

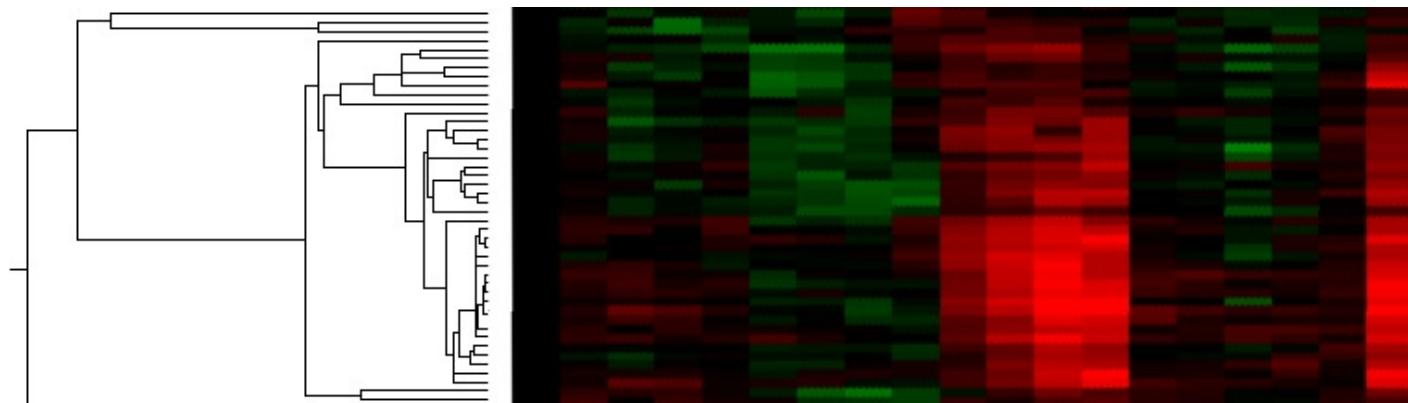
Oznamy

- Body z DÚ 1 budú časom v Moodle
- DÚ 2 je na stránke, odovzdať do 3.12.
- Stretnutia journal clubu: skupiny 1,2,4,7 oznámili dátum, 7 odovzdala aj krátku správu zo stretnutia. Chýbajú skupiny 5,6,8. Stretnutie má byť do 21.11. (budúci piatok)
- Ak máte nejaké otázky k článku, kontaktujte vyučujúcich

Regulácia génovej expresie

Broňa Brejová

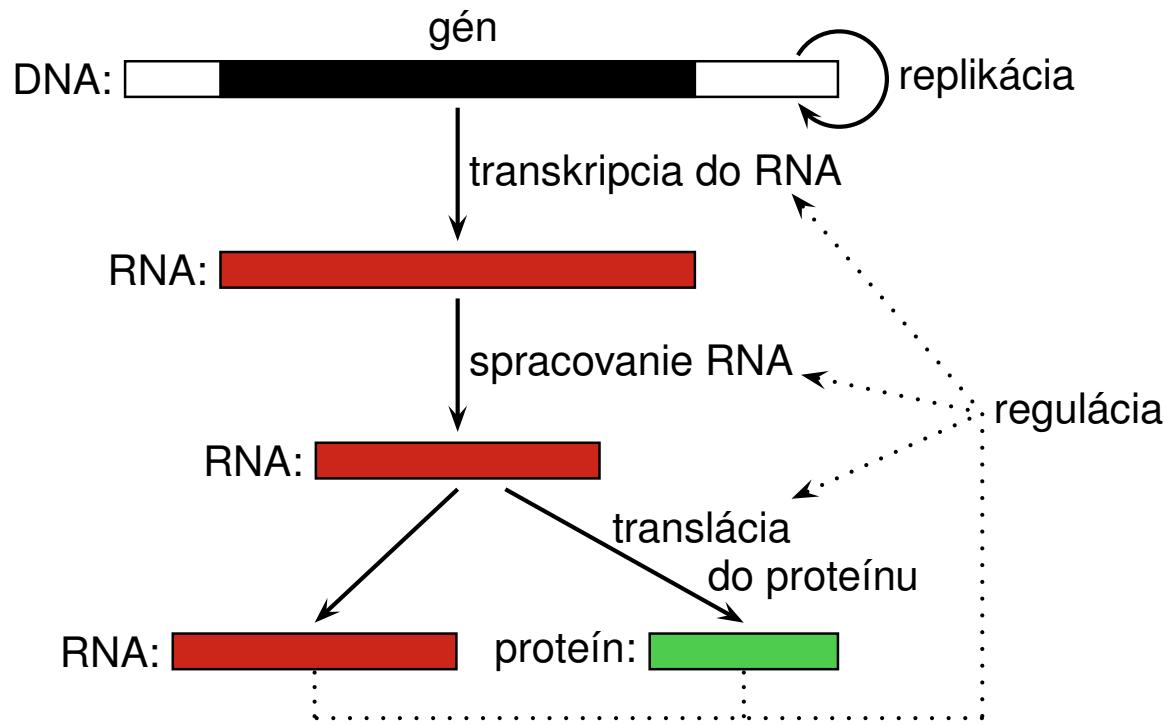
13.11.2025



Aká informácia je uložená v DNA?

Gény: Predpisy na tvorbu proteínov a funkčných RNA molekúl.

Riadenie ich expresie: kedy a koľko sa má tvoriť.



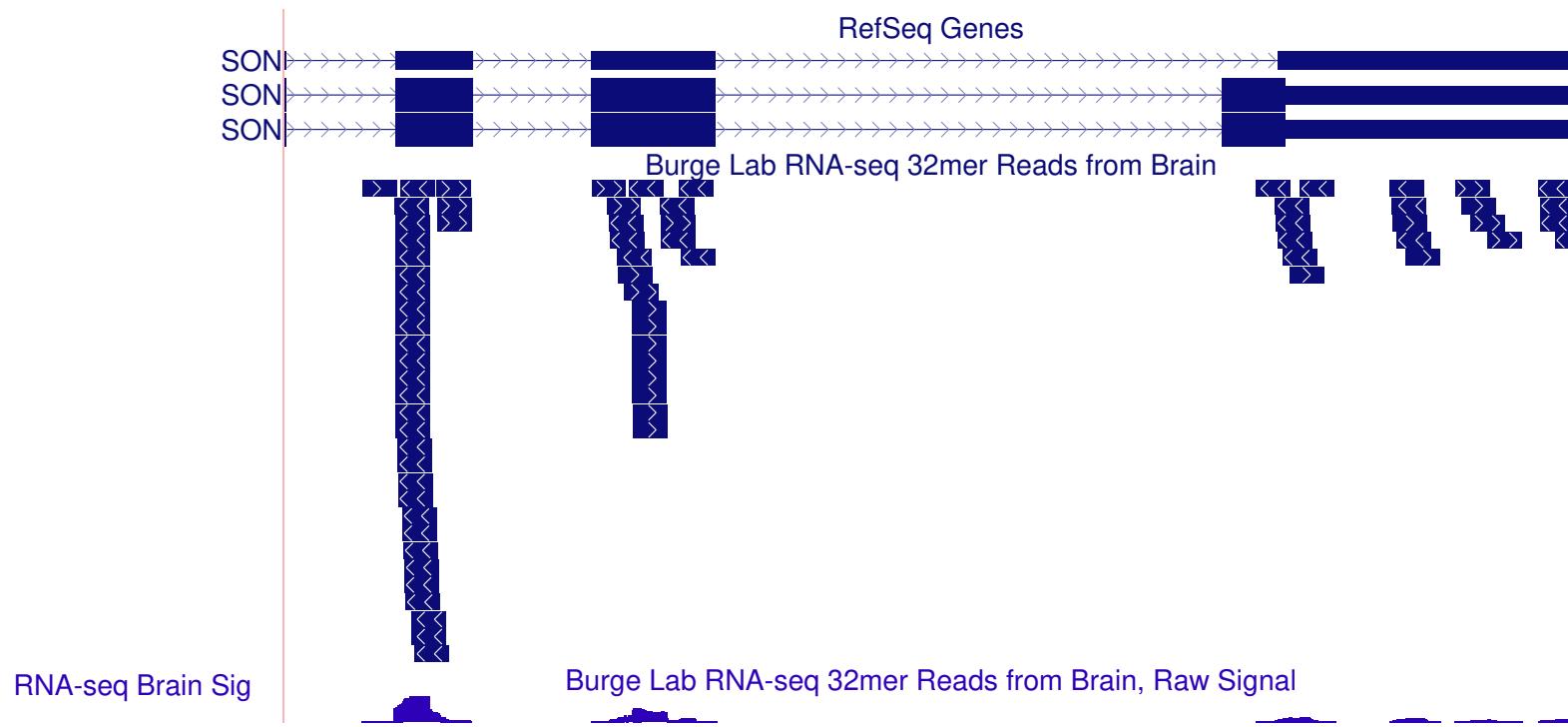
Regulácia na úrovni transkripcie, spracovania, translácie, posttranslačných modifikácií, ...

Ciele

- Zistiť, za akých podmienok je daný gén exprimovaný
(súvisí s funkciou génu)
- Ktoré gény ho regulujú
- Detaily regulačného mechanizmu
(väzobné miesta, zmeny v množstve expresie, . . .)

Technológia: RNA-seq

Sekvenujeme RNA extrahovanú z bunky,
mapujeme na genóm, hĺbka pokrycia zodpovedá úrovni expresie,
opakujeme za rôznych podmienok



Používa sa aj staršia technológia microarray (expression array)

Príklad dát o expresii

Riadky: gény, stĺpce: sekvenované vzorky, čísla: počet čítaní

	komora 1	komora 2	...	predsieň 1	predsieň 2	...
GJA5	377	348		6900	5912	
MAP3K5	1581	1326		254	237	
IDH2	19720	16734		5101	7059	
...						

Luo et al. 2021. Chamber-enriched gene expression profiles in failing human hearts with reduced ejection fraction. Scientific reports.

Celkovo 84 vzoriek ľudských sŕdc

My porovnávame vzorky z donorov bez porúch srdca

19 z komôr (ventricle), 17 z predsiení, 18 807 génov

Dnes: iný typ dát

Všetky ostatné prednášky: pracujeme so sekvenciami

- zostavovanie genómov
- zarovnávanie sekvencií
- hľadanie génov
- fylogenetické stromy, populačná a komparatívna genomika
- štruktúra a funkcia proteínov a RNA

Dnes: tabuľka čísel

- typické dáta v štatistike
- možno použiť všeobecné metódy štatistiky, strojového učenia

V druhej polovici prednášky sa aj dnes vrátime k sekvenciám

Normalizácia dát

Pozor: pôvodné počty sa **nedajú navzájom porovnávať**:

- v rôznych vzorkách sa získalo rôzne celkové množstvo dát
- RNA sa láme a teda dlhšie gény budú mať viac čítaní

Jednoduchá normalizácia: **transcripts per million (TPM)**:

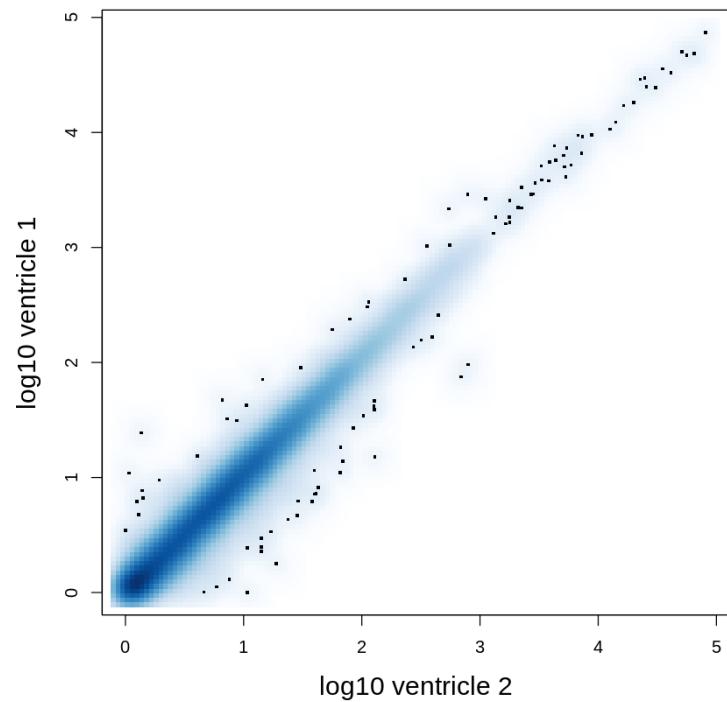
- počty vydelíme dĺžkou transkriptu (mRNA) v tisícoch
- potom každý stĺpec vydelíme jeho súčtom a vynásobíme miliónom

	komora 1	komora 2	...	predsieň 1	predsieň 2	...
GJA5	3.2	2.7		43.9	35.6	
MAP3K5	11.8	8.8		1.4	1.2	
IDH2	349.5	264.4		66.0	86.5	
...						

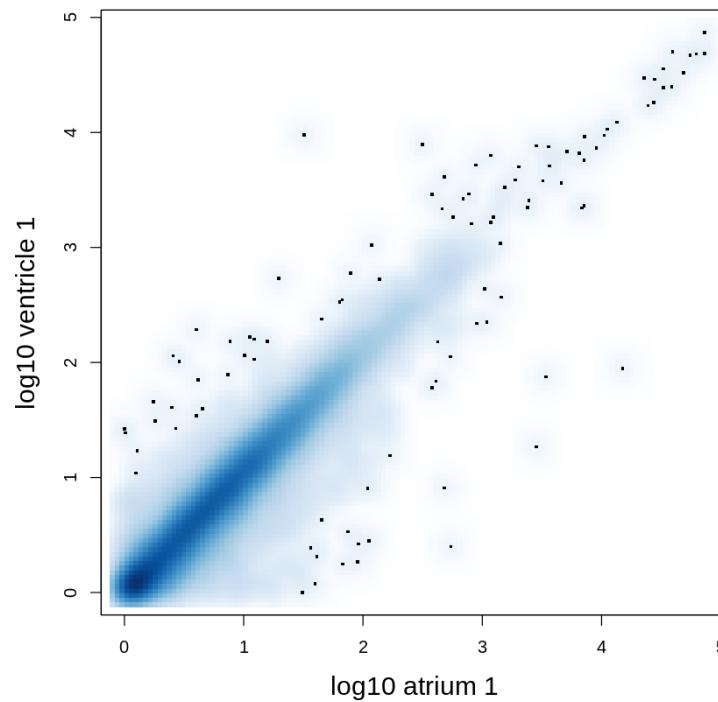
Existujú aj zložitejšie spôsoby normalizácie

Porovnanie dvoch vzoriek

Rovnaké tkanivá



Rôzne tkanivá



TPM hodnoty sme logaritmovali, lebo majú veľký dynamický rozsah

Pomerne veľká korelácia medzi vzorkami

Pri porovnaní rôznych tkanív niektoré hodnoty d'aleko od diagonály

– gény s odlišnou expresiou?

Porovnanie dvoch vzoriek: MA plot

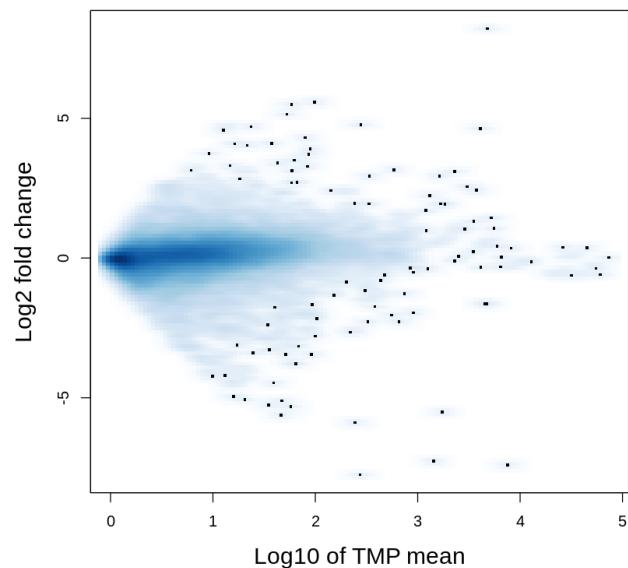
Tu: rôzne tkanivá

os x: \log_{10} priemerného TPM, $\log_{10} \left(\frac{x+y}{2} + 1 \right)$

os y: \log_2 pomera TPM, $\log_2 \left(\frac{x+1}{y+1} \right)$

– **log2 fold change (LFC)**

- LFC 1 znamená dvojnásobok, LFC -1 znamená polovicu
- LFC 2 znamená štvornásobok, LFC -2 znamená štvrtinu



Lepšia vizualizácia:

vodorovná čiara sa lepšie sleduje ako diagonálna

Ale pozor:

rozdiele videné v 2 vzorkách **môžu byť náhoda**

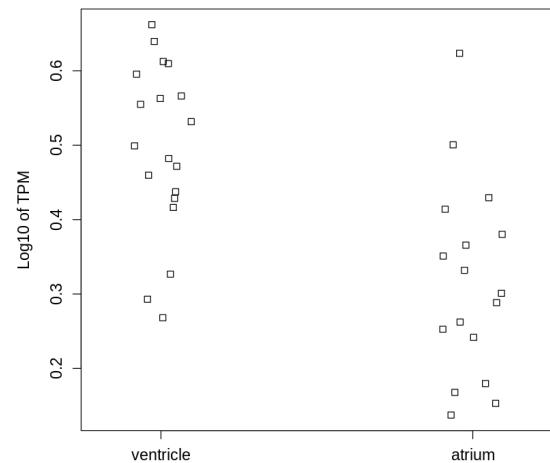
Porovnanie expresie génov

Rozdiely pozorované v dvojici vzoriek môžu byť náhoda,

Porovnávame väčšinou aspoň 3 vzorky v skupine

– napr. rôzni pacienti alebo opakovanie experimentu

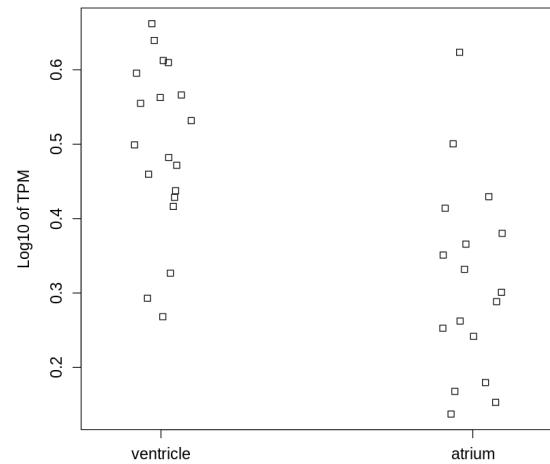
Príklad: Gén WNT11



Tu vidíme v komore (ventricle) posun k vyšším hodnotám
ale máme aj páry s opačným vzťahom

Differential gene expression analysis

Porovnávame väčšinou aspoň 3 vzorky v skupine
(napr. rôzni pacienti alebo opakovanie experimentu)



Štatistický test porovnávajúci dve skupiny čísel

Či má jedna významne vyššie hodnoty alebo môže ísť o náhodu

Veľa testov v klasickej štatistikе (napr. t-test)

Ale používajú sa špecializované nástroje zohľadňujúce **špecifiká dát**

– diskrétny počty, malý počet vzoriek, veľký šum

Nástroj DESeq2

Love et al 2014

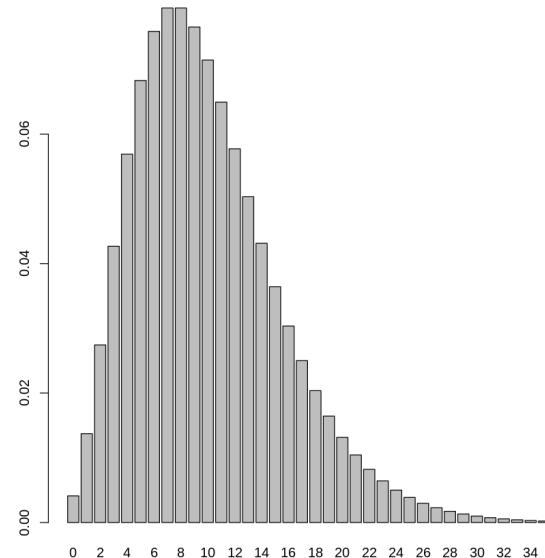
Vstupom sú pôvodné počty čítaní, nie TPM

Pravdepodobnosný model generujúci počty čítaní

Gén i , vzorka j , skupina r (u nás komora, predsieň)

Počet $K_{i,j}$ generovaný z **negatívneho binomického rozdelenia**

so strednou hodnotu $\mu_{i,j}$ a rozptylom $\sigma_{i,j}^2$



Príklad:

stredná hodnota $\mu = 10$

rozptyl $\sigma^2 = 30$

Rozdelenie vhodné na počty

Nástroj DESeq2

Gén i , vzorka j , skupina r (u nás komora, predsieň)

Počet čítaní $K_{i,j}$ generovaný z **negatívneho binomického rozdelenia**

so strednou hodnotou $\mu_{i,j}$ a rozptylom $\sigma_{i,j}^2$

$$\mu_{i,j} = s_j q_{i,r}$$

s_j normalizačná konštanta pre vzorku, podobná na výpočet TPM

$q_{i,r}$ odhad skutočnej expresie génu i v skupine r

$$q_{i,r} = 2^{\beta_{i,0} - \beta_{i,1}} \text{ alebo } q_{i,r} = 2^{\beta_{i,0} + \beta_{i,1}} \text{ podľa } r$$

$\beta_{i,0}$ je odhad logaritmu expresie génu i

$\beta_{i,1}$ je odhad (polovice) log2 fold change

Dôraz na modelovanie $\sigma_{i,j}^2$: kombinácia génu i a génov s podobnou expresiou

Zo všetkých dát odhadneme s_j

Pre dané $\beta_{i,0}$, $\beta_{i,1}$ a $\sigma_{i,j}^2$ vieme spočítať vierošodnosť počtov

Hľadáme najvierošodnejšie $\beta_{i,1}$

Nástroj DESeq2

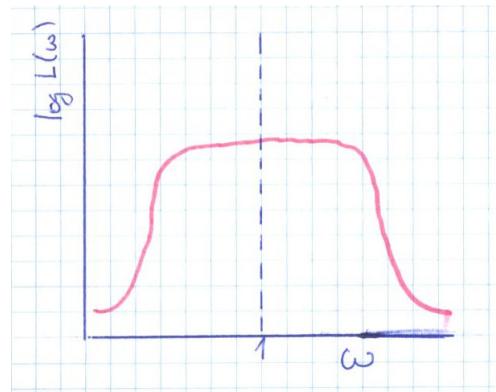
Z dát odhadneme najvierhodnejšie $\beta_{i,1}$ (log2 fold change, LFC)

Nulová hypotéza: $\beta_{i,1} = 0$

Podobné na pozitívny výber: pýtali sme sa, či $\omega > 1$, teraz $\beta_{i,1} \neq 0$

DESeq2 počíta p-hodnotu Waldovým testom

dal by sa aj test pomerov vierohodnosti



Viacnásobné testovanie

Pre každý gén máme p-hodnotu a odhad LFC

Nízka p-hodnota vrvá, že gén môže mať rôznu expresiu v dvoch skupinách
differentially expressed gene

Ale testovali sme veľa génov

Pri prahu 0.01 má každý 1% byť náhodou falošne vybraný

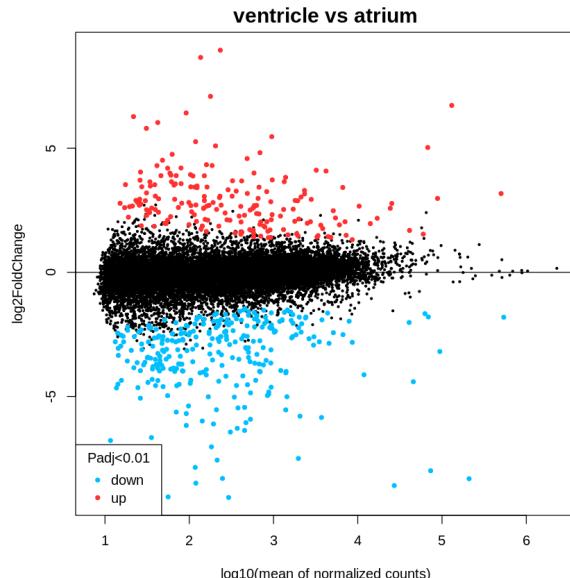
V našom príklade 18 807 génov, očakávame 188 falošných (náhodných)
DESeq2 našiel 711 génov, falošná môže byť až štvrtina z nich

Korekcia viacnásobného testovania:

Benjamini a Hochberg

Znižuj prah, kym očakávaný podiel falošných
neklesne pod napr. 1%

Po korekcii a filtri $|LFC| \geq 1$
zostáva 469 génov, cca 1% náhodné



Vizualizácia

50 génov s najnižšou p-hodnotou

Hodnoty zmenené na z-score:

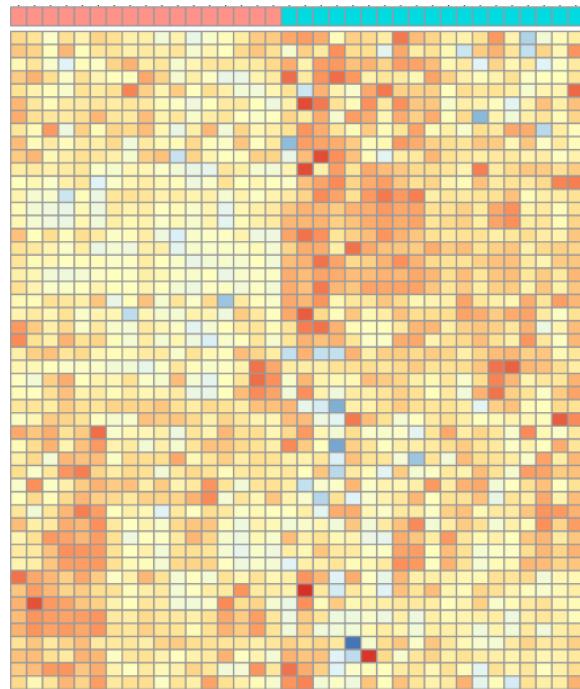
Profil génu: x_1, x_2, \dots, x_n

priemer \bar{x}

štandardná odchylka σ_x

$$\sigma_x = \frac{1}{n-1} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$



Zhlukovanie (clustering)

Ciel: nájsť skupiny génov s podobným profilom expresie.

Ak veľa génov v skupine má rovnakú funkciu,
ďalšie gény asi robia to isté

Meranie podobnosti profilov: napr. Pearsonov korelačný koeficient

Profil génu 1: x_1, x_2, \dots, x_n , priemer \bar{x}

Profil génu 2: y_1, y_2, \dots, y_n , priemer \bar{y}

$$C(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Číslo od -1 do 1, 1 pre lineárne korelované dátá

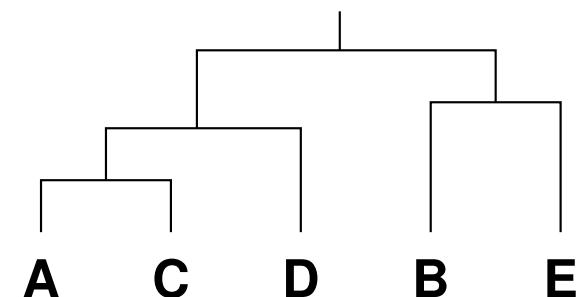
Vzdialenosť $d(x, y) = 1 - C(x, y)$

Aj iné možnosti, napr. Euklidovská vzdialenosť

Hierarchické zhlukovanie

- Podobné na metódu spájania susedov vo fylogenetických stromoch
- Začneme s každým génom v samostatnej skupinke
- Nájdeme dve najbližšie skupinky a spojíme ich do jednej
- Opakujeme, kým nie sú všetky gény spolu
- Vzdialenosť skupiek: napr. vzdialenosť najbližších génov z jednej a druhej, alebo priemer vzdialenosťí cez všetky páry
- Výsledkom je strom zobrazujúci postupnosť spájania

	A	B	C	D	E
gén A	0	0.6	0.1	0.3	0.7
gén B	0.6	0	0.5	0.5	0.4
gén C	0.1	0.5	0	0.6	0.6
gén D	0.3	0.5	0.6	0	0.8
gén E	0.7	0.4	0.6	0.8	0



Hierarchické zhlukovanie - príklad

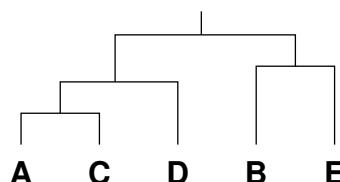
Vzdialenosť skupiniek ako vzdialenosť najbližších génov z jednej a druhej
(single linkage clustering)

	A	B	C	D	E	
gén A	0	0.6	0.1	0.3	0.7	
gén B	0.6	0	0.5	0.5	0.4	A C
gén C	0.1	0.5	0	0.6	0.6	
gén D	0.3	0.5	0.6	0	0.8	
gén E	0.7	0.4	0.6	0.8	0	

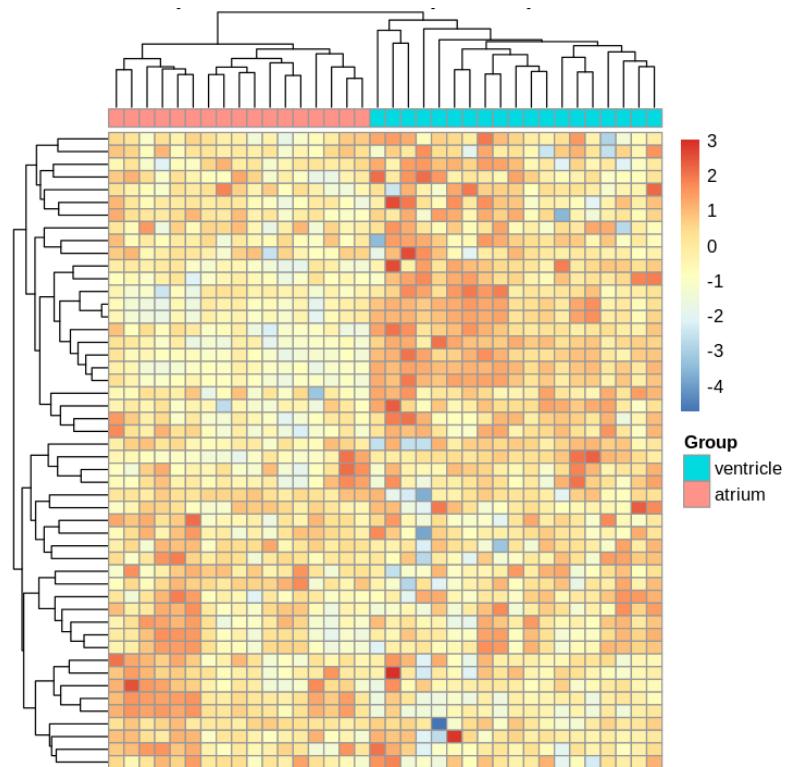
	A+C	B	D	E	
A+C	0	0.5	0.3	0.6	
B	0.5	0	0.5	0.4	A C D
D	0.3	0.5	0	0.8	
E	0.6	0.4	0.8	0	

	A+C+D	B	E	
A+C+D	0	0.5	0.6	
B	0.5	0	0.4	B E
E	0.6	0.4	0	

	A+C+D	B+E	
A+C+D	0	0.5	
B+E	0.5	0	



Príklad

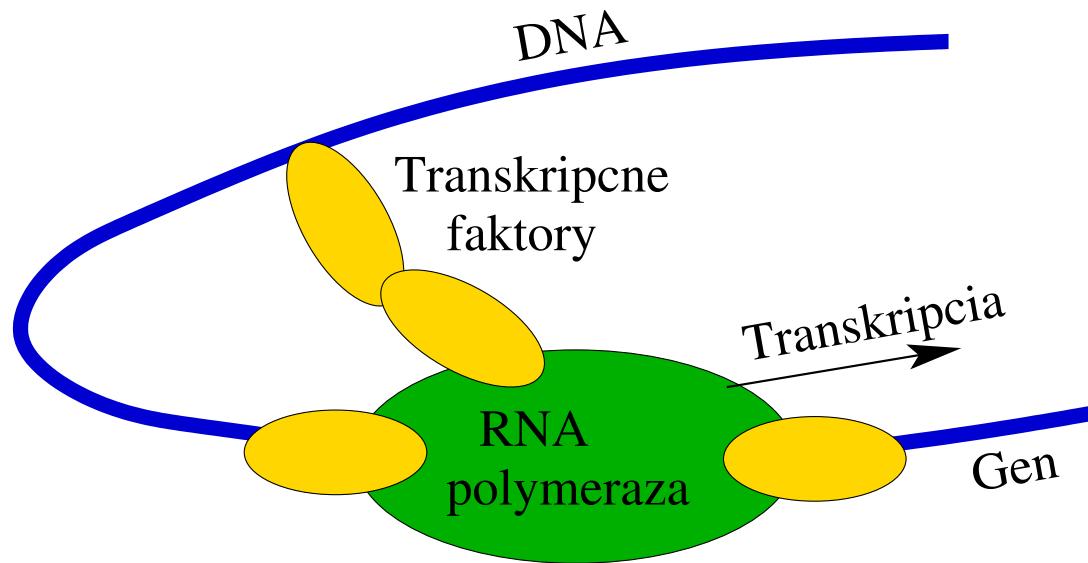


Zhlukovanie tiež pomáha vizualizácii dát,
podobné gény sa dostanú ku sebe

Zhlukovali sme aj pacientov, môže pomôcť objaviť subtypy choroby

Transkripčné faktory (TF)

Regulácia začatia transkripcie pomocou transkripčných faktorov:
proteíny viažúce DNA, pomáhajú pritiahnuť RNA polymerázu



Človek má vyše 2000 TF-ov

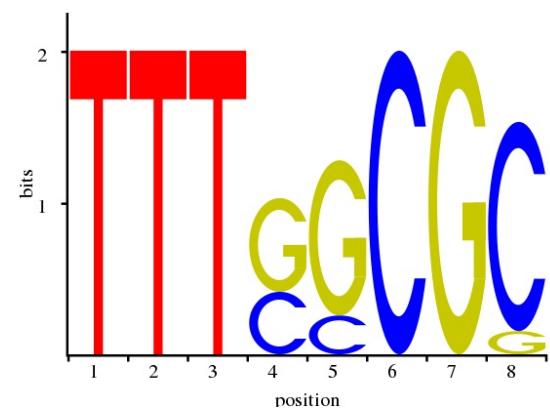
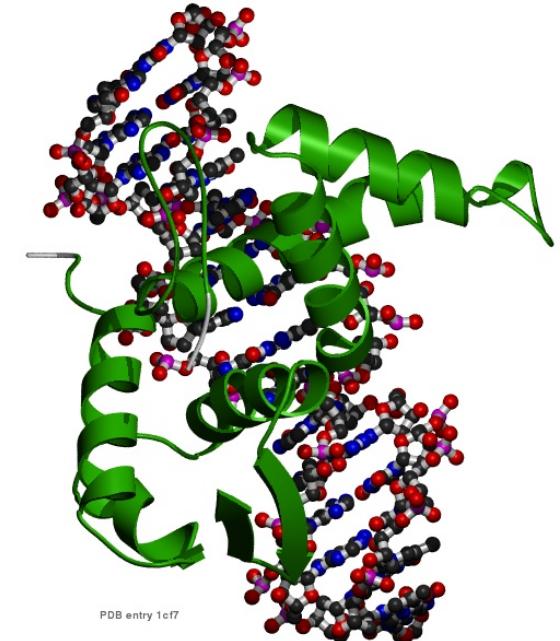
Môžu zvyšovať alebo znižovať mieru expresie,
fungovať v skupinách

Príklad: transkripčný faktor E2F1

- Reguluje bunkový cyklus
- Viaže TTTCCCGC alebo TTTCGCGC,
prípadne ďalšie varianty

A	0	0	0	0	0	0	0	0
C	0	0	0	4	2	10	0	9
G	0	0	0	6	8	0	10	1
T	10	10	10	0	0	0	0	0

- Sekvencie DNA, na ktoré sa viaže určitý TF
chceme **reprezentovať**
ako sekvenčný **motív**
a hľadať **ďalšie výskytu** v genóme



Reprezentácia väzobných motívov

Reťazec s nezhodami (konsenzus):

motív je reťazec, výskyty môžu mať vopred ohraničený počet nezhôd

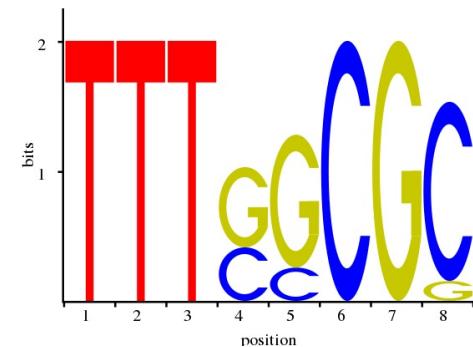
Príklad: motív TTTGGCGC + 1 nezhoda

TTTGGCGC, TTAGGCGC, TTTGC^CCGC sú výskyty motívu

TTTCCCGC nie je výskyt

Zostavenie motívu: napr. vezmi najčastejšie písmeno na každej pozícii

A	0	0	0	0	0	0	0	0
C	0	0	0	4	2	10	0	9
G	0	0	0	6	8	0	10	1
T	10	10	10	0	0	0	0	0



Reprezentácia väzobných motívov 2

Regulárny výraz:

niektoré pozície motívu dovoľujú výber z viacej možností

[GC] znamená pozíciu, na ktorej môže byť G alebo C

N znamená hociktorú bázu

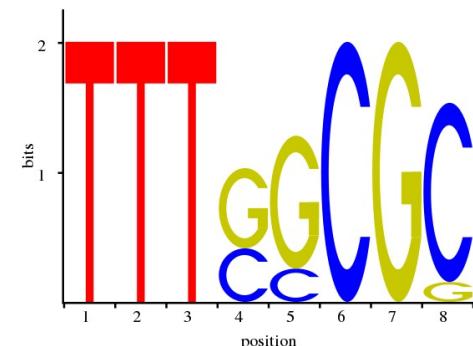
Príklad: motív TTT[CG][CG]CGC

TTTGGCGC, TTTCCCGC, TTTGCCGC sú výskyty motívu

TTAGGGCGC nie je výskyt

Zostavenie motívu: povoľ najčastejšie bázy na každej pozícii

A	0	0	0	0	0	0	0	0
C	0	0	0	4	2	10	0	9
G	0	0	0	6	8	0	10	1
T	10	10	10	0	0	0	0	0



Reprezentácia väzobných motívov 3

Position specific scoring matrix (PSSM, PWM):

skórovacia matica, skóre pre každú bázu na každej pozícii

Výskyty dosahujú skóre väčšie ako číslo T

Príklad: $T = 8$

A	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0
C	-1.6	-1.6	-1.6	0.6	0.0	1.5	-1.6	1.4
G	-1.6	-1.6	-1.6	1.0	1.3	-1.6	1.5	-0.5
T	1.1	1.1	1.1	-2.0	-2.0	-2.0	-2.0	-2.0

TTT**CC**CGC je výskyt: $1.1+1.1+1.1+0.6+0.0+1.5+1.5+1.4=8.3$

TTTGGCG**G** je výskyt: $1.1+1.1+1.1+1.0+1.3+1.5+1.5-0.5=8.1$

TT**A**GGCGC nie je: $1.1+1.1-2.0+1.0+1.3+1.5+1.5+1.4=6.4$

Zostavenie matice z frekvencií: budúca prednáška

Hľadanie výskytov motívu v genóme

- Zoberieme motív v niektornej reprezentácii:
 - Konsenzus, napr. TTTGGCGC + 1 nezhoda
 - Regulárny výraz, napr. TTT[CG][CG]CGC
 - Skórovacia matica, napr. prah $T = 8$ a matica:

A	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0
C	-1.6	-1.6	-1.6	0.6	0.0	1.5	-1.6	1.4
G	-1.6	-1.6	-1.6	1.0	1.3	-1.6	1.5	-0.5
T	1.1	1.1	1.1	-2.0	-2.0	-2.0	-2.0	-2.0

- Pre každú pozíciu v genóme testujeme, či je výskytom motívu
- Výskyty sú potenciálne väzobné miesta

Hľadanie výskytov motívu v genóme – problém

- Hľadanie motívu v genóme: skús každú pozíciu, či je výskytom
- Okrem **väzobných miest** často aj veľa **náhodných výskytov**
- Vieme spočítať E-hodnotu: koľko výskytov očakávame v náhodnej sekvencii
- Napr. TTT[CG][CG]CGC sa vyskytuje v priemere raz za 30 000 báz
- Na zlepšenie špecifickosti hľadáme
 - zhluhy väzobných miest,
 - miesta podporené experimentálne,
 - evolučne zachované
- Databázy motívov, napr. TRANSFAC, JASPAR

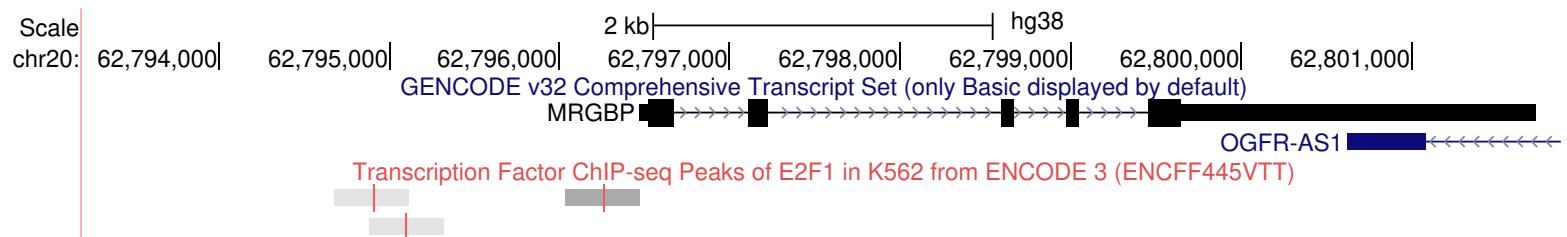
Ako nájsť väzobné miesta experimentálne?

Chromatin immunoprecipitation (ChIP)

Pomocou protilátky (antibody) na špecifický transkripčný faktor zistí, kde približne sa tento faktor viaže:

- Väzba medzi TF a DNA sa spevní formaldehydom
- DNA sa naseká na kusy
- Kusy, na ktorých je TF, sa zachytia na protilátke
- DNA sa izoluje a sekvenuje (**ChIP-seq**)

Problém: zistíme len približnú polohu väzobného miesta



Ako nájsť motívy výpočtovými metódami?

... ak nemáme niekoľko príkladov väzobného miesta

- Máme skupinu sekvencií, kde každá obsahuje väzobné miesto toho istého TF, ale väzobné preferencie TF nie sú známe
- Snažíme sa nájsť **čo najšpecifickejší** motív, ktorý sa vyskytuje vo všetkých týchto sekvenciách
resp. sa vyskytuje častejšie, ako by sme očakávali.
- **V súčasnosti:** zoberieme oblasti detegované pomocou ChIP-seq okolo väzobných miest, nájdený motív použijeme na presnejšie určenie polohy väzby TF
- **Pôvodne:** zoberieme skupinu génov s podobným profilom expresie a teda možno regulovaných tým istým TF, hľadáme motív v oblastiach pred týmito génmi

Príklad: Consensus Pattern Problem (CPP)

Jednoduchá formulácia problému hľadania motívov

Vstup: dĺžka motívu L , reťazce (sekvencie) S_1, S_2, \dots, S_k

Výstup: motív (reťazec) M dĺžky L

a výskyt motívu v každom S_i (reťazec s_i dĺžky L)

také, že celkový počet nezhôd medzi M a s_i je najmenší možný

Príklad:

Vstup: CAAACAT, AGTAGC, TAACCA, TCTCCTC, $L = 4$

Výstup: motív TAAC

výskyty a nezhody AAAC 1, TAGC 1, TAAC 0, TCTC 2

celkový počet nezhôd 4

Riešenie CPP

NP-ťažký problém

- **Idea 1:** Vyskúšaj všetky možné motívy dĺžky L

Problém: Nepraktické — prečo?

- **Idea 2:** Vyskúšaj všetky možné podreťazce dĺžky L reťazcov S_1, \dots, S_k

Problém: Nemusí fungovať — prečo?

Ale dá sa dokázať, že cena riešenia bude najviac dvojnásobok optima
(2-aproximačný algoritmus)

- **Ďalšie vylepšenie:** Skúšame všetky konsenzus sekvencie ℓ podreťazcov.
PTAS (polynomial-time approximation scheme)

Príklad:

Vstup: $L = 4$

CAAACAT,

AGTAGC,

TAACCA,

TCTCCTC

Výstup:

motív TAAC

výskyty a nezhody

AAAC 1,

TAGC 1,

TAAC 0,

TCTC 2

spolu 4 nezhody

Praktickejší prístup k hľadaniu motívov

Pravdepodobnostný model generujúci sekvenciu S pomocou matice frekvencií báz v motíve W a frekvencie báz q mimo motívu

A	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
C	0.01	0.01	0.01	0.39	0.19	0.97	0.01	0.01	0.89	
G	0.01	0.01	0.01	0.59	0.79	0.01	0.97	0.97	0.09	
T	0.97	0.97	0.97	0.01	0.01	0.01	0.01	0.01	0.01	

$$q(A) = 0.3, q(C) = 0.2, q(G) = 0.2, q(T) = 0.3$$

Pozícia motívu v S sa zvolí náhodne,
každá báza sa vygeneruje z q alebo z jedného stĺpca W

Tento model definuje rozdelenie $\Pr(S | W)$.

Hľadanie motívov cez pravdepodobnostné modely

Vstup: dĺžka motívu L , sekvencie S_1, S_2, \dots, S_k , frekvencie q

Výstup: spoločný motív ako matica frekvencií W maximalizujúca viero hodnosť dát $\Pr(S_1|W) \cdot \dots \cdot \Pr(S_k|W)$

- Ťažký problém, používajú sa heuristické algoritmy
- Napríklad EM (expectation maximalization)
- Lokálna optimalizácia, ktorá konverguje k lokálnemu maximu viero hodnosti
- Softvér: MEME

Schéma algoritmu EM

- **Incializácia:**

Zvoľ si počiatočnú maticu W
(napr. zostavenú podľa jedného okna dĺžky L)

- **Iterácia:**

1. Prirad každej pozícii j v sekvencii S_i váhu $p_{i,j}$, ktorá zodpovedá pravdepodobnosti, že na pozícii $S_i[j]$ začína výskyt motívu W
2. Spočítaj W zo všetkých možných výskytov v S_1, \dots, S_k váhovaných podľa $p_{i,j}$

Iterácie zvyšujú viero hodnosť dát, kým nedôjde ku konvergencii.

Skúšame veľakrát z rôznych počiatočných W

Príklad algoritmu EM

A	0.10	0.10	0.10	0.10	0.10
C	0.10	0.10	0.10	0.70	0.70
G	0.10	0.10	0.10	0.10	0.10
T	0.70	0.70	0.70	0.10	0.10

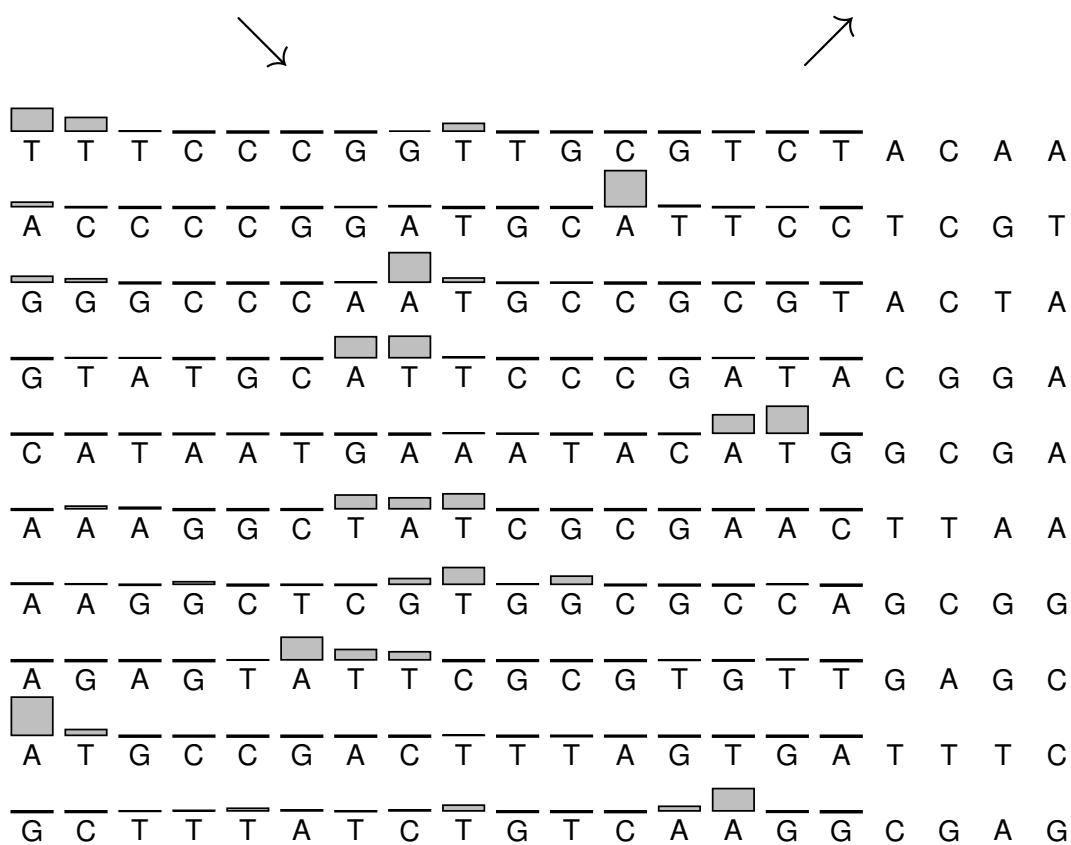
A	0.31	0.14	0.06	0.07	0.07
C	0.06	0.10	0.19	0.71	0.61
G	0.12	0.17	0.29	0.14	0.25
T	0.51	0.60	0.46	0.08	0.07

↓ ↗

T T T C C C G G T T G C G T C T A C A A
 A C C C C G G A T G C A T T C C C T C G T
 G G G C C C C A A T G C C C G C G T A C T A
 G T A T G C A T T C C C G A T A C G G G A
 C A T A A T G A A A T A C A T G G C G G A
 A A A G G G C T A T C G C G A A C T T A A A
 A A G G C T C G T G G C G C C C A G C G G G
 A G A G T A T T C G C G T G T T G A G A G C
 A T G C C G A C T T T A G T G A T T T T C
 G C T T T A T C T G T C A A G G C G A G

Príklad algoritmu EM: ďalšia iterácia

A	0.31	0.14	0.06	0.07	0.07		A	0.47	0.09	0.01	0.02	0.03
C	0.06	0.10	0.19	0.71	0.61		C	0.02	0.11	0.20	0.80	0.58
G	0.12	0.17	0.29	0.14	0.25		G	0.08	0.22	0.48	0.15	0.35
T	0.51	0.60	0.46	0.08	0.07		T	0.42	0.58	0.30	0.03	0.03



Príklad algoritmu EM: po 20 iteráciách

A	0.10	ϵ	ϵ	ϵ	ϵ
C	0.12	0.52	0.48	$1 - 3\epsilon$	ϵ
G	ϵ	0.48	0.52	ϵ	$1 - 3\epsilon$
T	0.78	ϵ	ϵ	ϵ	ϵ

The diagram shows 10 DNA sequences aligned vertically. Shaded boxes highlight specific matches between the first sequence and others. The first sequence is: T-T-T-C-C-C-G-G-T-T-G-C-G-T-C-T-A-C-A-A. The second sequence is: A-C-C-C-C-G-G-A-T-G-C-A-T-T-C-C-T-C-G-T. The third sequence is: G-G-G-C-C-C-A-A-T-G-C-C-G-C-G-T-A-C-T-A. The fourth sequence is: G-T-A-T-G-C-A-T-T-C-C-C-G-A-T-A-C-G-G-A. The fifth sequence is: C-A-T-A-A-T-G-A-A-A-T-A-C-A-T-G-G-C-G-A. The sixth sequence is: A-A-A-G-G-G-C-T-A-T-C-G-C-G-A-A-C-T-T-A-A. The seventh sequence is: A-A-G-G-C-T-C-G-T-G-G-C-G-C-C-A-G-C-G-G. The eighth sequence is: A-G-A-G-T-A-T-T-C-G-C-G-T-G-T-T-G-A-G-C. The ninth sequence is: A-T-G-C-C-G-A-C-T-T-T-A-G-T-G-A-T-T-T-C. The tenth sequence is: G-C-T-T-T-A-T-C-T-G-T-C-A-A-G-G-C-G-A-G.

Zhrnutie

- RNA-seq meria úroveň expresie pre všetky gény naraz, ale v dátach veľa šumu
- Častá úloha: nájsť gény, ktorých expresia sa výrazne líši v dvoch skupinách vzoriek
- Zhlukovanie (clustering) nájde podobné gény, nepotrebueme o dátach vopred nič vedieť (unsupervised learning)
- Väzobné motívy môžeme reprezentovať rôznym spôsobom (reťazec, regulárny výraz, skórovacia matica)
- Tieto motívy nie sú dosť špecifické, okrem väzobných miest môžu mať aj ďalšie náhodné výskytu
- EM algoritmus na hľadanie nových motívov v sekvenciách