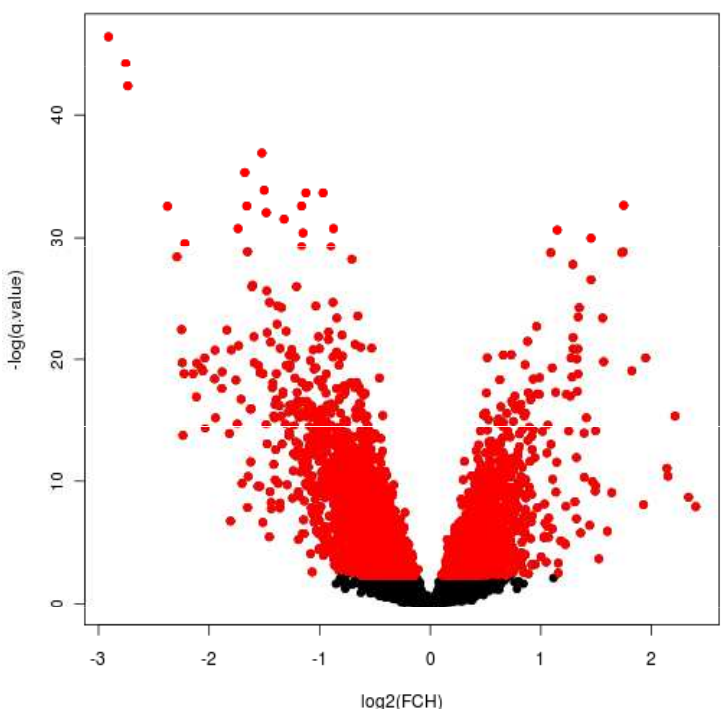
DESeq2

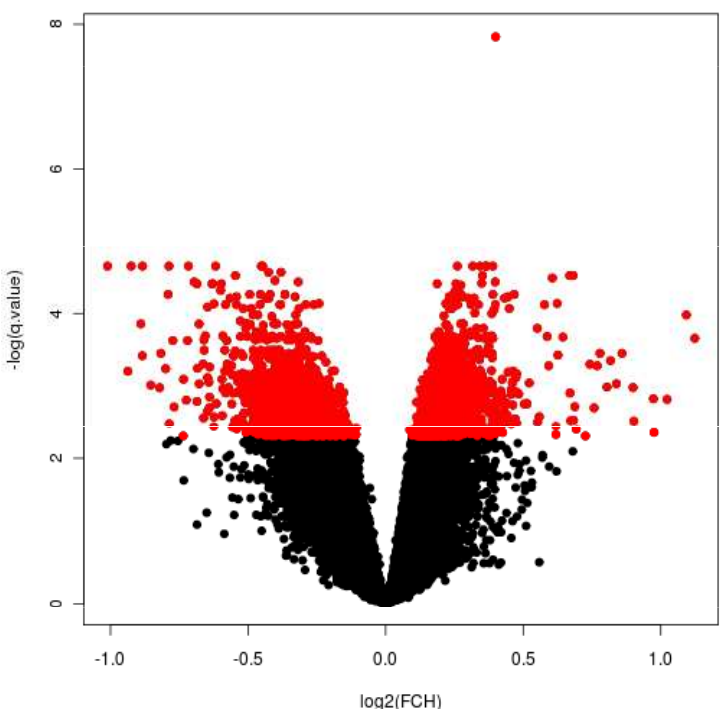
- Love et al., 2014
- DeSeq – metoda pracuje s daty RNAseq, neprovádí převod na kontinuální škálu
- Pracuje s daty jako s negativním binomickým rozdělením
- Disperze je odhadována pro každý gen a následně je modelována jako funkce průměrné exprese genu
- K této disperzi se aplikuje **empirický Bayesovský přístup**, který „stáhne“ individuální odhady disperze směrem k hladkému modelu (sdílené disperzi). Tento krok je podobný Bayesovské shrinkage v limmě.
- Pro výpočet rozdílů používa Waldovy testy

Metoda	DESeq	Limma-voom	SAM	edgeR	DEXSeq
Typ dat	RNA-seq (počty čtení)	RNA-seq (log-CPM)	Mikroarray (intenzity)	RNA-seq (počty čtení)	RNA-seq (počty čtení, exony)
Model variability	Negativní binomické rozdělení	Lineární model s vážením	Permutační T-statistics	Negativní binomické rozdělení	Negativní binomické rozdělení
Normalizace	Size factors	TMM nebo CPM	Kvantilová nebo loess	TMM	Size factors
Disperze/variabilita	Modelovaná jako funkce průměru	Vážený mean-variance vztah	Implicitně v T-statistics	Modelovaná jako funkce průměru	Modelovaná jako funkce průměru
Testování hypotéz	Waldovy/přesné testy	Moderované t-statistics	T-statistics s FDR	Quasi-likelihood/Fisherovy testy	Generalizovaný lineární model
Hlavní výhoda	Přesné modelování počítacích dat	Rychlost a flexibilita	Citlivost na slabě exprimované geny	Rychlé i pro velké datasety	Analýza alternativního sestřihu
Výpočetní náročnost	Střední až vysoká	Nízká až střední	Nízká	Nízká až střední	Vysoká
Kdy použít	Malé datasety, přesná variabilita	Velké datasety, rychlá interpretace	Mikroarray data, málo vzorků	Univerzální pro RNA-seq	Alternativní sestřih/exony

Typické zobrazení významnosti genů
Volcano plot

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





Volcano plot