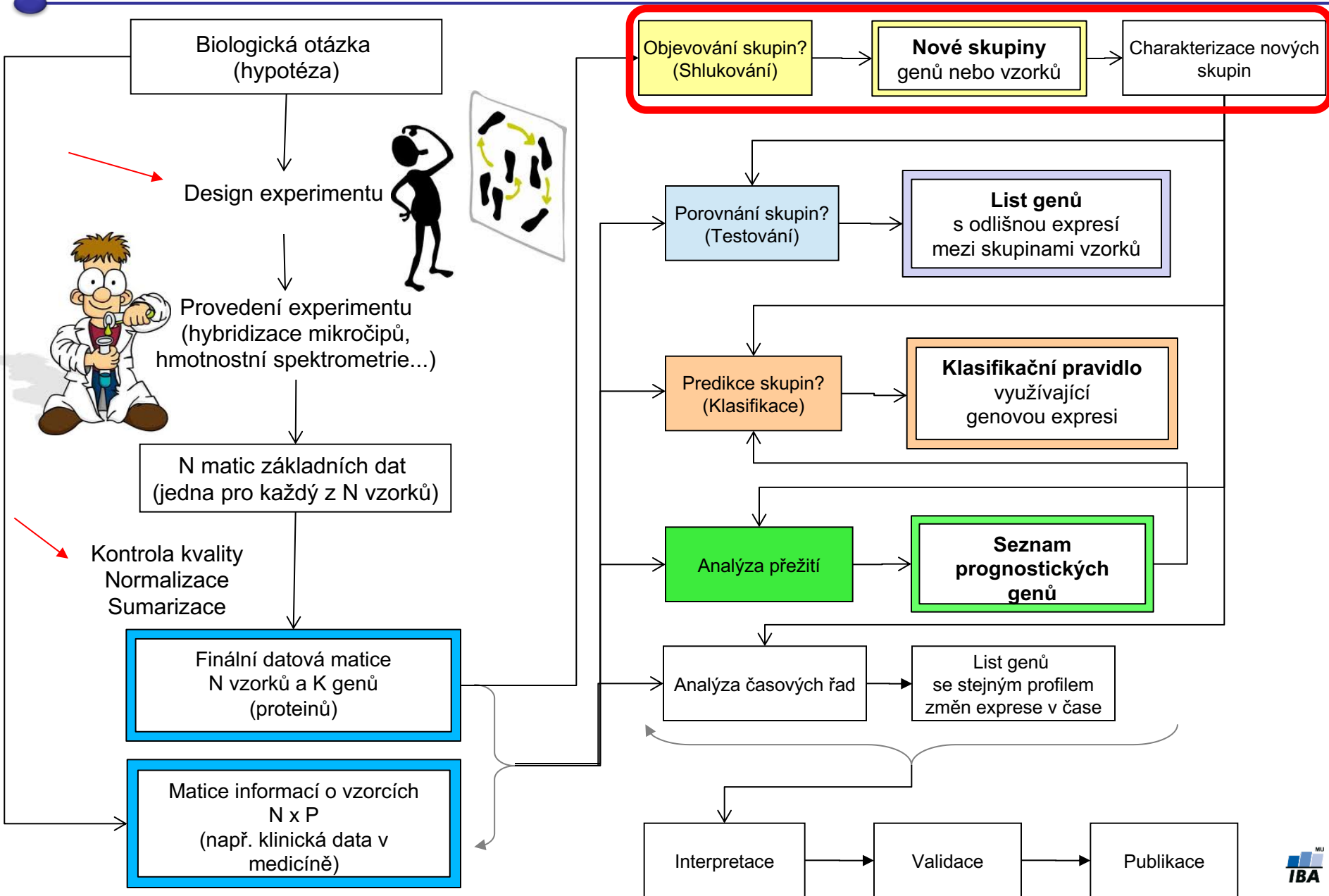


---

# Objevování skupin (class discovery)

# Společná schéma analýzy dat



# Tradiční schéma analýzy

- **Učení s učitelem (supervised learning)**
  - V tomto případě zobecňujeme známou strukturu dat na nové data
  - **Porovnávání skupin (class comparison)**
    - hledáme rozdíly v expresi, počtu kopií genů nebo abundanci proteinů mezi již definovanými skupinami
  - **Předpovídání skupin (class prediction)**
    - na známých skupinách se snažíme vytvořit klasifikátor, který by dokázal zařadit nového pacienta do jedné ze skupin
- **Učení bez učitele (unsupervised learning)**
  - V tomto případě struktura v datech není známá a musíme ji objevit
  - **Objevování skupin (class discovery)**
    - na základě informací o genech/proteinech hledáme nové skupiny
    - onemocnění X je velmi heterogenní a snažíme se identifikovat specifitější podtypy, které by mohli být cílem cílené terapie

# Společné znaky analýzy dat

---

- Velké množství proměnných
- Malé množství vzorek
- Proměnné jsou často korelované, s velmi komplexními vztahy
- Data obsahují množství šumu – biologická i technická variabilita

# Objevování skupin

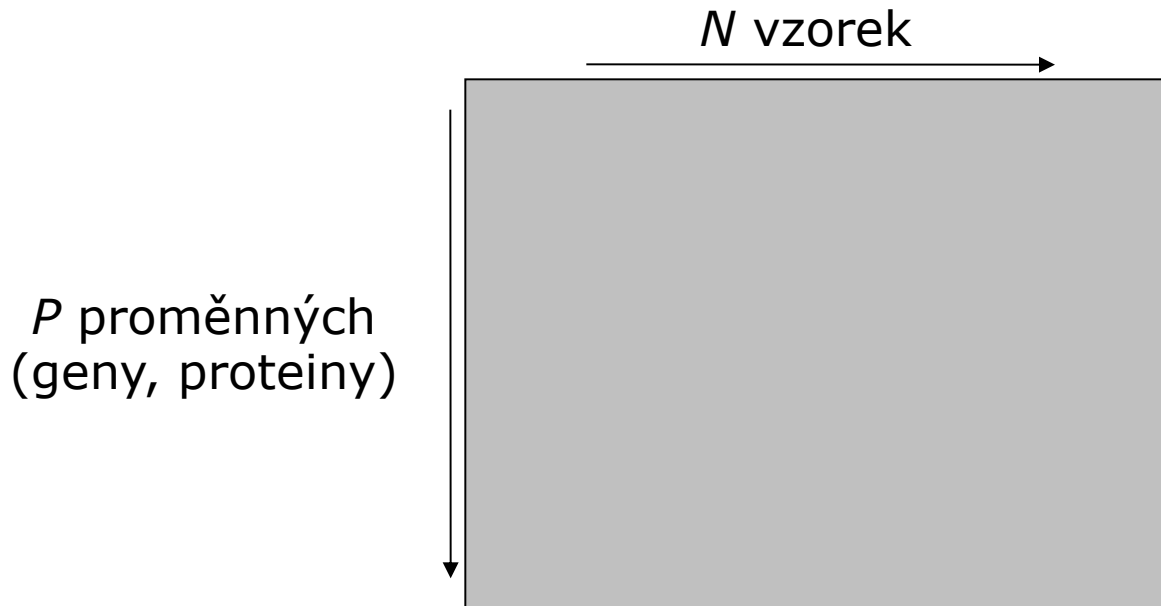
- Snažíme se vytvořit závěry o datovém souboru bez (braní do úvahy) jakékoliv předchozí znalostí biologických skupin (=shlukování)
- Cílem je vytvořit skupiny objektů na základě jejich vzájemné podobnosti
- Objekty uvnitř skupiny mají být co nejpodobnější a objekty z různých skupin mají být tak odlišné, jak jen je to možné
- Skupina metod pro objevování skupin je představovaná metodami shlukování bez učitele

# Co shlukujeme v molekulární biologii

- Geny/proteiny
  - Chceme identifikovat skupiny ko-regulovaných genů/proteinů
  - Chceme zredukovat dimenzi dat na základě funkčních genových/proteinových skupin
- Vzorky
  - Kontrolujeme kvalitu vzorků
  - Chceme najít nové skupiny vzorků (například podtypy)
  - Chceme zkontrolovat diskriminační schopnost genů vybraných při porovnávání známých skupin do vzorek

# Princip

- Máme datovou matici  $X$  velikosti  $N \times P$ 
  - $N$  – počet objektů (vzorek)
  - $P$  – počet proměnných (geny/proteiny)



- Hledáme nejlepší rozdělení dat na skupiny tak, aby nalezené skupiny byly uvnitř skupiny vysoce homogenní a mezi sebou vysoce heterogenní

# Typy shlukovacích metod

- Shlukovací metody se dělí na dvě hlavní skupiny:
  - 1. Metody založené na vzdálenostech**
    - neparametrické
    - nejčastěji používané, intuitivní
    - hierarchické a nehierarchické shlukování
  - 2. Metody založené na modelování**
    - parametrické, kladou silné předpoklady na rozložení dat
    - založeny na statistickém modelování – přiřazují každému objektu pravděpodobnost s jakou patří do daného shluku



# Metody založené na vzdálenostech I.

- Princip:
  1. Vypočítáme matici vzdáleností mezi objekty
  2. Vybereme shlukovací algoritmus
  3. Stanovíme počet shluků – jen u některých metod
  4. Aplikujeme shlukovací algoritmus na matici vzdálenosti  
→ získáme shluky
- Shlukovací algoritmy:
  - Hierarchické
    - Aglomerativní – Single, Complete, Average, Ward linkage, ...
    - Divizní - DIANA
  - Nehierarchické
    - K-means
    - PAM

# Metriky vzdáleností I.

Máme 2 vektory hodnot  $\mathbf{x} = (x_1, \dots, x_n)$ ,  $\mathbf{y} = (y_1, \dots, y_n)$

- Euklideovská vzdálenost: 
$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$
- Standardizovaná Euklideovská vzdálenost:

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2 / \sigma_{i^2}}$$

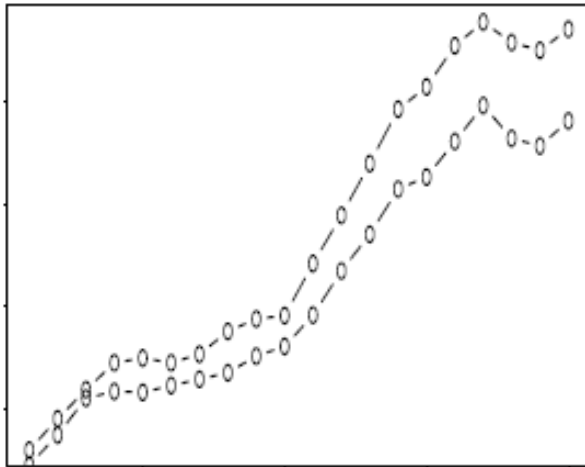
Metrika penalizuje – snižuje vzdálenost mezi objekty s velkou variabilitou, předpokládajíc, že jsou důležitější než objekty s malou variabilitou.

- Manhattanovská vzdálenost: 
$$d_M(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Robustnější vůči odlehlým hodnotám.

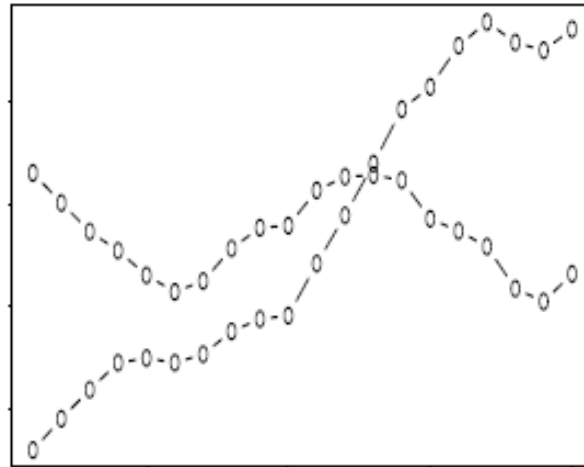
# Metriky vzdáleností II.

- Metriky založené na korelačním koeficientu  $r(x,y)$
- Můžeme odvodit dvě různé metriky:  $d_1(x,y) = [1 - r(x,y)]^2$   
 $d_2(x,y) = 1 - [r(x,y)]^2$
- Ukázka rozdílu mezi metrikama



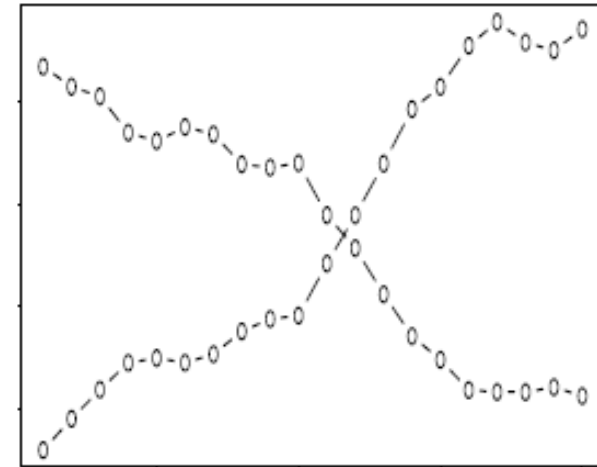
$r = 0.9$

$d_1=0.05, d_2=0.19$



$r = 0.0$

$d_1=0.5, d_2=1$



$r = -0.9$

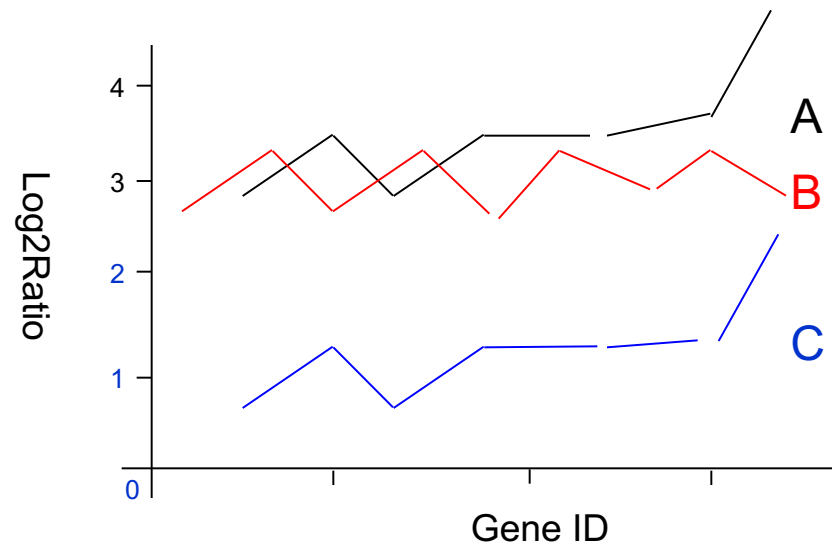
$d_1=0.95, d_2=0.19$

Při použití  $d_1$  budou geny s opačnými profily patřit do odlišných shluků, zatímco při použití metriky  $d_2$  budou patřit do toho stejného shluku.

Pokud chceme shluky interpretovat jako množiny genů ze stejné regulační sítě, použijeme raději  $d_2$ .

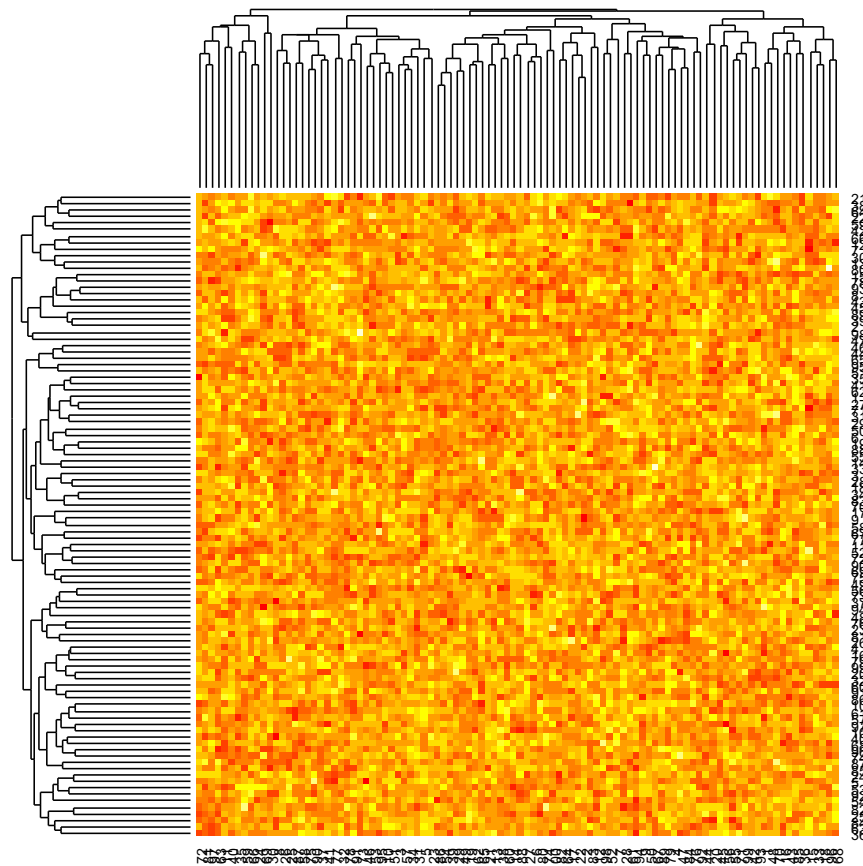
# Výběr metriky

- Výběr metriky záleží na tom, jaký typ podobnosti nás zajímá
  - Pokud nás zajímá průměrná exprese genů (A a B jsou podobné), aplikujeme Euklidovskou vzdálenost
  - Pokud nás zajímá vzor exprese genů (A a C jsou podobné), aplikujeme vzdálenost založenou na korelaci



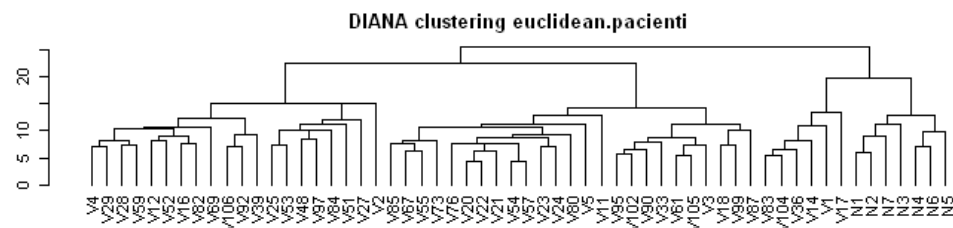
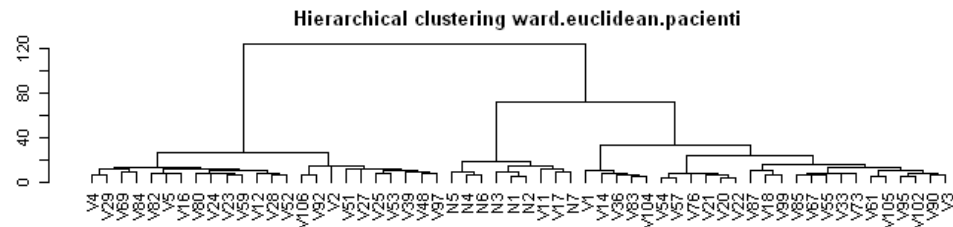
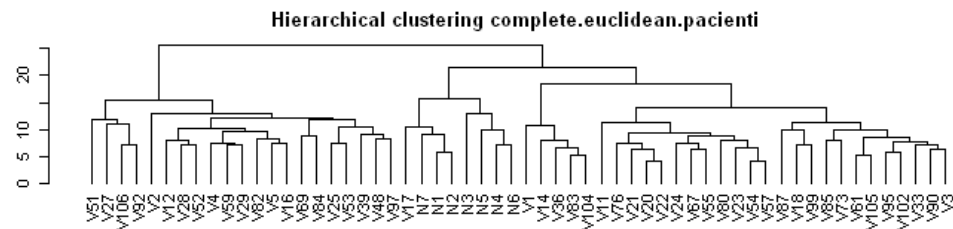
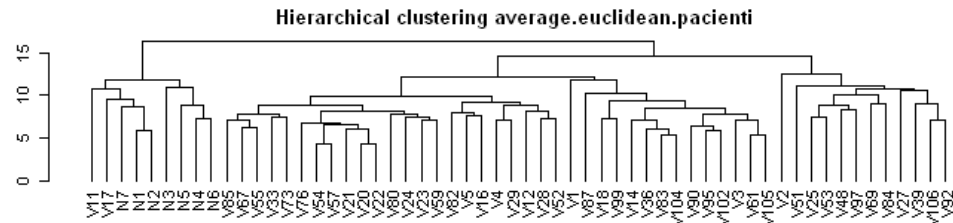
# Na co si dávat pozor I.

- Mnoho shlukovacích technik najde shluky i v datech, ve kterých nejsou žádné přirozené shluky, jen proto, že byly pro tento účel zkonstruované



# Na co si dávat pozor II.

- Výsledek jediného shlukování by nikdy neměl být považovaný za objektivní reprezentaci informace skryté v datech, protože je závislý od použité metody a také v rámci metody od nastavení!



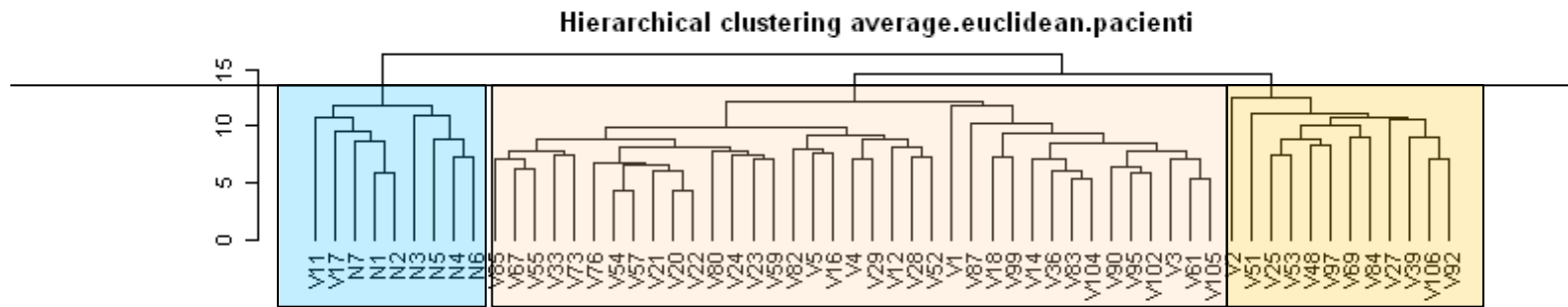
# Další problémy

---

- Výběr shlukovacího algoritmu a metriky ovlivňuje konečné výsledky
- Výsledky jsou závislé na samotných datech
- Kolik shluků?
- Potřebujeme odhad jistoty, že nalezené shluky jsou správné
- Odhad kvality shluků je založen na metrikách z dat z kterých byli shluky vytvořené

# Kolik shluků?

- V případě nehierarchických metod počet shluků určujeme dopředu
- V případě hierarchického shlukování vytváříme strom, dendrogram, který se potom prořezává

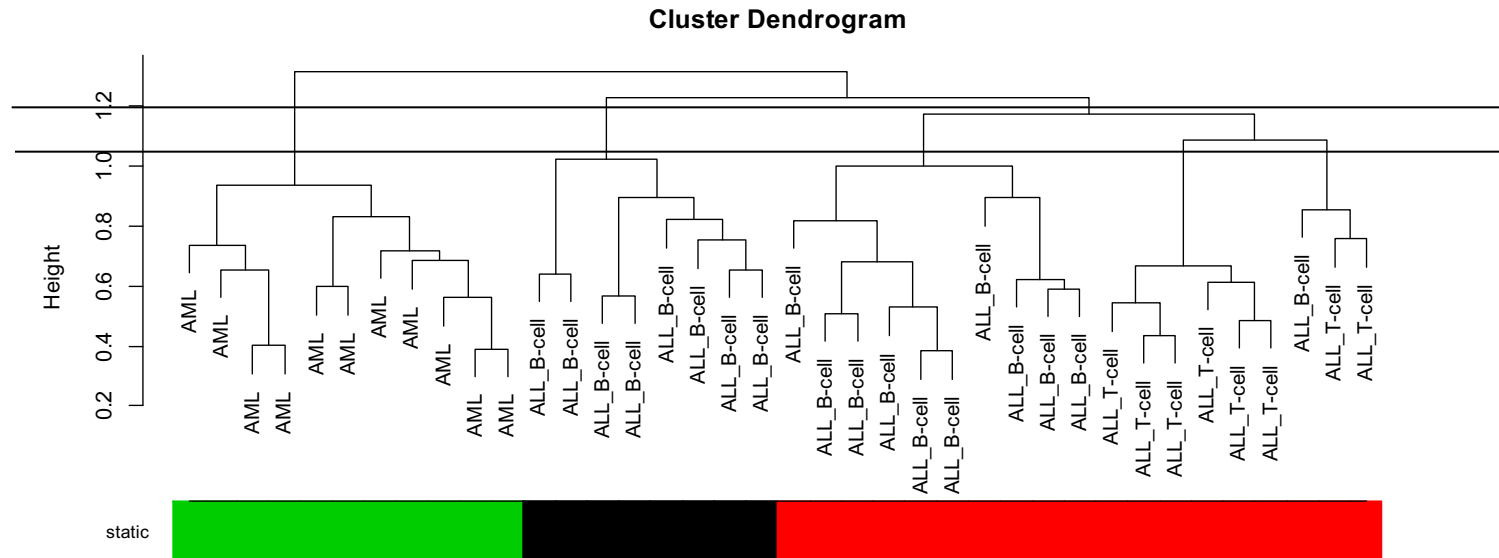


- Počet shluků je následně určený tak, aby heterogenita v rámci shluků byla co nejmenší a mezi shluky co největší
- Různé metriky heterogenity shluků – variabilita, Silhouette, ...



# Řezání dendrogramu jeho problém

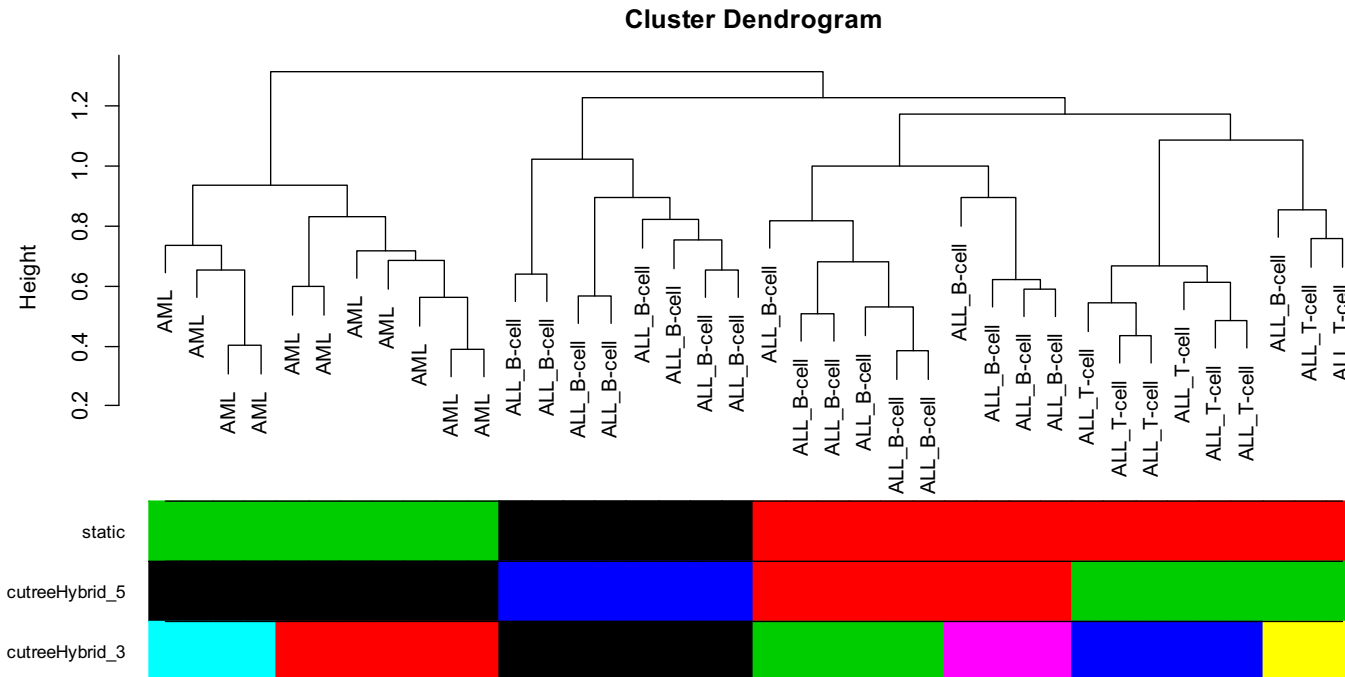
- U hierarchického shlukování se stanovuje fixní výška řezu dendrogramu `>cutree()`
- Problém: u genomických dat se často vyskytují shluky v různých výškách řezu



# Dynamic tree cut

- Metoda prořezávání dendrogramu (Langfelder et al, 2007)
- Dynamické řezání dendrogramu na základě minimální velikosti shluků, maximální výšky řezu a dalších parametrů

>library(dynamicTreeCut)



# Robustní shlukování

- V analýze vysokopokryvných molekulárních dat mají výše uvedené problémy větší váhu
- Malý počet vzorek a vysoký počet genů/proteinů spolu s vyšším množstvím šumu v datech jsou důvodem, proč je shlukování těchto dat citlivé na přeučení (overfitting)
- Shlukování je méně robustní (více ovlivněné variabilitou dat)
- Variabilita dat a výsledky shlukování se dají simulovat opakovaným náhodným výběrem z dat

# Consensus clustering

- Forma robustního shlukování (Monti et al., 2003)
  - Opakované vzorkování a shlukování jako způsob nalezení konsenzusu mezi jednotlivými výsledky shlukování za účelem:
    - Určení počtu a stability shluků v datech
    - Vytvoření nové metriky vzdálenosti - konsenzusu
  - Základní princip:
    1. Rozrušení struktury originální  $N \times P$  datové matice pomocí náhodného výběru podmnožiny vzorků a/nebo genů
    2. Na novém datovém souboru aplikujeme shlukovací algoritmus se stejnou mírou similarity a počtem shluků
- Oba body jsou opakované  $L$  krát pro jiný počet shluků.

# Consensus clustering II

- V každém výběru (pro daný počet shluků) vznikají dvě matice  $N \times N$ :
- *Matice konektivity*  $\mathbf{C}^{(l)}$  – pro každý pár vzorků  $i, j$  ukládá informaci, zda byly ve stejném shluku

$$C^l(i, j) = \begin{cases} 1 & \text{pokud } i \text{ a } j \text{ patří do stejného shluku} \\ 0 & \text{jinak} \end{cases}$$

- *Indikátorová matice*  $\mathbf{I}^{(l)}$  – pro každý pár vzorků  $i, j$  ukládá informaci, zda byly vybrány ve společném výběru

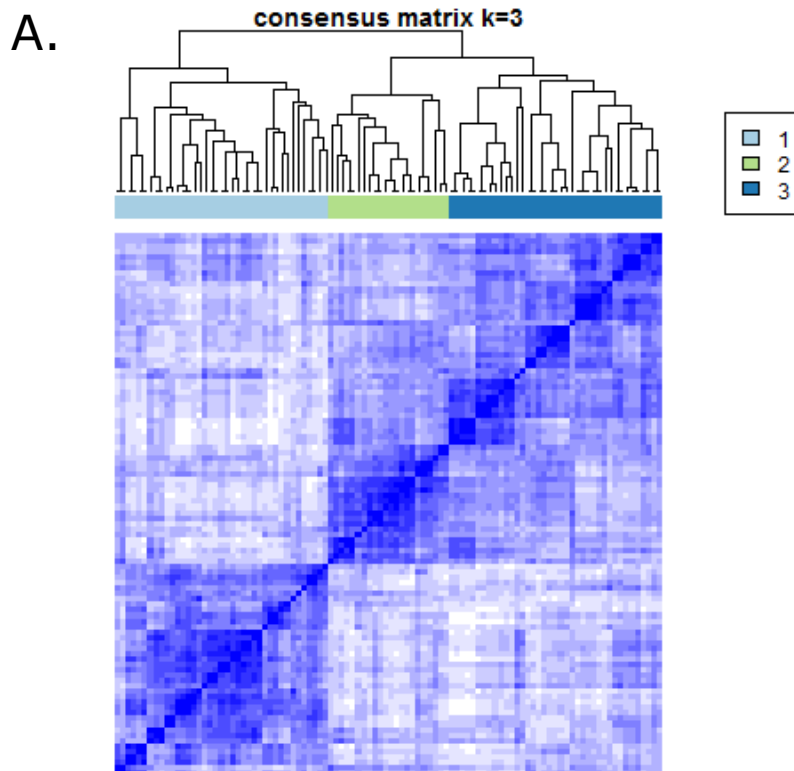
$$I^l(i, j) = \begin{cases} 1 & \text{pokud } i \text{ a } j \text{ patří do stejného výběru} \\ 0 & \text{jinak} \end{cases}$$

- **Matice konsenzusu**  $\mathbf{M}$  je definovaná jako:

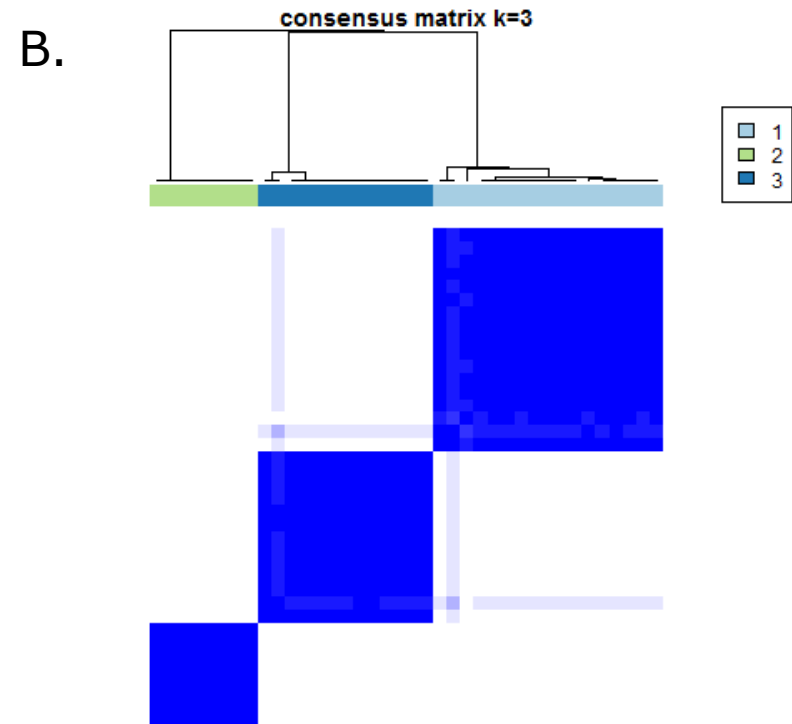
$$M_{ij} = \frac{\sum_{l=1}^L C_{ij}^{(l)}}{\sum_{l=1}^L I_{ij}^{(l)}}$$

# Consensus clustering II – myšlenka

- Pokud se dva vzorky v jednotlivých výběrech nacházejí často spolu ve shluku, jsou důvěryhodnějšími členy shluku než ty, které se ve shluku nacházejí méně často



*Data bez struktury (náhodný výběr z normálního rozložení)*



*Data se třemi skupinami*

# Consensus clustering IV – další metriky

*Konsenzus shluku  $k$*

$$m^k = \frac{1}{N_l(N_l - 1) / 2} \sum_{\substack{i, j \in I_k \\ i < j}} M_{ij}$$

*Konsenzus vzorku  $s_i$  v shluku  $k$*

$$m_i^k = \frac{1}{N_l - 1 \{s_i \in I_k\}} \sum_{\substack{j \in I_l \\ j \neq i}} M_{ij}$$

kde  $1\{s_i \in I_k\}$  je indikátorová funkcia

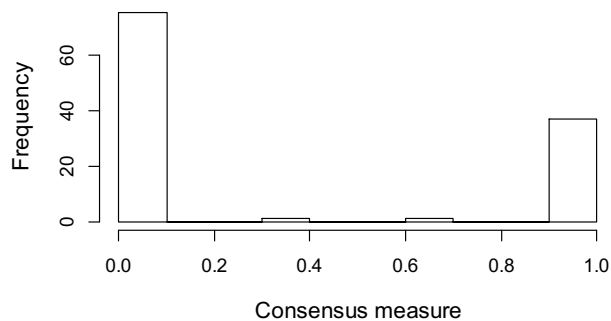
- Obě míry se používají pro identifikaci odlehlých hodnot (vzorky s nízkou mírou konsenzusu k jakémukoliv jinému vzorku v jinak homogenním shluku; shluky s nízkou mírou konsenzusu všeobecně)

# Consensus clustering V - výběr počtu shluků I

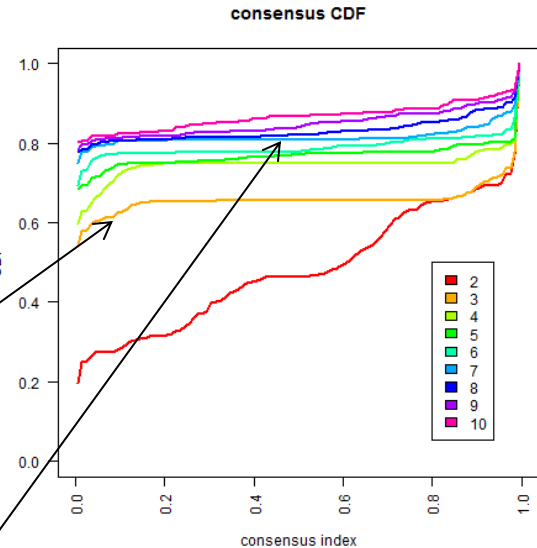
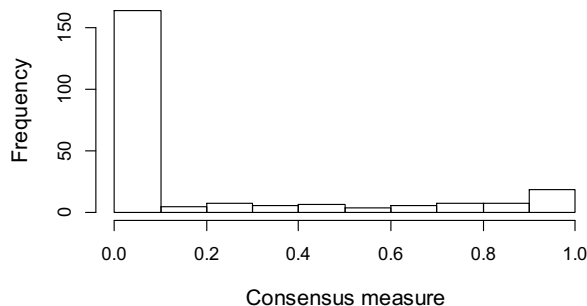
Empirická kumulativní distribuční funkce  
(pravděpodobnost, že proměnná  $M_{ij}$  nabyde hodnoty menší anebo rovné jako  $x$ )

$$CDF^x = \frac{\sum_{i < j} 1\{M_{ij} \leq x\}}{N(N-1)/2}$$

golub data, k=3



golub data, k=6



6 shluků má podstatně míň vzorků s konsenzusem 1 a tím pádem jsou tyto shluky míň důvěryhodné

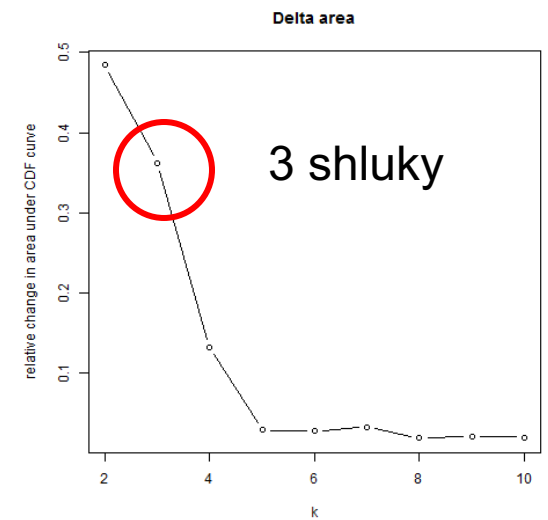
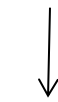
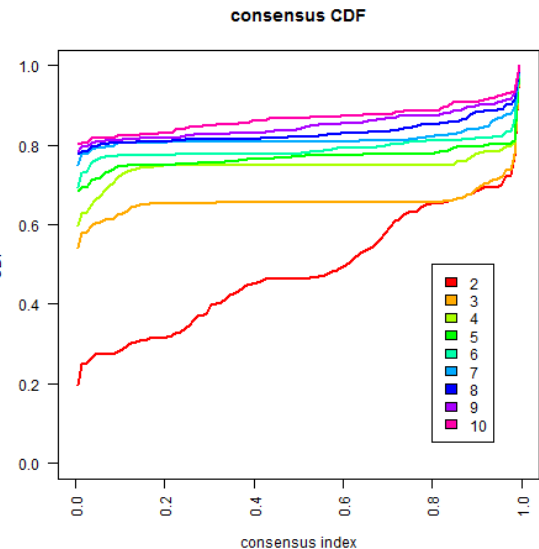
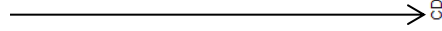
Struktura s 3 shluky naopak vypadá jako optimum

Jako rozhodovací pravidlo –  
**rozdíl v plochách pod CDF  
křivkami**



# Consensus clustering V - výběr počtu shluků II

$$CDF^x = \frac{\sum_{i < j} 1\{M_{ij} \leq x\}}{N(N-1)/2}$$



Optimální počet shluků je určený vypočtením ploch pod CDF křivkami jednotlivých počtů shluků a porovnáním relativní změny mezi různým shlukováním (plocha **delta**).

# Consensus clustering VI – R balík

---

```
> source("http://bioconductor.org/biocLite.R")  
> biocLite("ConsensusClusterPlus")
```

# Metody založené na modelech

- **Modely Gaussových směsí (mixture models)**
  - Předpokládají, že naměřené hodnoty genu/proteinu  $g$  ve všech vzorkách ( $X_g$ ) jsou náhodným výběrem a jejich rozložení závisí na skupině do které gen  $g$  patří
  - Náhodnost  $X_g$  souvisí s pozorovanou variabilitou v datech z genomických a proteomických experimentů
  - Na rozdíl od metod založených na vzdálenosti poskytují tyto modely:
    - odhad parametrů, které charakterizují každou skupinu (průměr, rozptyl, ...)
    - pravděpodobnost příslušnosti genu ke každé ze skupin
    - statistická kritéria pro výběr počtu skupin

# Modely Gaussových směsí

- Skupina  $G$  genů pochází ze smíšeného rozdělení  $K$  skupin (populací):  $C_1, \dots, C_k$ . Každý gen má marginální pravděpodobnost  $\pi_k$  ( $\sum \pi_k = 1$ ) příslušnosti ku skupině  $C_k$ .
- V závislosti na skupině, do které patří, genový/proteinový profil  $X_g$  genu  $g$  má smíšené rozdělení  $\Phi(\cdot; \theta_k)$ :

$$(X_g | g \in C_k) \sim \Phi(\cdot; \theta_k) \quad X_g \sim \sum_k \pi_k \Phi(\cdot; \theta_k),$$

kde parametr  $\theta_k$  je specifický pro skupinu  $C_k$

- Podmíněná věrohodnost  $X_g$  ( $g=1, \dots, n$ ):

$$\log \mathcal{L}(\{X_g\}; \{\pi_k, \theta_k\}) = \sum_g \log[\sum_k \pi_k \Phi(X_g, \theta_k)]$$

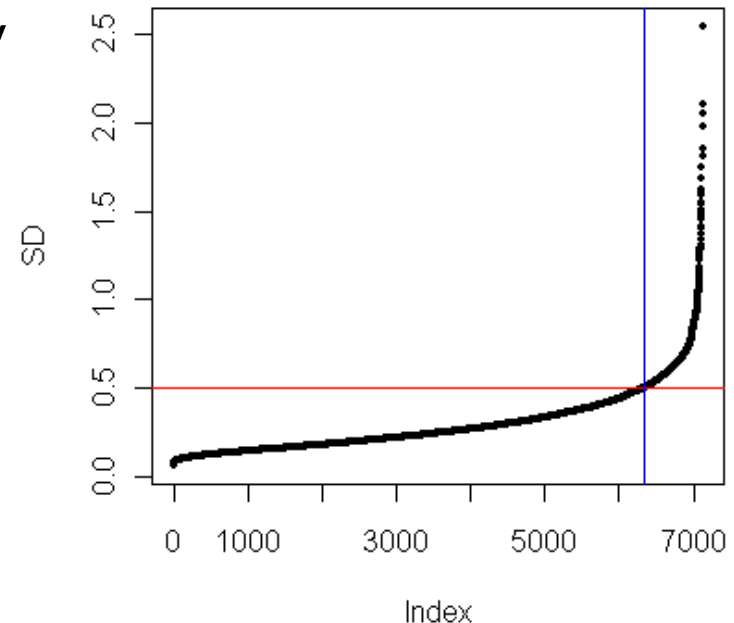
- Obvykle se uvažuje mix normálních rozložení
- Odhad parametrů a pravděpodobnosti pomocí Expectation maximization (EM)

# Pokud objekt patří do více skupin shluků

- Většina shlukovačích technik vytváří disjunktní shluky: každý objekt je součástí jediného shluku
- Toto zvláště v genomice a proteomice nemusí být nejlepší přístup, protože většina proteinů/genů je součástí více biologických drah -> proto by měli patřit do více skupin
- Jak zohlednit tuto informaci:
  - Aplikujeme speciální shlukovací metody (například fuzzy clustering)
  - Aplikujeme metody založené na modelech a vyvodíme závěry z přiřazených pravděpodobností
- Biclustering (two-way clustering) shlukuje zároveň řádky i sloupce

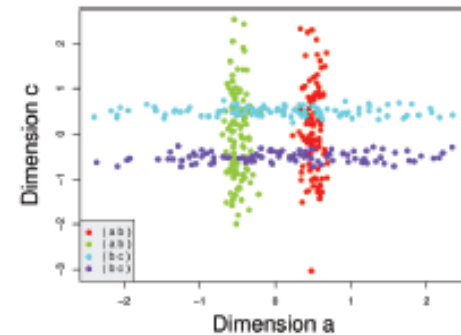
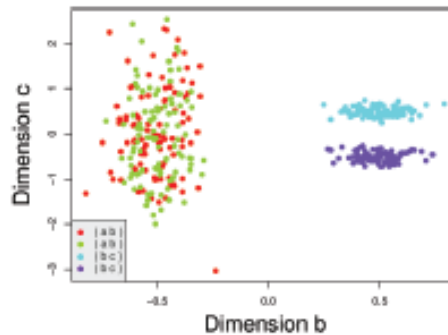
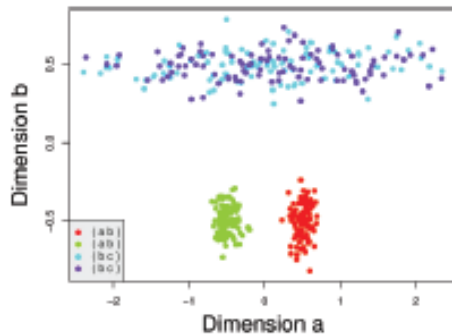
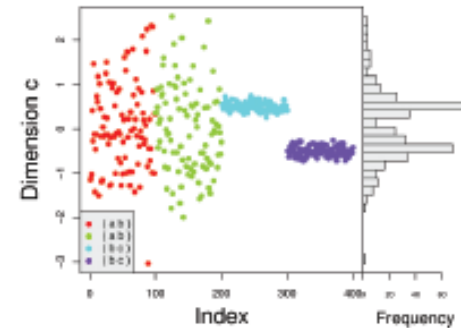
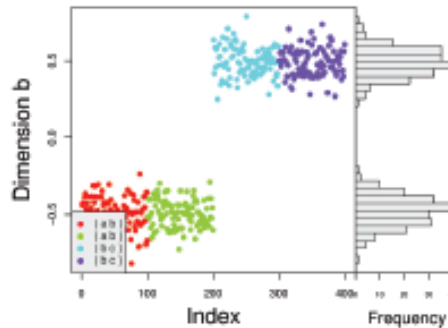
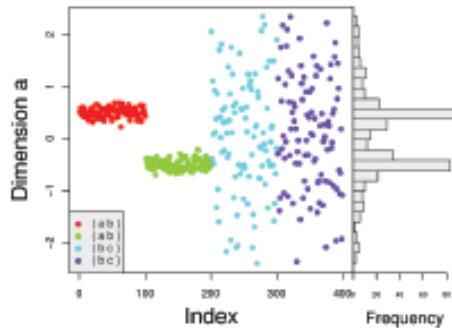
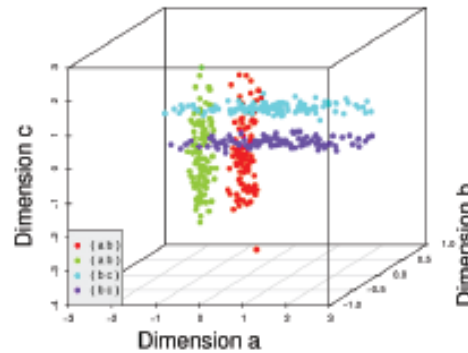
# Jak shlukovat efektivně

- V genomice a proteomice obvykle nemá význam shlukovat úplně všechny objekty (proteiny/geny)
  - Většina z nich není významná
  - Vnášejí do procesu šum, který zakryje pravou strukturu dat
- Je vhodné zredukovat dimenzi dat:
  - PCA, gene-shaving, ... - dokáží extrahovat informaci o genech/proteinech s podobnými charakteristikami, stačí potom ve shlukování reprezentovat charakteristikami těchto skupin
  - Redukce na základě SD anebo CV



# Kde hledat shluky I.

- Data můžou vytvářet shluky v odlišných dimenzích



from Giovanni Montana's presentation

# Kde hledat shluky II.

- V případě, že předpokládáme shlukování v nižších dimenzích, můžeme:
  - Hledat v nižších dimenzích vytvořených PCA
  - Použijeme podprostorové shlukovací algoritmy, které jsou schopné detekovat shluky, které existují ve více podprostorech a mohou se překrývat

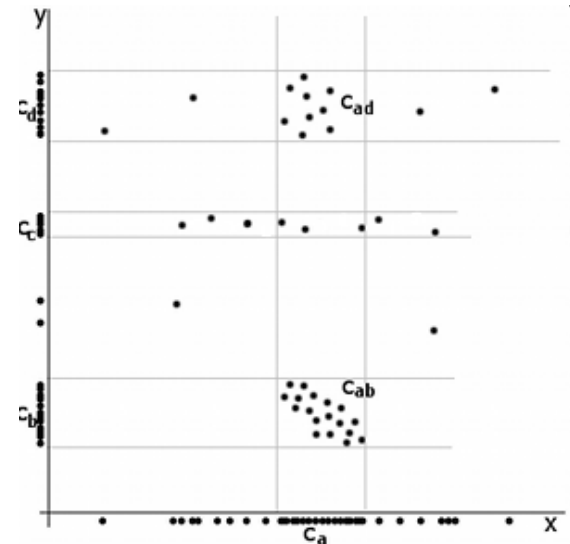


# Podprostorové shlukování

- Hledá shluky ve všech podprostorech
- Počet podprostorů je  $2^d$ , kde  $d$  je počet dimenzí (počet genů/proteinů)
- Typy algoritmů:
  - Top-down – najde iniciální rozložení na všech dimenzích a potom se dívá na podprostory každého shluku, iterativně zlepšují výsledky
  - Bottom-up – najdou regiony v nižších dimenzích a potom je zkombinují a vytvoří shluky

- MAFIA (Nagesh, 1999)
- ENCLUS (Chen, 1999)
- COSA (Damian et al., 2007)
- SMART (Jing et al., 2009)

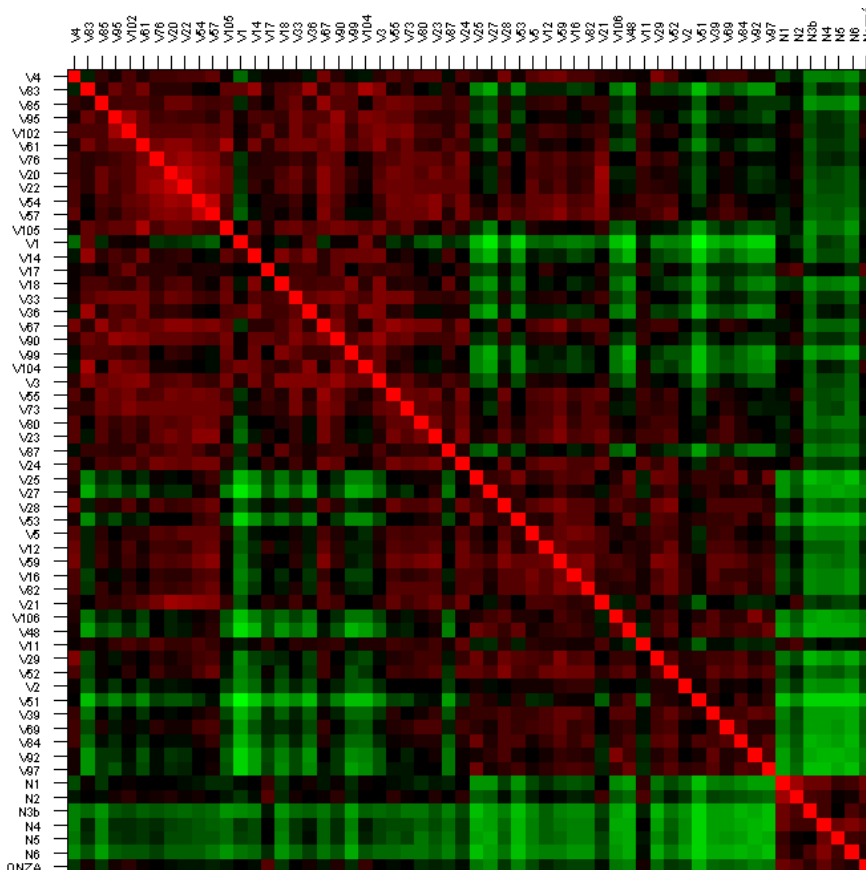
> library(orclus)



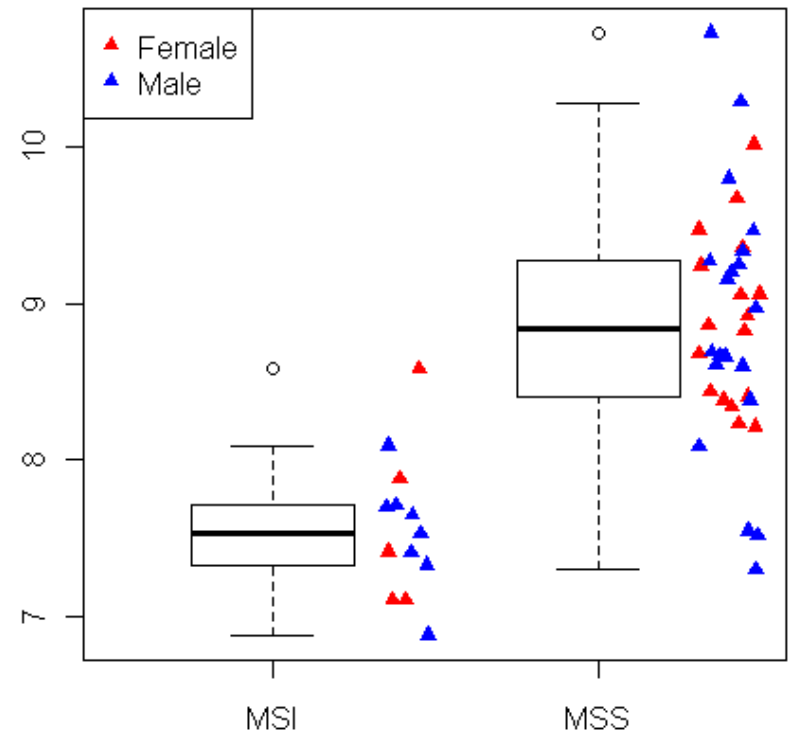
# Vizualizace výsledků

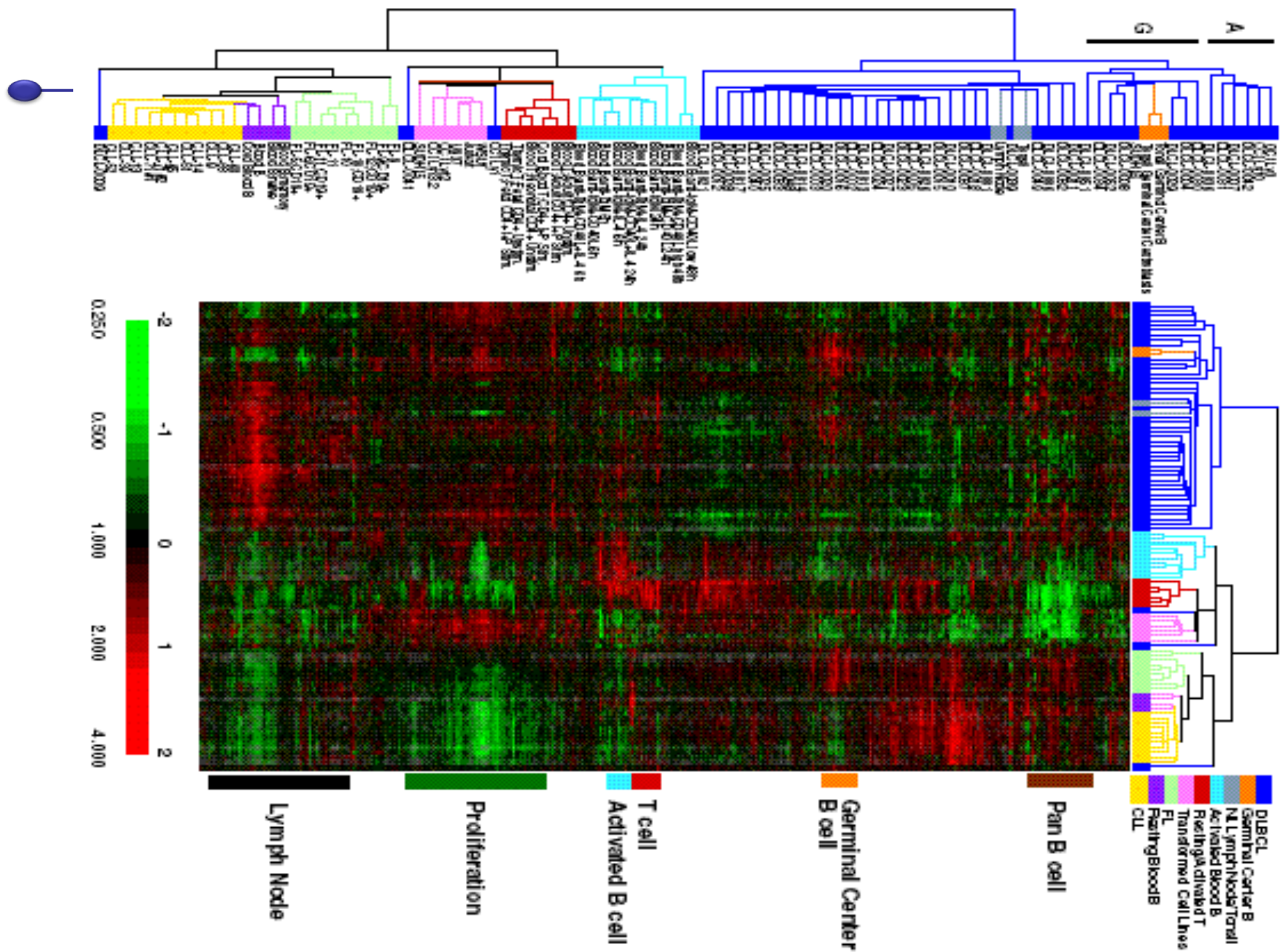
- Správná vizualizace výsledků je nejdůležitější součástí analýzy!

## Vizualizace korelací mezi vzorky



## Boxploty exprese genů





Alizadeh et al., Nature 403:503-11, 2000

# Shrnutí

---

- Více metod v rámci jedné studie
- Konsenzuální shlukování
- Dynamické řezání stromu
  
- Vizualizace výsledků
- Propojení výsledků s biologickými či klinickými proměnnými
- Validace výsledků na testovém souboru!