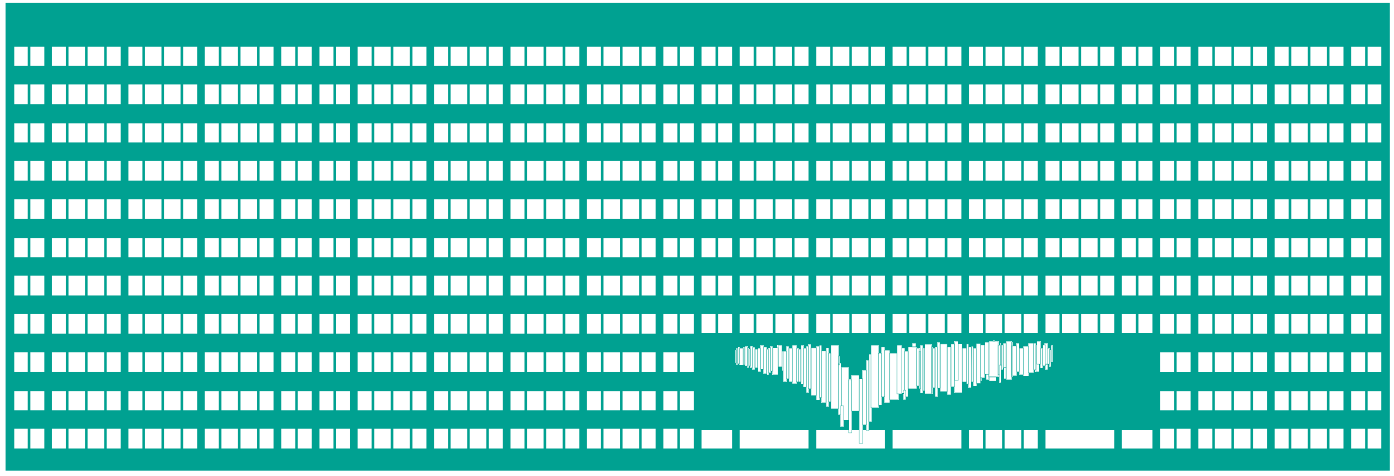
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Algoritmy pro Bioinformatiku

Exprese genů

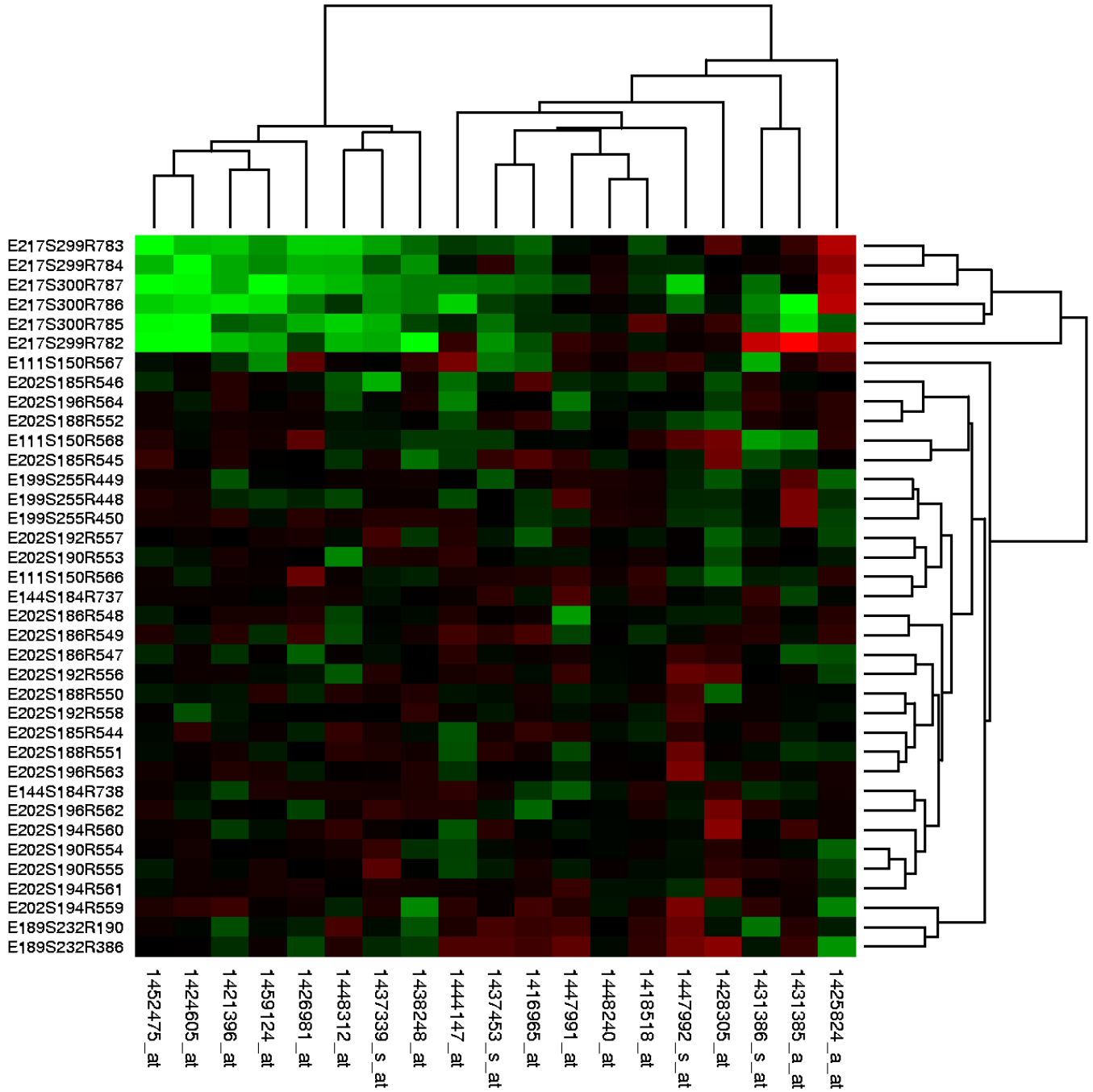
Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

20. listopadu 2019


E217S299R783
E217S299R784
E217S300R787
E217S300R786
E217S300R785
E217S299R782
E111S150R567
E202S185R546
E202S196R564
E202S188R552
E111S150R568
E202S185R545
E199S255R449
E199S255R448
E199S255R450
E202S192R557
E202S190R553
E111S150R566
E144S184R737
E202S186R548
E202S186R549
E202S186R547
E202S192R556
E202S188R550
E202S192R558
E202S185R544
E202S188R551
E202S196R563
E144S184R738
E202S196R562
E202S194R560
E202S190R554
E202S190R555
E202S194R561
E202S194R559
E189S232R190
E189S232R386
1452475_at
1424605_at
1421396_at
1459124_at
1426981_at
1448312_at
1437339_s_at
1438248_at
1444147_at
1437453_s_at
1416965_at
1447991_at
1448240_at
1418518_at
1447992_s_at
1428305_at
1431386_s_at
1431385_a_at
1425824_a_at

Přehled přednášky

- Co to je exprese genů.
- Postup při analýze.
- Ukázka v R.

Základní pojmy

- Genom - množina genů jednoho organismu.
- Transcriptom - mRNA v buňce v jednom konkrétním okamžiku.
- Proteom - všechny polypeptidy v bunče v jednom konkrétním okamžiku.
- Interactom - množina všech protein-protein interakcí v buňce v jednom konkrétním okamžiku.

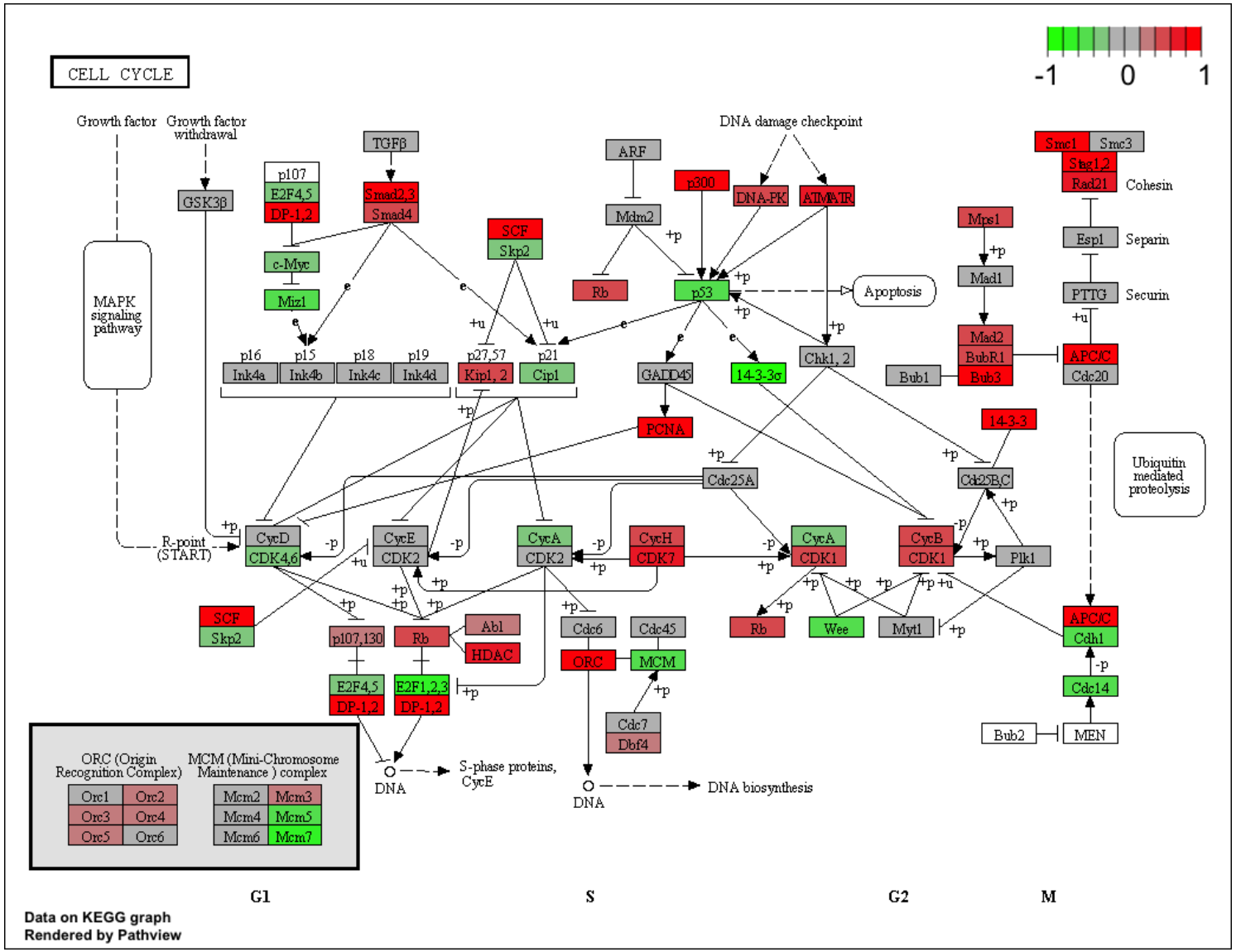
Exprese genů

- Proces během kterého je informace uložená v DNA převedna na buněčnou funkci nebo strukturu formou přepisu DNA do proteinu.
- Exprese genu regulována na mnoha úrovních:
  - Navázání RNA polymerázy ovlivněno přítomností represorů a aktivátorů - váží se poblíž promotérů a ovlivňují schopnost polymerázy zahájit transkripci.
  - Regulace aktivátorů a represorů pomocí koregulátorů - koregulátory jsou přenašeči signálu, který způsobí navázání nebo odpojení regulačních proteinů.
  - Mutace v oblasti promotérů mohou ovlivnit schopnost polymerázy nasednout na DNA.
  - MicroRNA naváže se na mRNA zabrání translaci do proteinu.

Dva pohledy zkoumání

- Které geny jsou aktivní při běžných buněčných procesech?
- Jaký je rozdíl aktivity genů mezi dvěmi skupinami vzorků (např. pacienti vs. zdraví jedinci).

Výstup - ukázka



RNA sekvenování

RNA sekvenování využívá standardních nástrojů sekvenování nové generace (Next-Generation Sequencing - NGS).

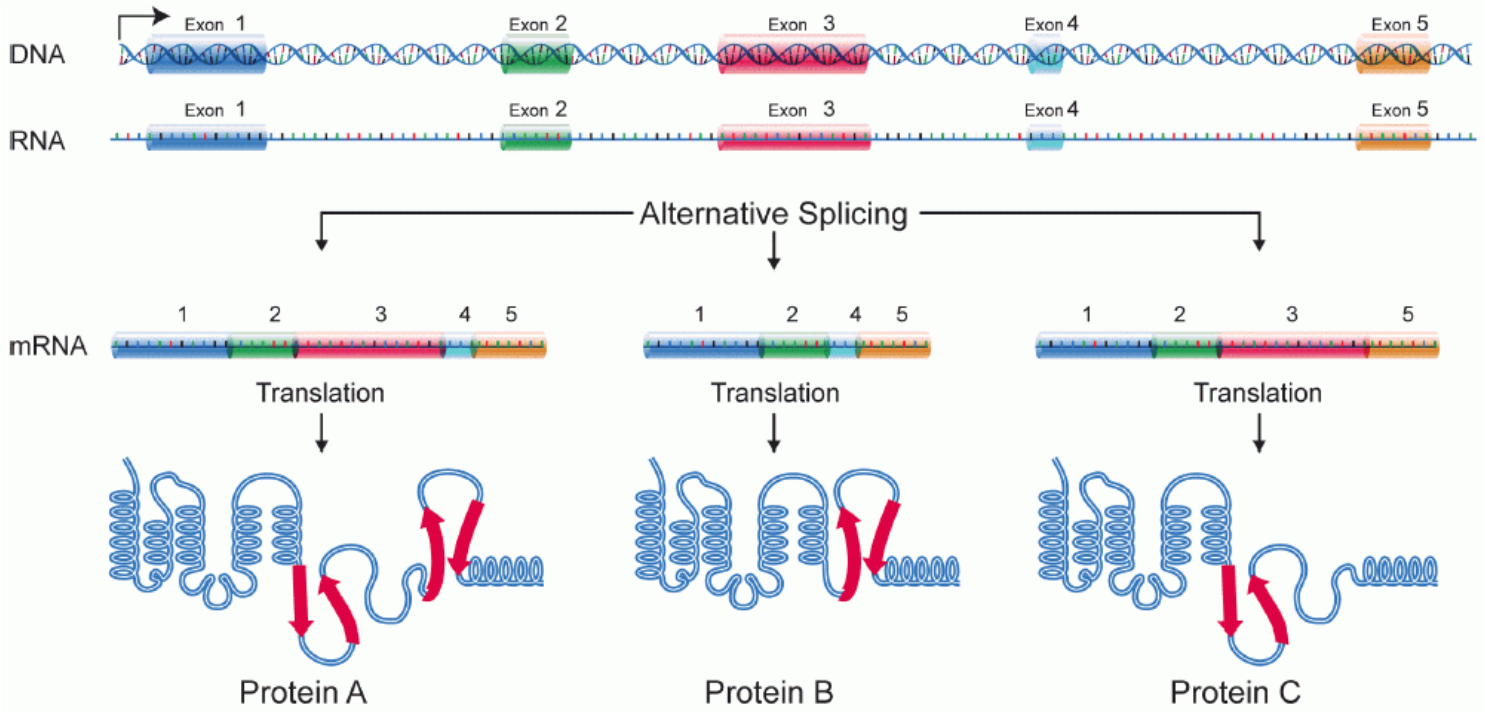
- Vzorek RNA je odebrán ze zkoumané tkáně, buňky, skupiny buněk.
- V daném okamžiku je množství RNA závislé na stavu buňky (je vystavena útoku, stresu, změně okolních podmínek, reakce na signály).
- Některá RNA se může vyskytovat více, tvorba jiné naopak potlačena.
- Chemickými reakcemi se RNA molekula reverzně přepíše na cDNA (chybí úseky DNA, které se nepřepisují na RNA).
- cDNA lze osekvenovat standardními nástroji NGS.

RNA sekvenování

- RNA-seq technologie umožňuje detekovat transkripty s velmi nízkou i velmi vysokou úrovní exprese.
- Experimenty jsou reprodukovatelné - mají menší variabilitu způsobenou způsobem měření.
- Technologie RNA-seq není limitována na organismy s popsaným genomem.
- RNA-seq kombinuje znalost o mutacích v přepisovaných oblastech DNA s množstvím exprimovaných genů (prekurzorů proteinů), díky tomu technolgie umožňuje nalezení dosud neobjevených genů, či úseků, které se přepisují do nějakého typu RNA.

RNA-seq aplikace

- Rozdíl v expresi genů mezi různými vzorky.
- Studium alternativních spojení exonů při transkripci z DNA => vede k jiným proteinům.

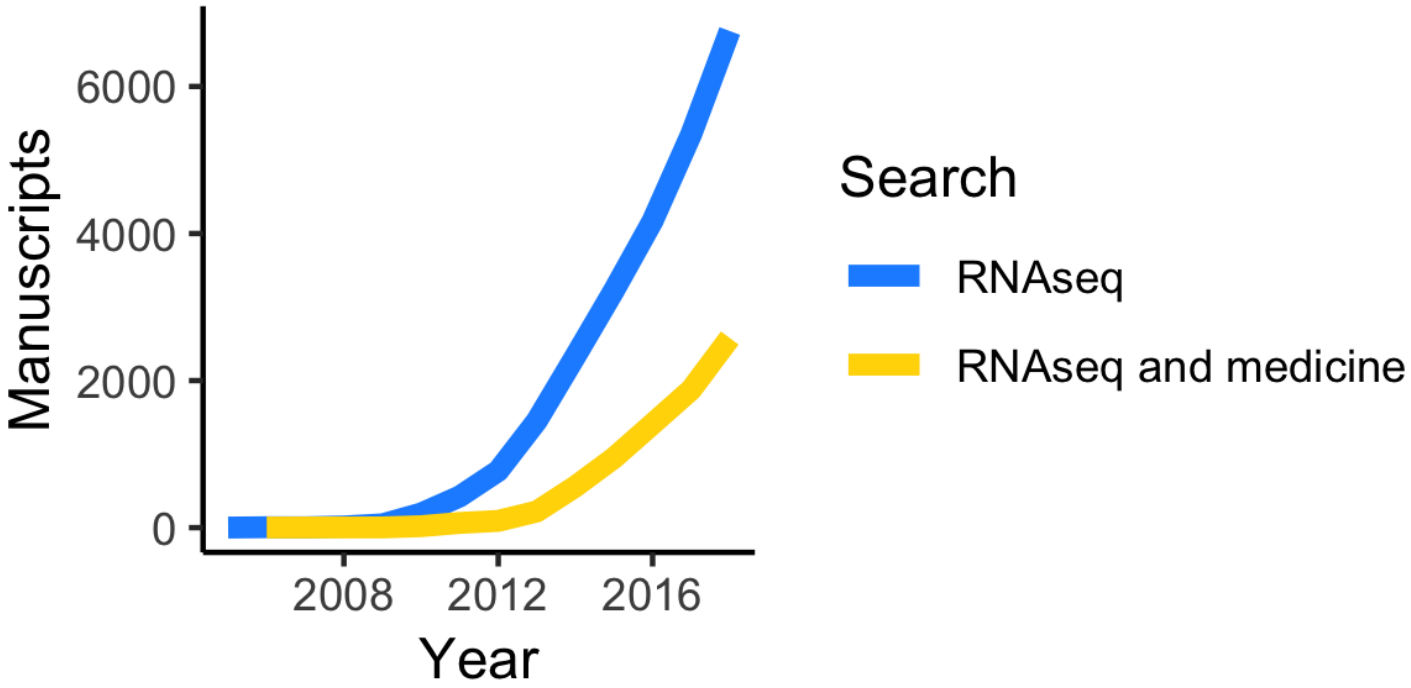


RNA-seq aplikace

- Identifikace exprese specifické pro konkrétní alelu u bodových záměn, charakteristika rakovinných procesů.
- Další aplikací: RNA-seq jedné molekuly - nová oblast zkoumání komplexních biologických procesů zaměřených na chování jedné individuální buňky.
  - Výzkum kmenových buněk - kmenová buňka se diferencuje ve specifický typ buňky, jak je proces diferenciace řízen, které geny (proteiny) jsou exprimovány, které potlačeny.

Historie RNA-seq

- Vyvinuto v 2005.
- Exponenciálně rostoucí počet vědeckých výstupů.

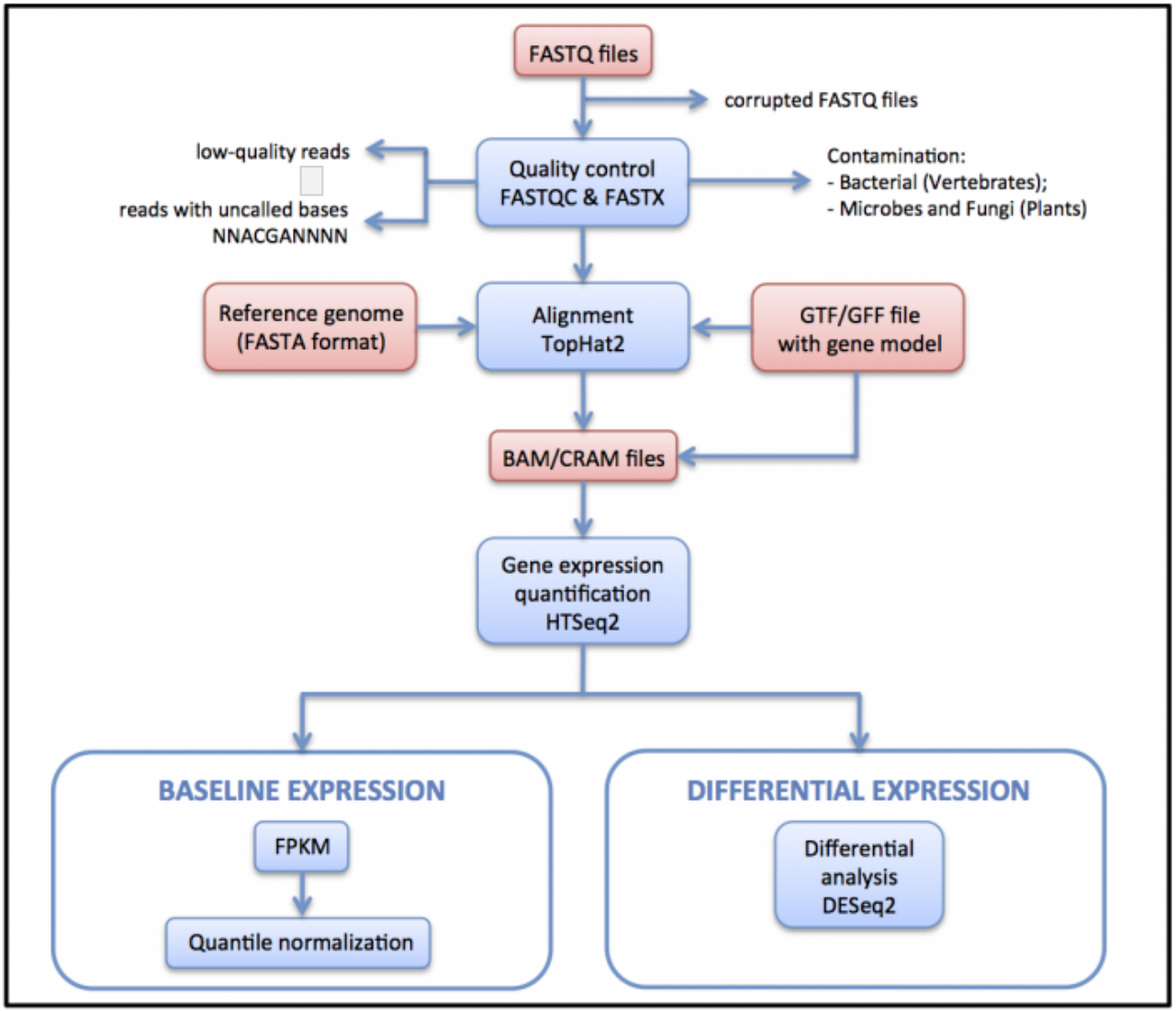


Zpracování RNA-seq dat

Ve chvíli, kdy je vzorek cDNA osekvenován, obdržíme k analýze standrdní FASTQ soubory. Konkrétní metody se mohou lišit pro různé experimenty, ale v základu obsahují následující kroky:

- Kontrola kvality sekvencí ve FASTQ souborech
- Přiložení sekvencí k referenční sekvenci, alternativně de-novo sestavení cDNA genomu.
- Kvantifikování exprese genů (výpočet konkrétních počtů sekvencí RNA překládaných na proteiny).
- Detekce diferenciálně exprimovaných genů (meziskupinové porovnání).

Zpracování RNA-seq dat

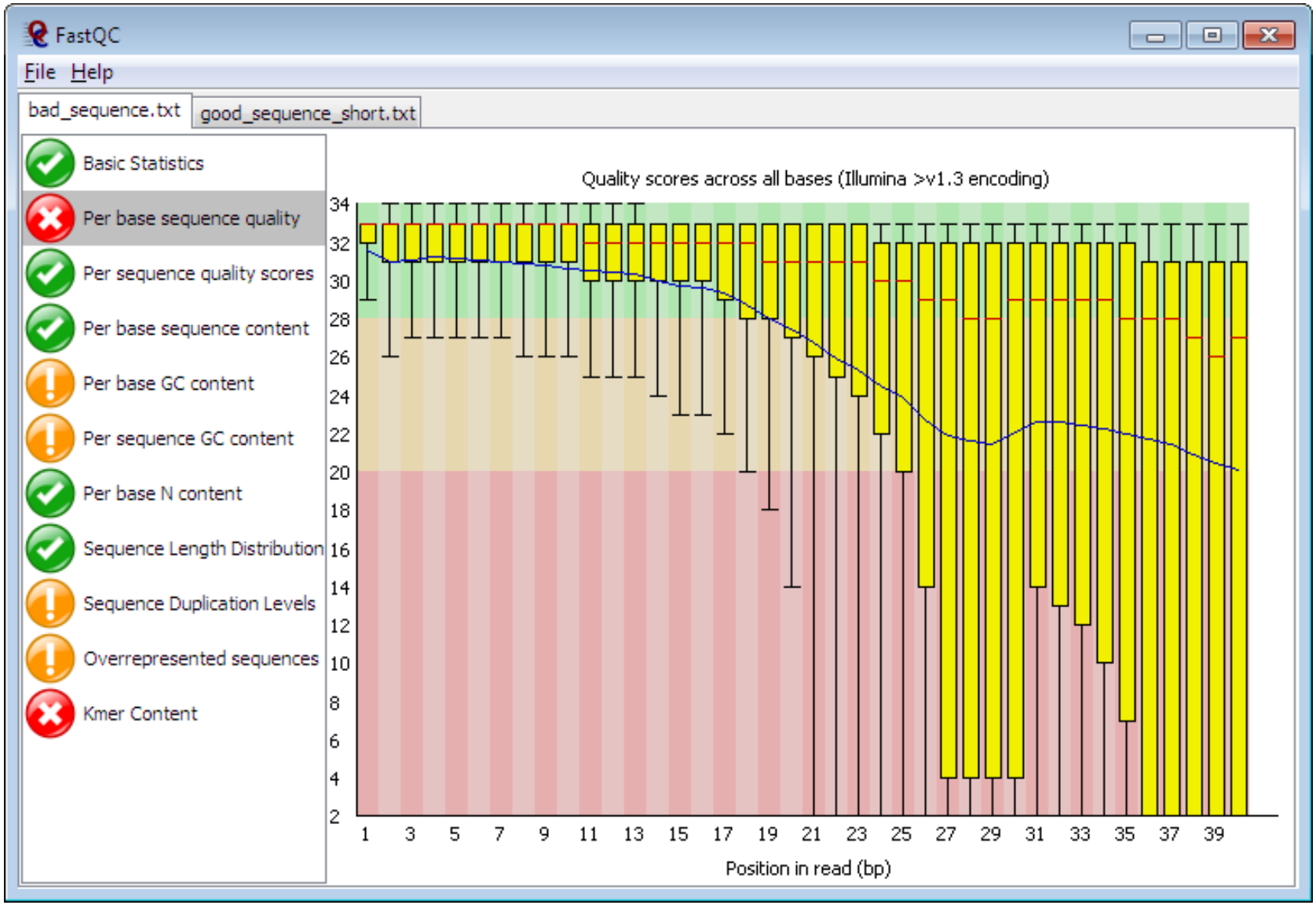


Kontrola kvality

- Odstranění adaptérových sekvencí. Adaptérové sekvence obsahují specifické sekvence nukleotidů[1].
- Vyfiltrování sekvencí s nízkou kvalitou.
- Vyfiltrování nebo přepracování sekvencí s nepřečtenými nukleotidy.
- Odfiltrování kontaminovaných sekvencí (sekvence jiných organismů, bakteriií, virů).
- Průmyslovým standardem ověření kvality je použití nástroje FastQC.
- Odstranění primerů, adaptérových sekvencí, případně odstranění nepřečtených bazí - standardem nástroj Trimmomatic.

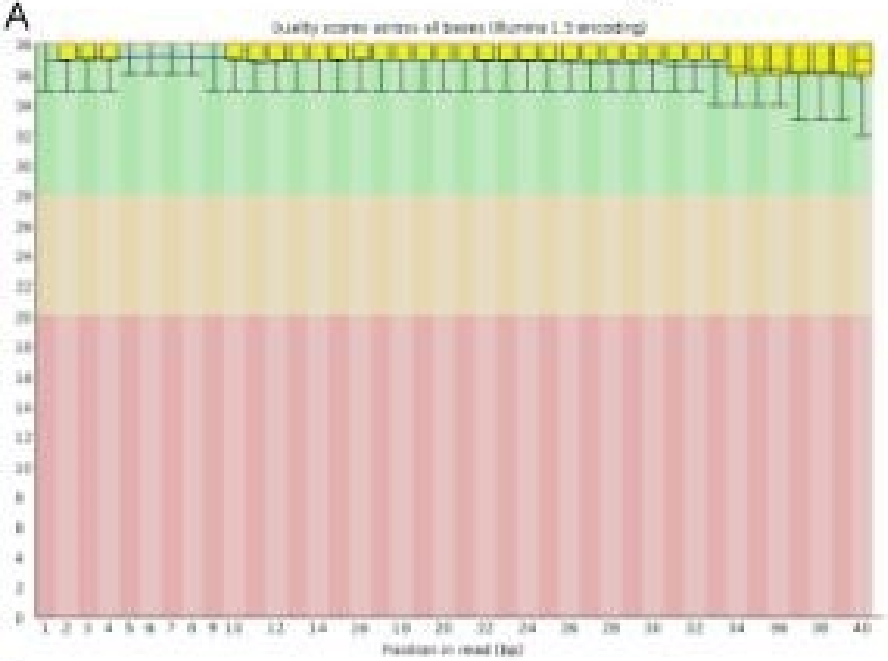
[1] `https://www.youtube.com/watch?v=fCd6B5HRaZ8`

FastQC - UI


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

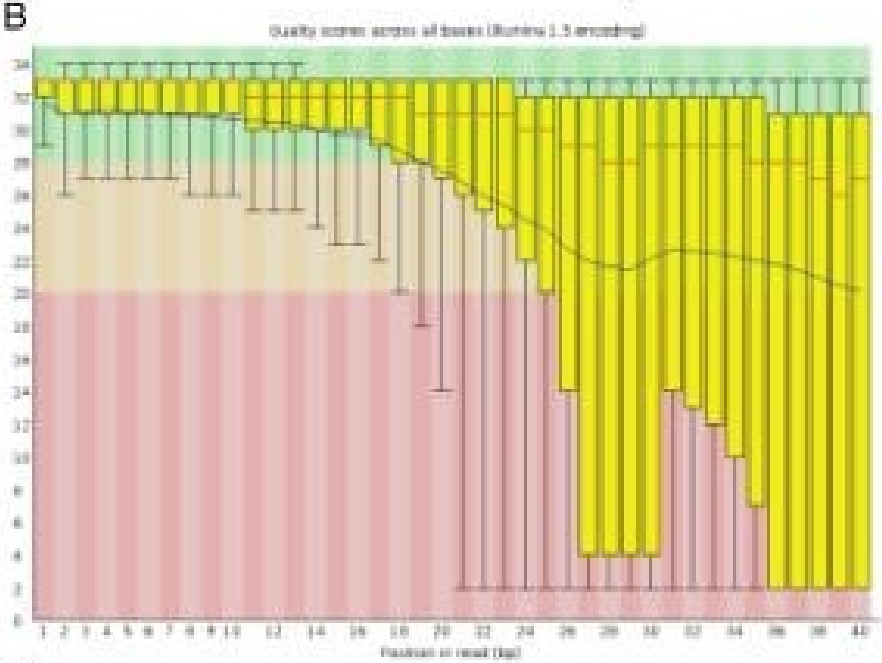
Good Quality

A



Bad Quality

B



Alignment či assembly

- Na vstupu již předpokládáme sekvence o vysoké kvalitě. Dva možné postupy v závislosti na existenci referenční sekvence:
  - Namapování na existující referenční sekvenci
    - Namapování sekvencí nevyžaduje znalost pozic genů, nebo různých variant přepisu exonů.
    - Tento přístup zároveň umožňuje nalezení nových, neoanotovaných transkriptů.
    - Nutnost řešit sekvence překrývající více exonů.
  - De-novo sestavení transkriptů.
    - Sestavení sekvencí do skupiny překrývajících se sekvencí => tzv. contigy.
    - Na tyto contigy pak původní sekvence znovu namapujeme pro určení počtu sekvencí překrývajících určitý transkript.

Alignment tools

Existuje celá řada nástrojů, které umožňují namapování krátkých sekvencí na referenční sekvenci, velmi populární TopHat, BWA, Bowtie, Hisat. TopHat například pracuje ve dvou krocích:

- Nejdříve jsou sekvence nepřekrývající více exonů namapovány na referenční sekvenci.
- Ve druhém kroku se pokoušíme o alignment dosud nenamapovaných sekvencí. Například využijeme pigeonhole principle, část sekvence se nám podaří namapovat na jeden exon, druhou část na druhý exon. Očekáváme, že se exony vyskytují relativně blízko sobě.

- Tento krok je následován kontrolou kvality namapování sekvencí (bereme v potaz především počet mutací a existence alternativních alignmentů). Používaný nástroj Picard.

Různé počty readů

Mějme dva vzorky C a P. První vzorek obsahuje 5 milionů readů a druhý 10 milionů readů.

Gen	C	P
gen1	4000	4000
gen2	4000	8000
...	...	...
Celkem	5 mil.	10 mil.

- Jsou geny 1 a 2 u vzorku P jinak exprimovány než u vzorku C.

Různé celkové počty readů u různých vzorků vedou na první úlohu o normalizaci:

Count per milion reads - CPM

Udává počet readů namapovaných na gen vztažený k jednomu milionu readů.

$$CPM(gene) = \frac{N_{gene} \times 10^6}{N} \tag{1}$$

- N_{gene} je počet readů namapovaných na konkrétní gen,
- N je celkový počet readů.

Efekt CPM

Mějme dva vzorky C a P. První vzorek obsahuje 5 milionů readů a druhý 10 milionů readů.

Gen	C	P	CPM(C)	CPM(P)
gen1	4000	4000	800	400
gen2	4000	8000	800	800
...	...	...	...	...
Celkem	5 mil.	10 mil.	1mil.	1 mil.

Normalizace - problémy

Různé délky genů

Mějme ve vzorku 5 milionů readů a např gen1 s počtem pokrývajících readů $N_{gen1} = 1000$ a délkou $|gen| = 2000$bp.

Gen	N_{gen}	$\|gen\|$
gen1	1000	2000
gen2	1000	4000
gen3	2000	4000

Tabulka: Table caption

- Dva geny se stejným pokrytí ready.
- Určení, který gen je exprimován více.

Řešení různých délek genů

Různé délky genů vedou k potřebě normalizace počtu sekvencí:

RPKM - reads per kilo base of transcript per million mapped reads

RPKM udává počet readů na kilobázi transkriptu (přepsaného genu) v miliónu namapovaných readů.

$$RPKM(gen) = \frac{CPM(gen) \times 10^3}{|gen|}$$

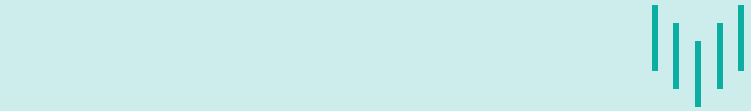
Normalizace není zapotřebí pro vyhodnocení exprese mezi různými skupinami vzorků, nicméně je nezbytná pro vzájemné porovnání exprese genů v jednom vzorku.

Normalizace - problémy

Mějme ve vzorku 5 milionů readů a např gen1 s počtem pokrývajících readů $N_{gen1} = 1000$ a délkou $|gen_1| = 2000$bp.

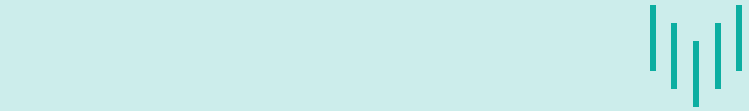
Gen	N_{gen}	$\|gen\|$	RPKM
gen1	1000	2000	100
gen2	1000	4000	50
gen3	2000	4000	100

Diferenciální analýza exprese genů



- Diferenciální analýza má na vstupu normalizované počty readů.
- Předpokládá alespoň dvě skupiny vzorků.
- Diferenciální analýza se snaží pomocí statistických metod prokázat změny v expresi genů oproti kontrolní skupině vzorků.
- Statistické testování slouží k ohodnocení významnosti pozorované změny, jinými slovy zda-li je rozdíl větší, než jaké hodnoty bychom očekávali na základě přirozené náhodné variability.

Metody analýzy diferenciální exprese

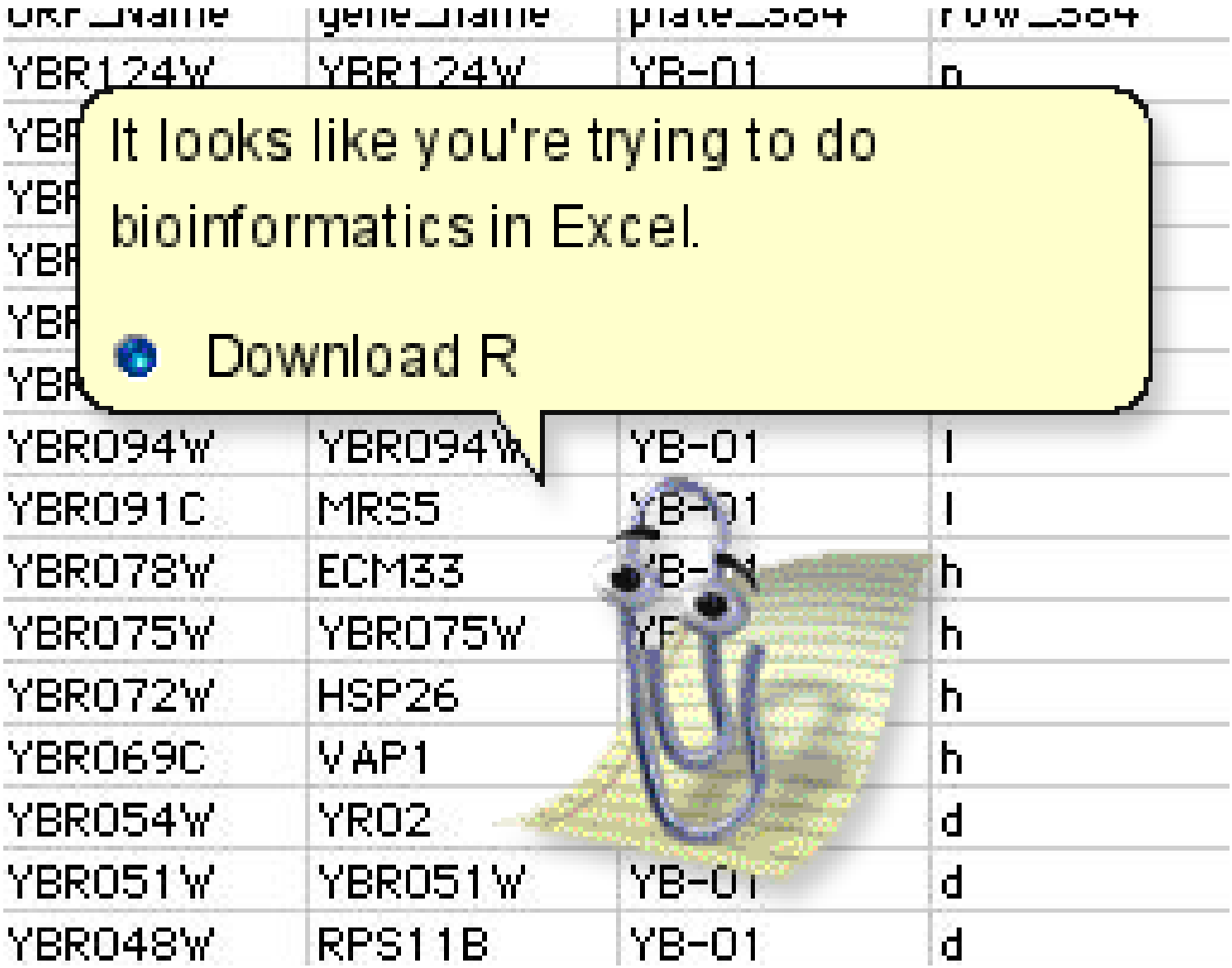


- Vytvořena celá řada nástrojů.
- limma - moderated t test.
- edgeR a DESeq založeny na modelování pomocí negativního binomického rozdělení.
- baySeq a EBSeq založeny na Bayesovské statistice.
- Samotný experiment obvykle předurčí jednu z metod, některé metody dokáží provést diferenciální analýzu pouze mezi dvojicí skupin. Nástroje jako edgeR, limma-voom, DeSeq či maSigPro dokáží testovat rozdílnost více skupin.

Biologická interpretace dat z genové exprese

"Data don't make any sense,
we will have to resort to statistics."

Biologická interpretace dat z genové exprese



Normální rozdělení

- Náhodná proměnná z z normálního rozdělení má střední hodnotu μ a roztyl σ^2.
- Říkáme, že z má rozdělení $N(\mu, \sigma^2)$.
- Hustota pravděpodobnosti je dána:

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

χ^2 rozdělení

Definition

Nechť $z_1, z_2, \ldots, z_n$ je náhodný výběr z normálního rozdělení $N(0,1)$. Náhodný vektor $z = (z_1, z_2, \ldots, z_n)$ má rozdělení $N_n(0,1)$, pak dle definice:

$$\sum_{i=1}^{n} z_i^2 = z'z \text{ má rozdělení } \chi^2(n)$$

- Součet druhých mocnin n nezávislých normálně rozdělených náhodných proměnných má χ^2 rozdělení s n stupni volnosti.

Definition

Nechť z je $N(0,1)$, u je $\chi^2(p)$ a z a u jsou nezávislé náhodné proměnné, pak je:

$$t = \frac{z}{\sqrt{u/p}} \text{ má rozdělení } t(p)$$

tzv. t rozdělení s p stupni volnosti.

t test o shodě průměrů

- Mějme dvě skupiny vzorků, uvažujme, že tyto dvě skupiny se liší nějakým rysem, který nás zajímá.
- Pro obě skupiny vzorků jsme provedli sekvenování RNA.
- Chceme zjistit, které geny jsou charakteristické pro zkoumanou odlišnost,
- Předpokládejme, že rozdělení počtu sekvencí je normální náhodná proměnná a pro obě skupiny platí, že mají stejný rozptyl.
- Můžeme použít t test o shodě průměrů náhodných veličin se stejným rozptylem.

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

t test o shodě průměrů

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

- n_x, n_y jsou počty vzorků v jednotlivých skupinách.
- Nulová hypotéza: oba průměry jsou shodné, tedy $\overline{x} - \overline{y} = 0$.
- Statistika t má rozdělení $T(n_x + n_y - 2)$.
- Odhad rozptylu

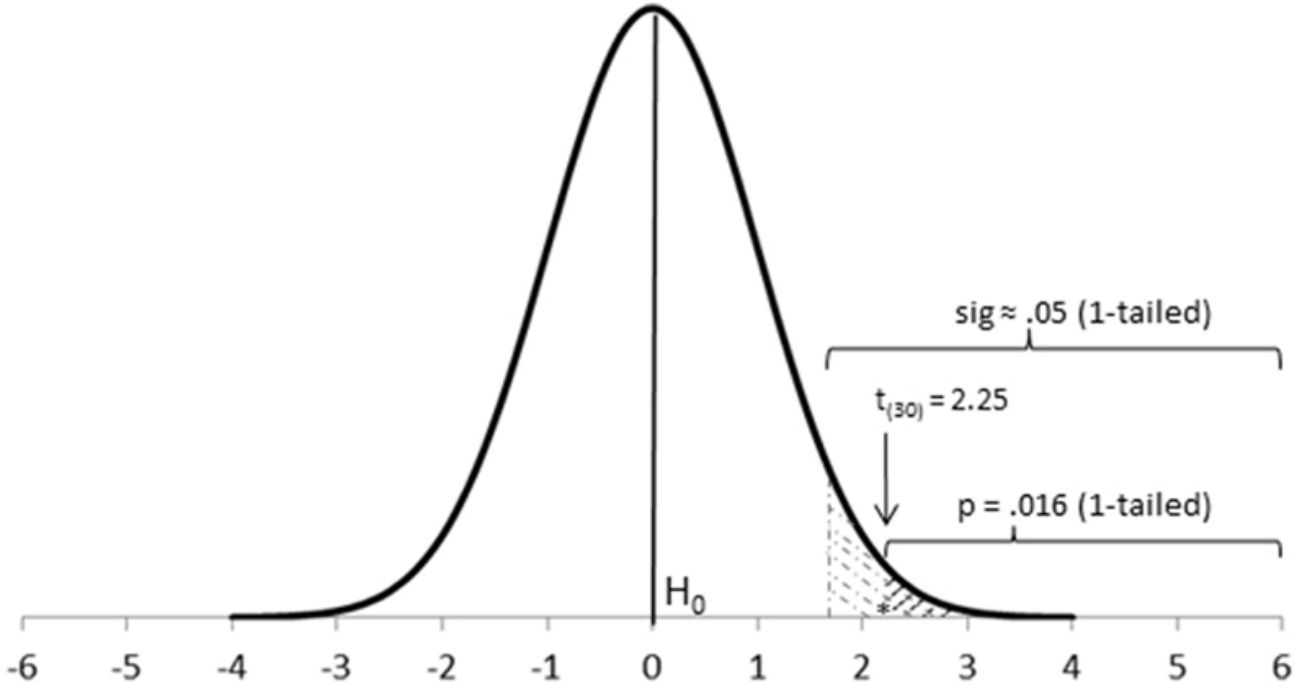
$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1 s_y^2}{(n_x + n_y - 2)}$$

t test o shodě průměrů

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

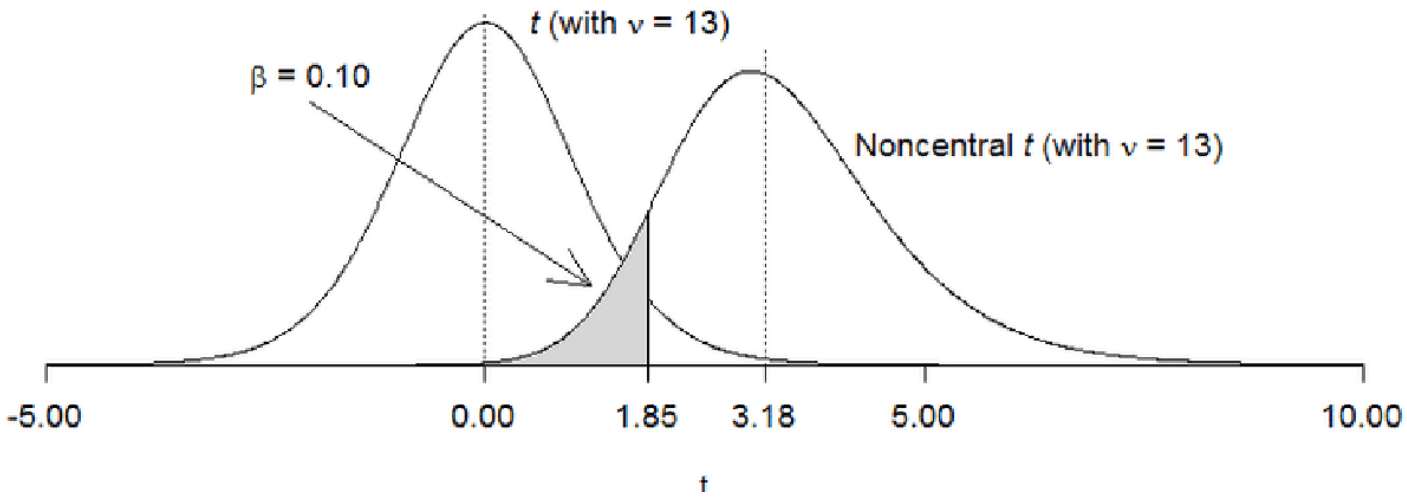
- Pro vypočtenou hodnotu t spočítáme pravděpodobnost, že hodnota je větší nebo rovna t tzv. p-hodnota (p-value).
- Podle předem zvolené hladiny významnosti α porovnáme zda je $p(X > t) < \alpha$, pokud ano, pak zamítáme nulovou hypotézu.
- Hladina významnosti nám udává chybu, že jsme nesprávně zamítli nulovou hypotézu.

t test o shodě průměrů



- Pro vypočtenou hodnotu t spočítáme pravděpodobnost, že hodnota je větší nebo rovna t tzv. p-hodnota (p-value).
- Podle předem zvolené hladiny významnosti α porovnáme zda je $p(X > t) < \alpha$, pokud ano, pak zamítáme nulovou hypotézu.
- Hladina významnosti nám udává chybu, že jsme nesprávně zamítli nulovou hypotézu.

t test o shodě průměrů



- Nulová hypotéza má centrální t rozdělení.
- Alternativní hypotéza má necentrální t rozdělení.
- Rozdělení alternativní hypotézy neznáme, máme pouze odhady.

t test o shodě průměrů

- Normální rozdělení předpokládá spojité hodnoty počtu readů.
- Geny s nižší expresí mají větší roztyl a mají tendenci být zešikmené.
- Pro geny s nižší expresí neplatí předpoklad normálního rozdělení.
- Doporučuje se aplikovat $\log_2$ transformaci počtu readů před vyhodnocením významnosti.

Jednoduchá lineární regrese

Předpokládejme, že následující model popisuje nějakou modelovou lineární závislost pro n pozorování.:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y a ϵ jsou náhodné proměnné a předpokládejme, že x je vždy nějaká konstanta.
- $E(\epsilon) = 0$ jinými slovy $E(y) = \beta_0 + \beta_1 x$.
- $var(\epsilon) = var(y) = \sigma^2$

Jednoduchá lineární regrese

Předpokládejme, že máme náhodný vzorek n pozorování $y_1, y_2, \ldots, y_n$, Neznáme β_0 ani β_1 a potřebujeme získat nějaký jejich odhad $\hat{\beta}_0$, $\hat{\beta}_1$ a pomocí těchto odhadů vytvoříme predikci $\hat{y}$ hodnoty y pro nějakou hodnotu x.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\beta}_0$ a $\hat{\beta}_1$ získáme metodou nejmenších čtverců - minimalizací rozdílu mezi skutečnou hodnotou y a predikovanou $\hat{y}_i$.

$$\hat{\epsilon}\hat{\epsilon} = \sum_i \hat{\epsilon_i}^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- $y_i - \hat{y}_i$ nazýváme residuum y, nebo také chyba odhadu.
- $(y_i - \hat{y}_i)^2 = SSE$ nazýváme residuum součtu čtverců, nebo také chybu odhadu součtu čtverců.

Odhady $\hat{\beta}_0$ a $\hat{\beta}_1$

Derivací podle $\hat{\beta}_0$ a $\hat{\beta}_1$ a položením derivací rovno 0, nalezneme odhady $\hat{\beta}_0$ a $\hat{\beta}_1$, které minimalizují $\hat{\epsilon}\hat{\epsilon} = SSE$:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x_i})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Jednoduchá linární regrese t-statistika

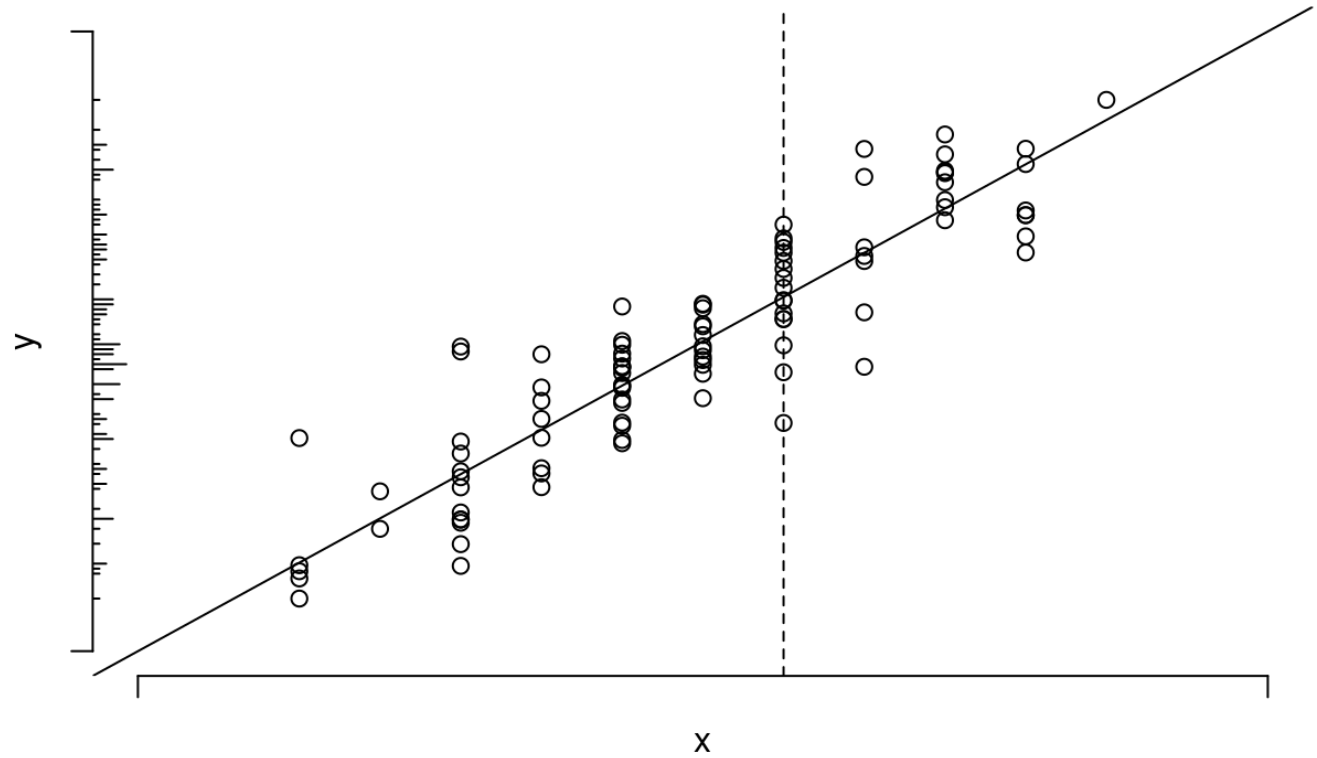
$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_i (x_i - \overline{x})^2}}$$

- Statistické testování je obvykle založeno na předpokladu, že pokud mezi dvěma proměnnými neexistuje lineární vztah, pak je $\beta_1 = 0$.
- Hypotéza H_0 je tedy $H_0 : \beta_1 = 0$.
- Odhad rozptylu:

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

- Hodnoty x_i volíme např. $x_i = 0$ pokud vzorek je z kontrolní skupiny a $x_i = 1$ pokud je ze skupiny pacientů.

Statistické testování



- Umožňuje testovat lineární závislost pouze mezi dvěmi kategoriemi, např. zdraví vs. nemocní.

Příklad výsledku

	A	B	C	D	E
1		baseMean	log2FoldChange	stat	pvalue
2	APOL1	2290.56530882017	3.10832963628912	8.61937720535332	6.73192495579598e-18
3	LYZ	77870.2427655811	-7.0510192436375	-7.74791509820844	9.34135604883871e-15
4	SLC6A8	678.327168089741	1.90544175497903	7.52749712316208	5.17221477079242e-14
5	SEMA4B	1741.57724244313	1.49340137828406	7.15320066774613	8.47775311396553e-13
6	STATH	3313.87839012742	-8.16963354067458	-7.06518543076518	1.60401654729597e-12
7	ZG16B	1487.34560622635	-5.7128835515497	-6.80027227078752	1.04421612857286e-11
8	KCNMA1	259.585606740072	-2.76284495857023	-6.79096191874476	1.11388327853096e-11
9	ATP2A3	730.994954105378	-3.69920010406926	-6.78428505072016	1.16662786413202e-11
10	EGLN3	706.580777036715	3.49475068397002	6.77969308994643	1.20431442080478e-11
11	SOX4	1503.45484443017	1.32943866871453	6.7756078105378	1.23884320257993e-11
12	PRH2	9053.73856673143	-7.84574482823239	-6.74997021728914	1.47875506823875e-11
13	MARK1	459.174938188555	0.785729633569354	6.62769799546138	3.40962140042419e-11
14	FERMT2	166.730996298386	-2.0893443692225	-6.4386924529594	1.20507094321743e-10
15	APOL2	943.627078393778	1.62708093956464	6.40488809297558	1.50479712245154e-10
16	NDUFA4L2	372.555153565341	3.77626062525379	6.35736307736515	2.05246477081785e-10
17	DMBT1	3703.80190168882	-8.62250414133728	-6.29771090292796	3.02073212034546e-10

Log 2 fold change

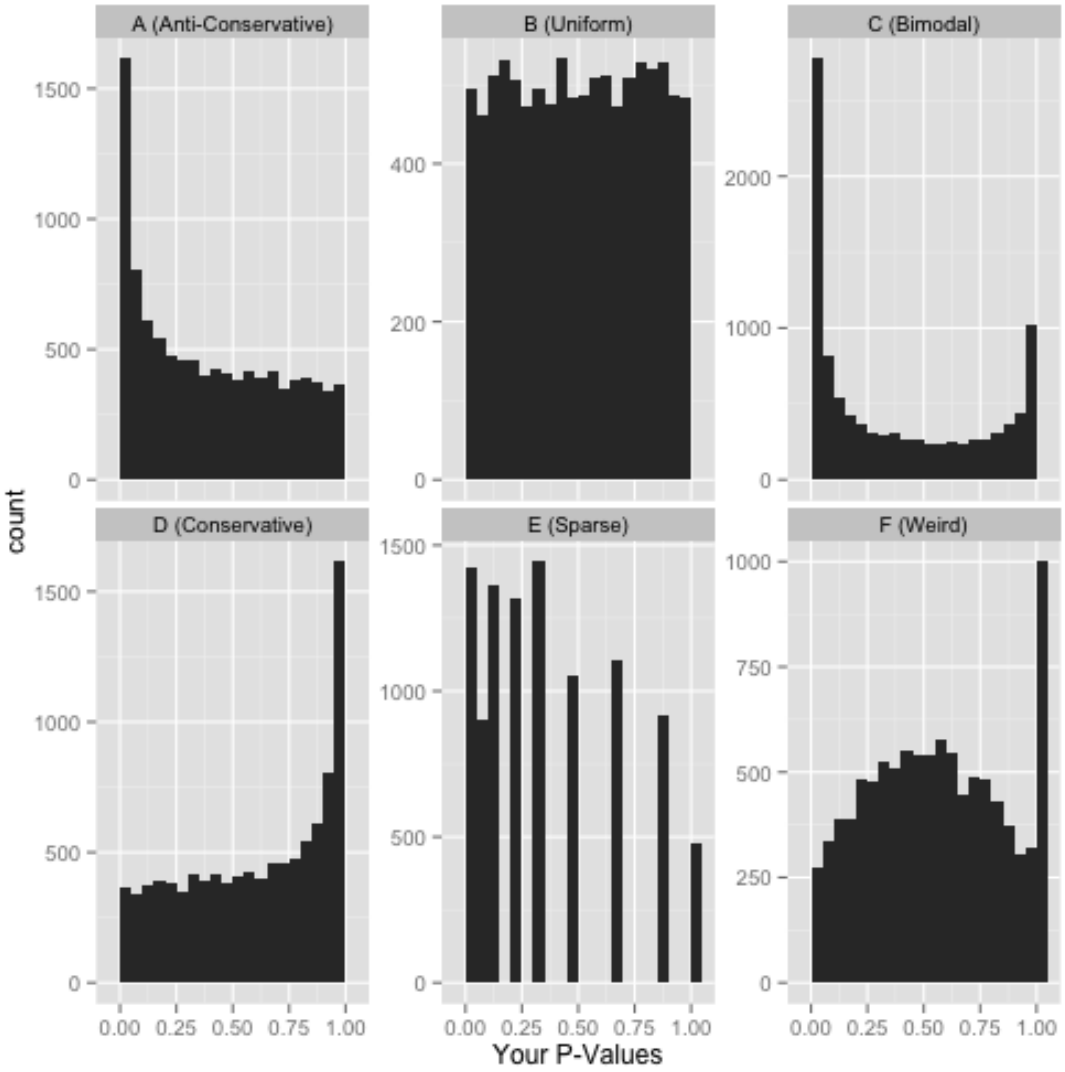
- Porovnání míry exprese mezi skupinami vzorků.
- Vztaženo ke kontrolní skupině.
- Průměrné exprese: kontrolní skupina $\overline{y_c}$, zájmová skupina (pacienti) $\overline{y_p}$
- Fold change:

$$fc = \frac{\overline{y_p}}{y_c}$$

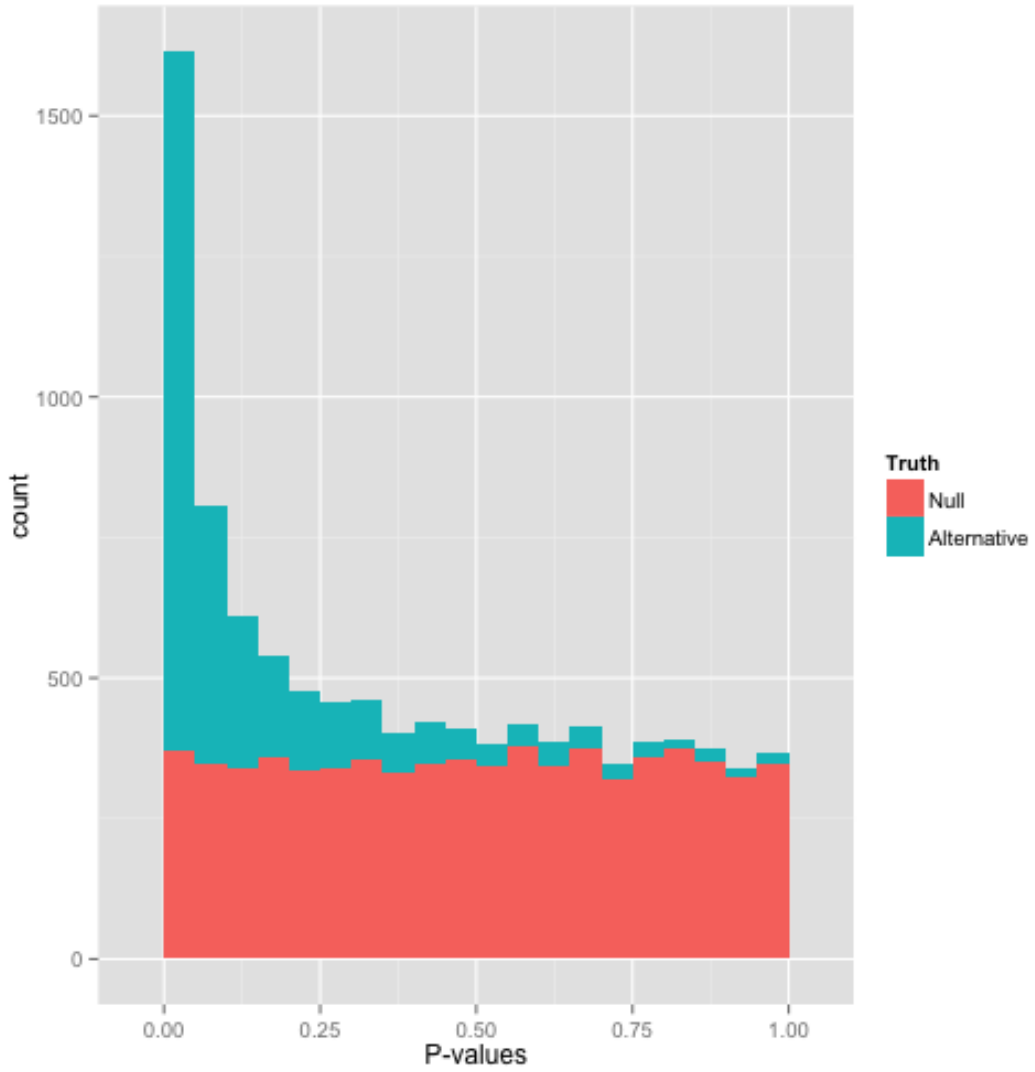
- V logaritmickém tvaru čteme změnu exprese stejně pouze s opačným znaménkem.

$$l2fc = \log_2 \frac{\overline{y_p}}{y_c}$$

Ověření rozdělení p-hodnot



Intuice rozdělení p-hodnot



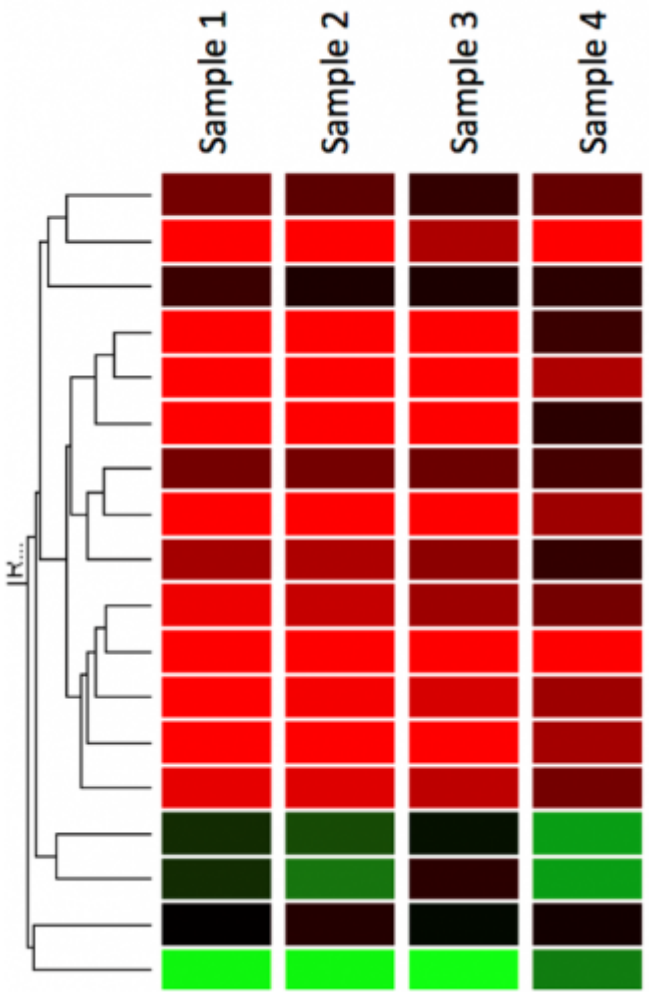
RNA Seq data

- Obecně počty readů spíše odpovídají negativnímu binomickému rozdělení.
- Sestavujeme obecný linární regresní model.
- Waldův test.

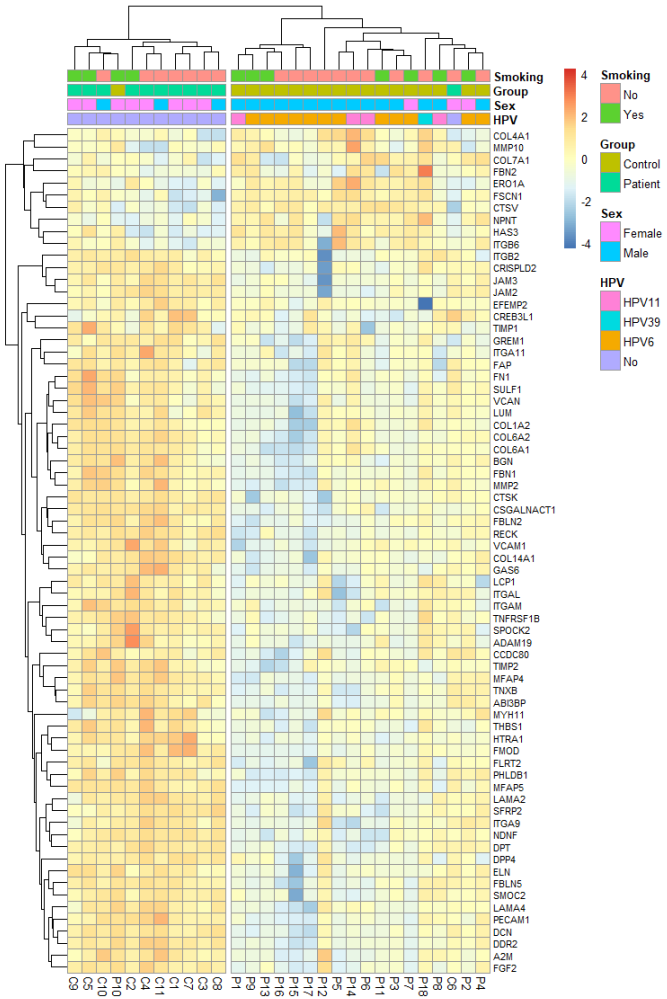
Biologická interpretace dat z genové exprese

Nejbvyklejší metodou vizualizace (vědecko-výzkumná úroveň) exprimovaných dat jsou tzv. heatmapy.

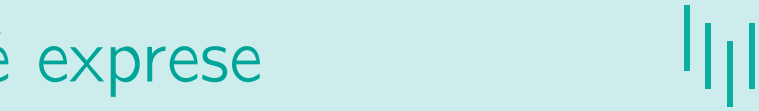
- Použito hierarchické shlukování.
- Geny shlukovány podle podobné exprese.
- Pacienti shlukováni podle podobné exprese genů.



Biologická interpretace dat z genové exprese - příklad



Biologická interpretace dat z genové exprese



- Obvykle se nám nepodaří rozdělit vzorky do skupin.
- Vzorky nejsou dokonalé.
- Vzorky nemusí v konkrétní časový okamžik odběru a sekvenování popisovat stejný stav buněk.
- Proces analýzy zahrnuje mnoho kroků, každý krok vnáší určitou chybu.
- Exprese genů je komplexní proces, v experimentu máme pouze RNA sekvence, důvodů exprese však může být více (např. fenotypy, nemoci).
- Vzorky měřeny na různých strojích => potřeba mezistrojového přizpůsobení, plánování experimentu - technické replikáty.
- Magické statistické metody pro zahrnutí variability měření a skrytých proměnných, RUV (remove unwanted variance), SVA (surrogate variable analysis)

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

20. listopadu 2019

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY