MUNI | RECETOX

Analýza genomických a proteomických dat

Objevování skupin

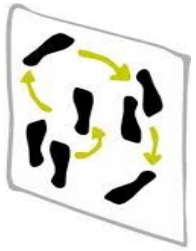
Jaro 2021
28 duben 2021
Eva Budinská (budinska@recetox.muni.cz)

MUNI | RECETOX

Společná schéma analýzy dat

Biologická otázka
(hypotéza)

Design experimentu



Provedení experimentu
(hybridizace mikročipů,
hmotnostní spektrometrie...)

N matic základních dat
(jedna pro každý z N vzorků)

Kontrola kvality
Normalizace
Sumarizace

Finální datová matice
N vzorků a K genů
(proteinů)

Matice informací o vzorcích
N x P
(např. klinická data v medicíně)

Objevování skupin?
(Shlukování)

Nové skupiny
genů nebo vzorků

Charakterizace nových
skupin

Porovnání skupin?
(Testování)

List genů
s odlišnou expresí
mezi skupinami vzorků

Predikce skupin?
(Klasifikace)

Klasifikační pravidlo
využívající
genovou expresi

Analýza přežití

Seznam
prognostických genů

Analýza časových řad

List genů
se stejným profilem
změn exprese v čase

Interpretace

Validace

Publikace

CETOX

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 případe 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 specifičtější podtypy, které by mohli být cílem cílené terapie

Společné znaky analýzy dat

- Velké množství proměnných
- Malé množství vzorků
- 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 v potaz) jakékoliv předchozí znalostí biologických skupin
- 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

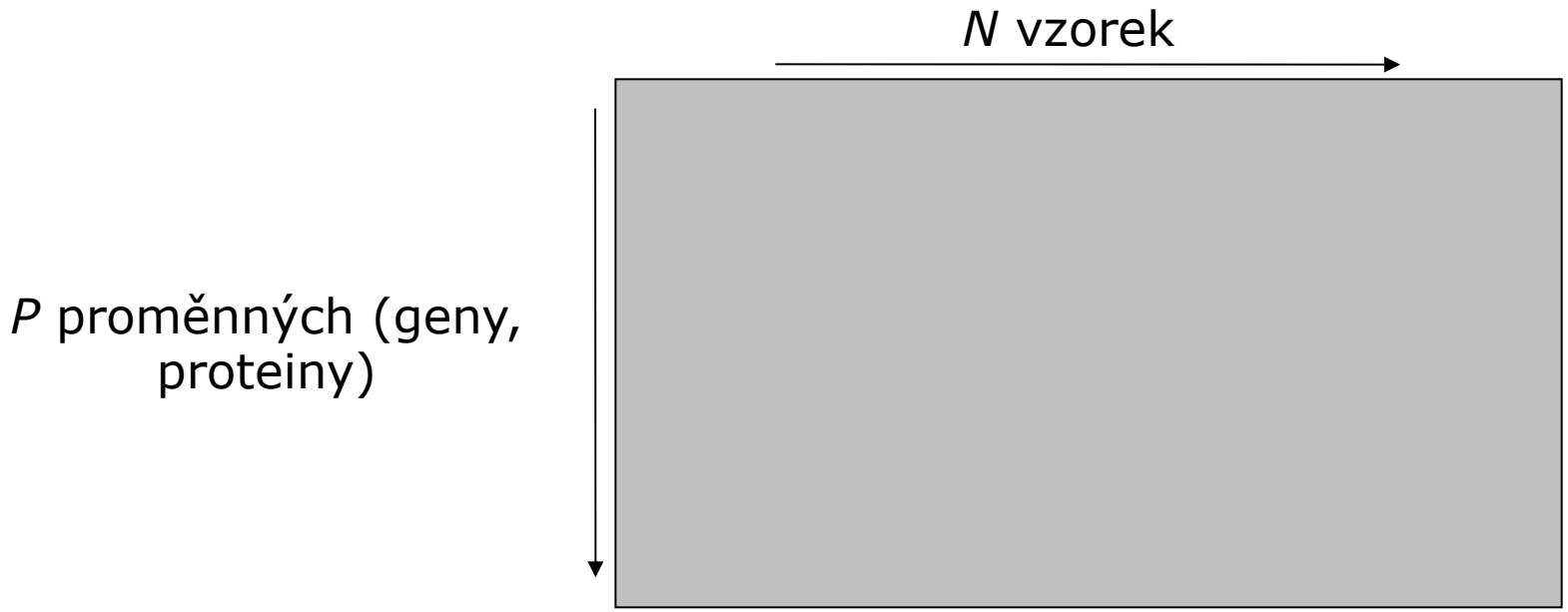
Co shlukujeme v molekulární biologii

- 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 vizualizovat diskriminační schopnost genů vybraných při porovnávání známých skupin

Princip

- Máme datovu matici ***X*** velikosti ***N x P***
  - *N* – počet objektů (vzorky)
  - *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 vzdálenosti**
   - 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. Zvolíme metriku vzdálenosti (jak vzdálenost měříme?)
  2. Vypočteme matici vzdáleností mezi objekty (každý s každým)
  3. Vybereme shlukovací algoritmus
  4. Stanovíme počet shluků – jen u některých metod
  5. Aplikujeme shlukovací algoritmus na matici vzdálenosti → získáme shluky

- Shlukovací algoritmy:
  - Hierarchické
    - Aglomerativní – Single, Complete, Average, Ward linkage, …
    - Divizivní – DIANA,...
  - Nehierarchické
    - K-means, PAM...

Metriky vzdáleností I.

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

- Euklideovská vzdálenost:

- Standardizovaná Euklideovská vzdálenost:

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

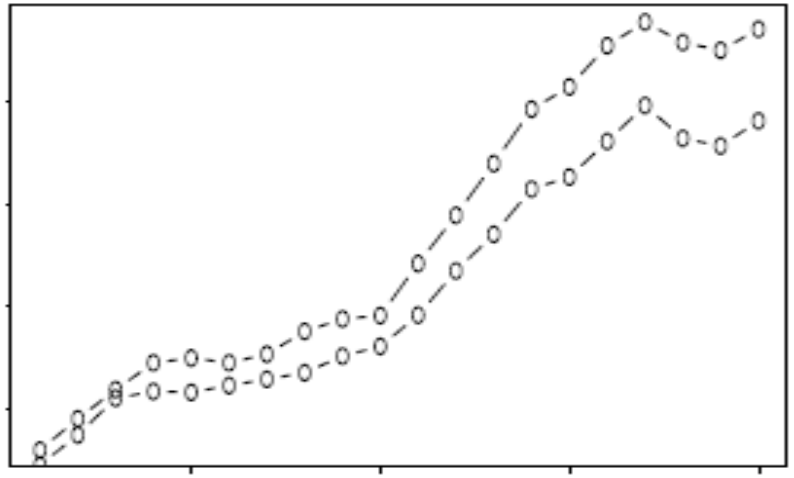
- Manhattanovská vzdálenost:

  Robustnější vůči odlehlým hodnotám.

Metriky vzdáleností II.

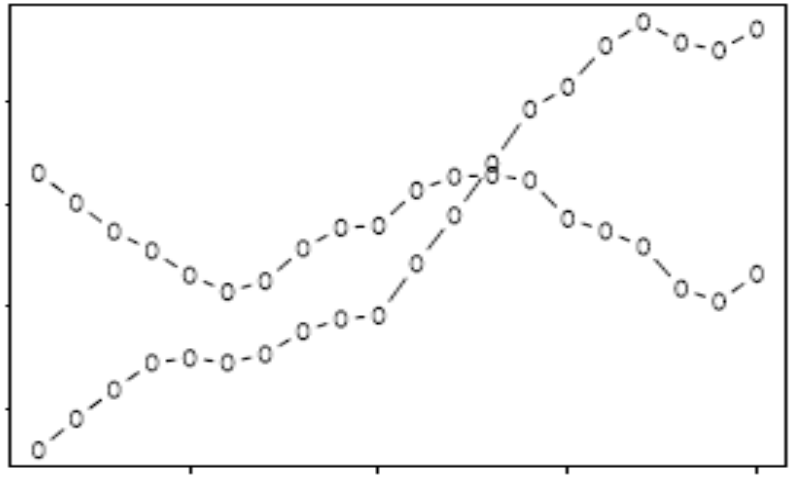
- Metriky založené **na korelačním koeficientu *r(x,y)***
- *Můžeme odvodit dvě různé metriky:*

- Ukázka rozdílu mezi metrikama



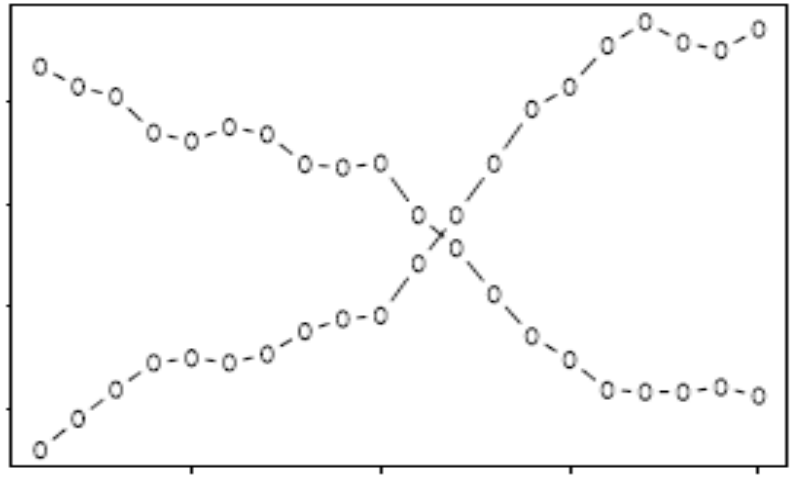
$r = 0.9$

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

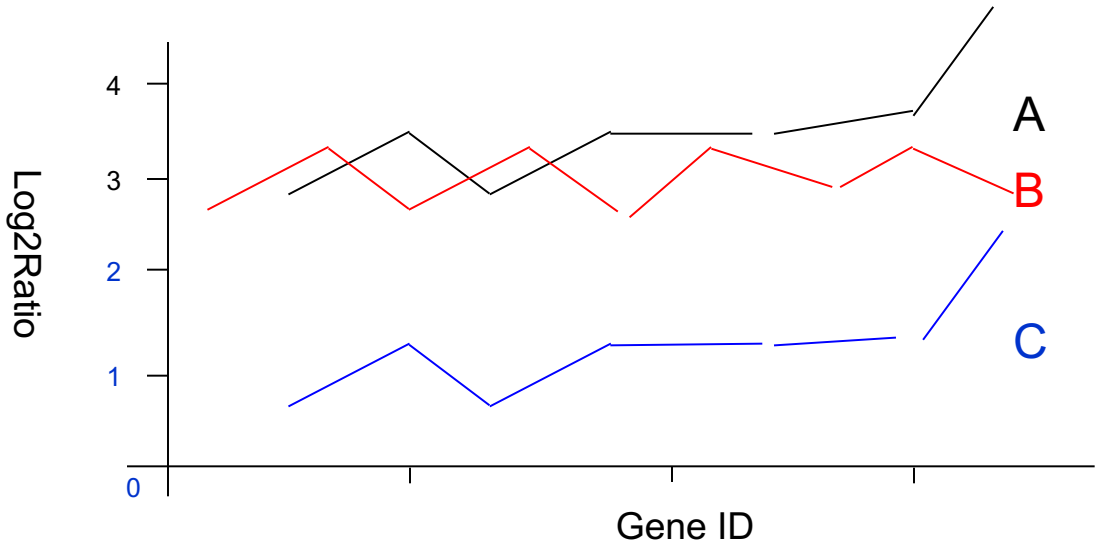
Pří použití ***d_1*** budou geny s opačnými profily patřit do odlišných shluků, zatímco při použití metriky ***d_2*** budou patřit do toho stejného shluku.
Pokud chceme shluky interpretovat jako množiny genů ze stejné regulační sítě, použijeme raději d_2.

Výběr metriky

Výběr metriky záleží na tom, jaký typ podobnosti nás zajímá

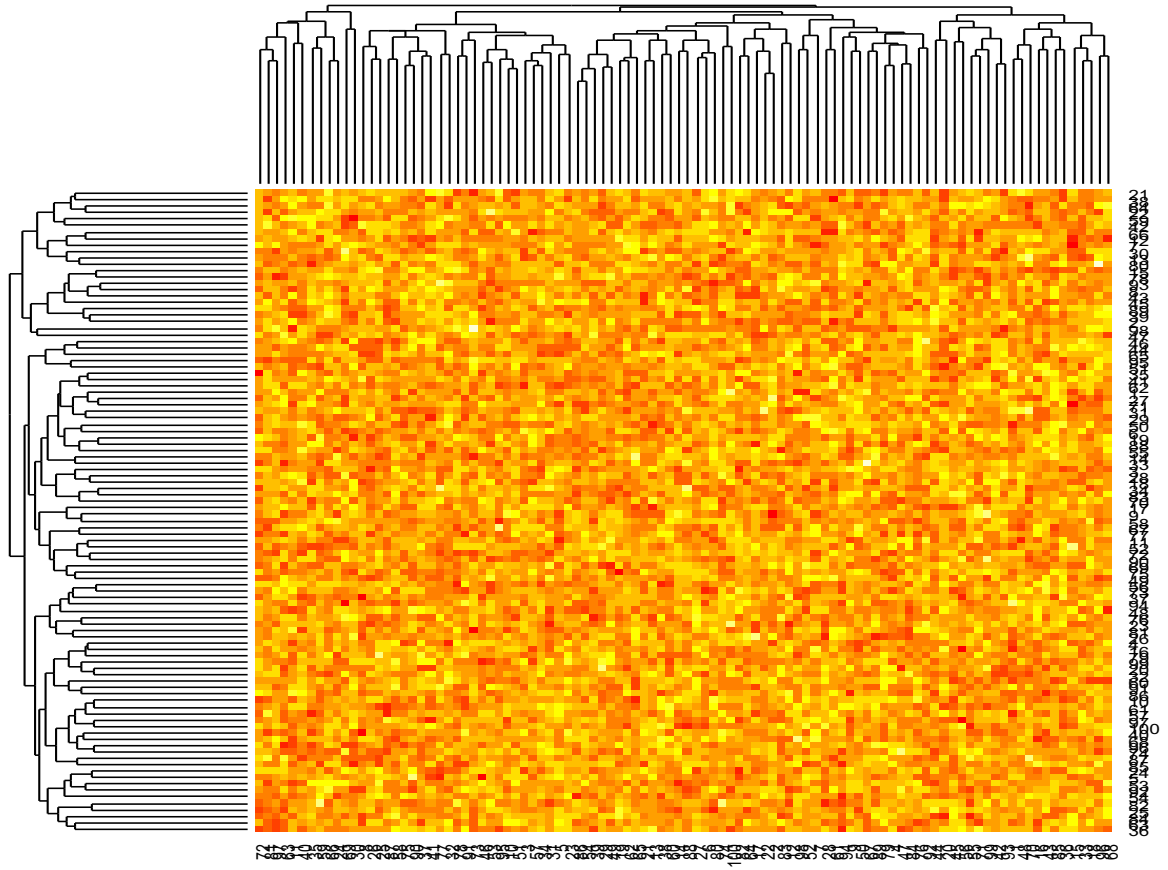
Pokud nás zajímá průměrná exprese genů (A a B jsou podobné), aplikujeme Euklidovskou vzdálenost

Pokud nás zajímá vzor exprese genů (A a C jsou podobné), aplikujeme vzdálenost založenou na korelaci



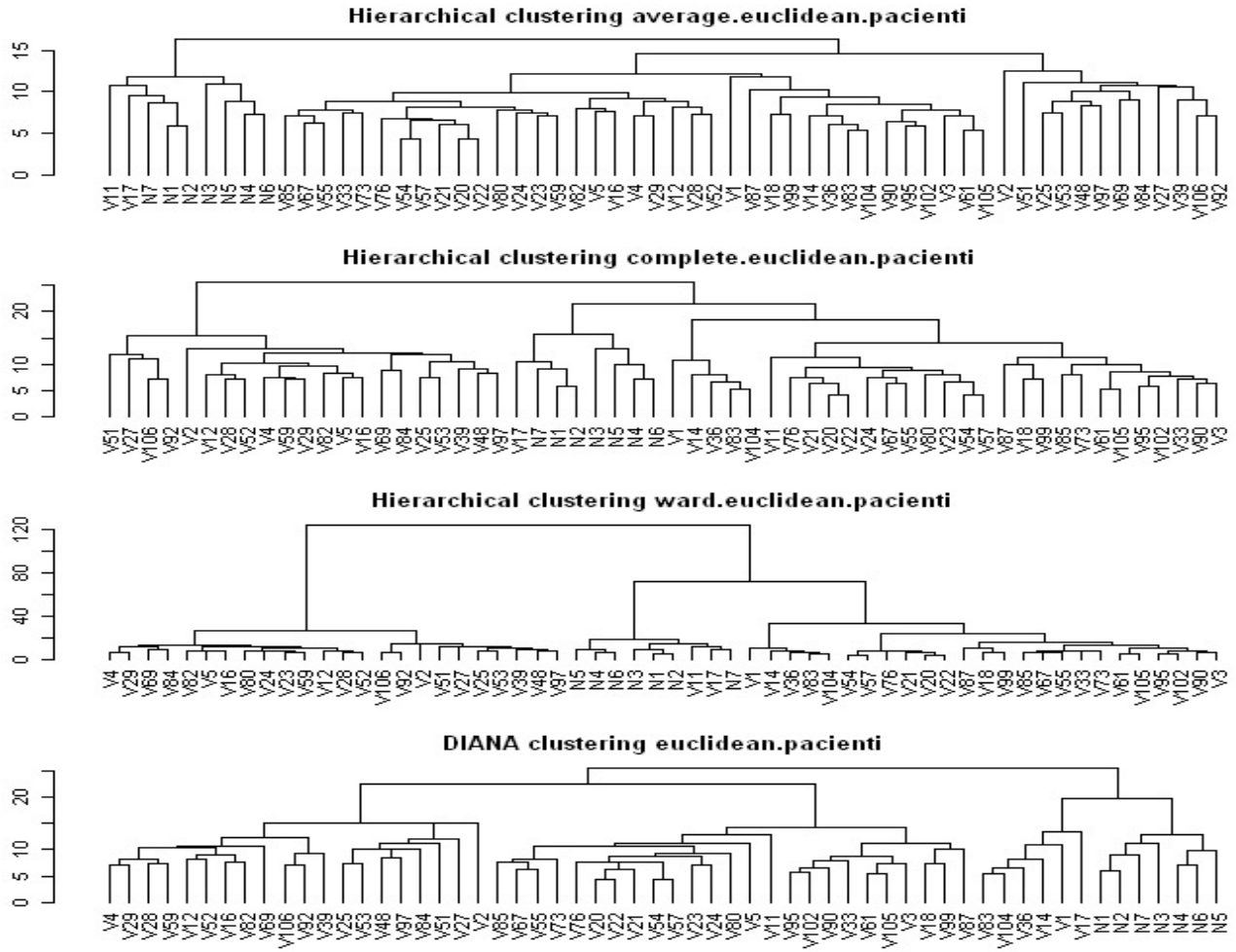
Na co si dávat pozor I.

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!

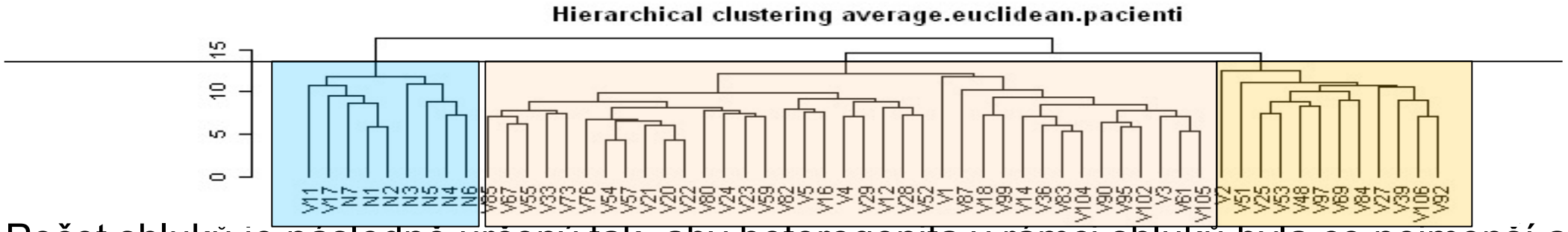


Další problémy

- 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?

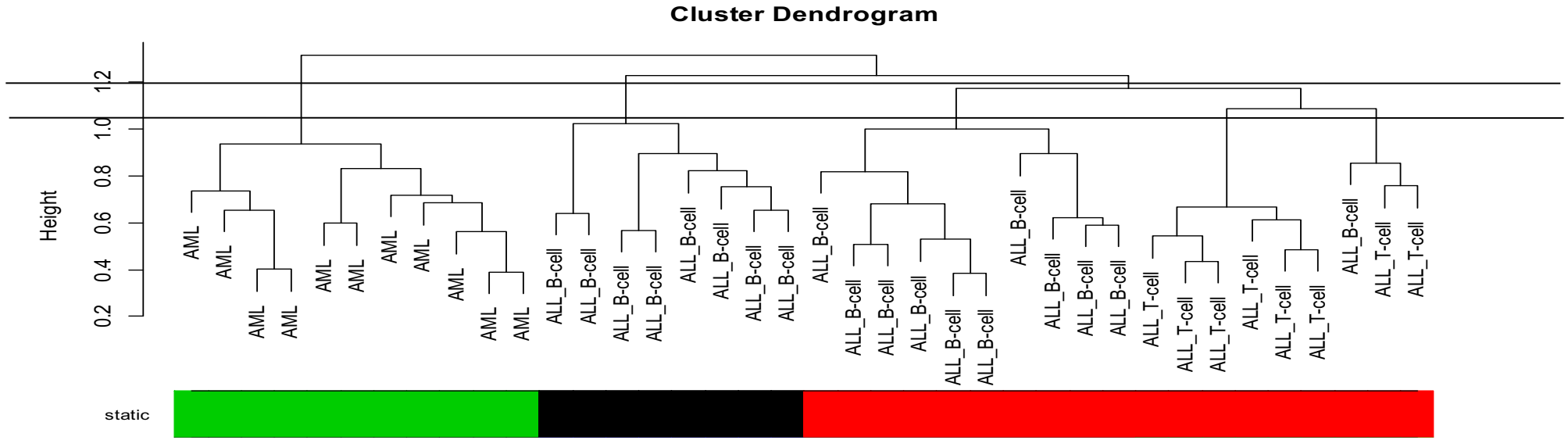
- V případě nehierarchických metod počet shluků určujeme dopředu
- V případě hierarchického shlukování vytváříme strom, dendrogram, který se potom prořezává



- 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 dendrogamu jeho problém

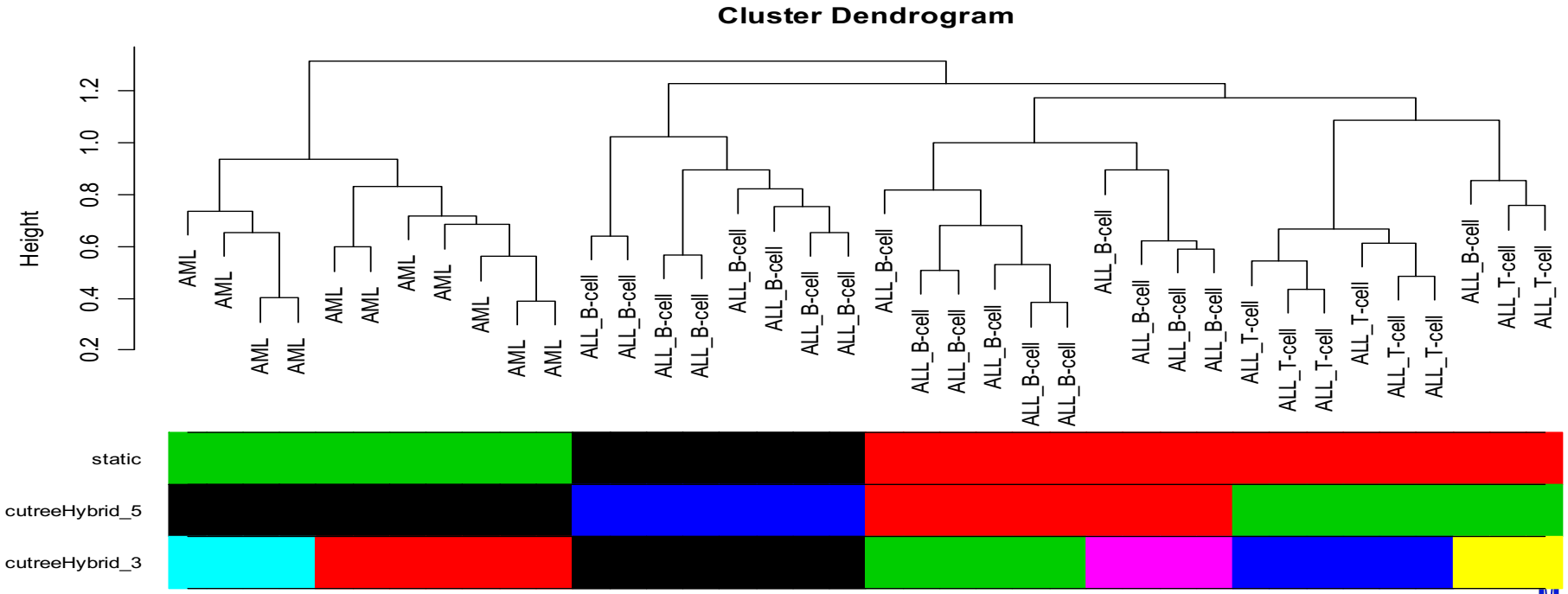
- U hiararchického shlukování se stanovuje fixní výška řezu dendrogramu
  >cutree()
- Problém: u genomických dat se často vyskytují shluky v různých výškách řezu



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

Konsenzuální shlukování

- Forma robustního shlukování (Monti et al., 2003)

- Opakované vzorkování a shlukování jako způsob nalezení **konsenzusu** mezi jednotlivými výsledkami shlukování za účelem:
  - **Určení počtu a stability shluků** v datech
  - **Vytvoření nové metriky vzdálenosti** – konsenzusu

- Základní princip:
  - Rozrušení struktůry originální N x P datové matice pomocí náhodného výběru podmnožiny vzorků a/nebo genů
  - Na novém datovém souboru aplikujeme shlukovací algoritmus se stejnou mírou similarity a počtem shluků
  - Oba body jsou opakované L krát pro jiný počet shluků.

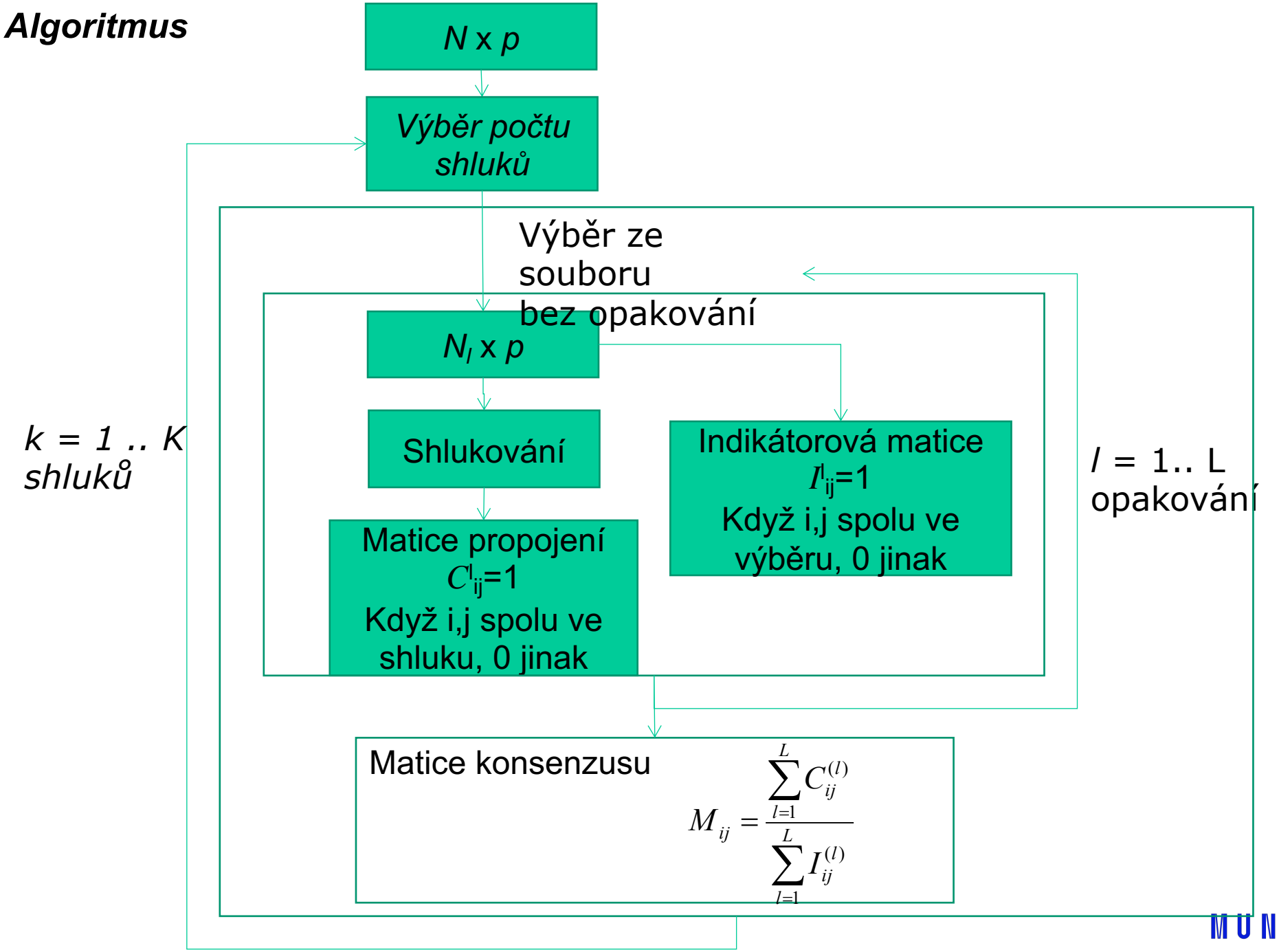
Základní princip konsenzuálního shlukování

(i) Narušení struktury originální N x P datové matice pomocí náhodného výběru podmnožiny vzorků a/nebo proměnných

(ii) Na novém datovém souboru aplikace shlukovacího algoritmu se stejnou mírou similarity a počtem shluků

(i) a (ii) opakuj L-krát pro různé počty shluků (1, …, k)

Algoritmus


N x p
Výběr počtu shluků
Výběr ze souboru bez opakování
N_l x p
Shlukování
Matice propojení C^l_ij=1 Když i,j spolu ve shluku, 0 jinak
Indikátorová matice I^l_ij=1 Když i,j spolu ve výběru, 0 jinak
Matice konsenzusu
$M_{ij} = \frac{\sum_{l=1}^{L} C_{ij}^{(l)}}{\sum_{l=1}^{L} I_{ij}^{(l)}}$
k = 1 .. K shluků
l = 1.. L opakování

Konsenzuální shlukování

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

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

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

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

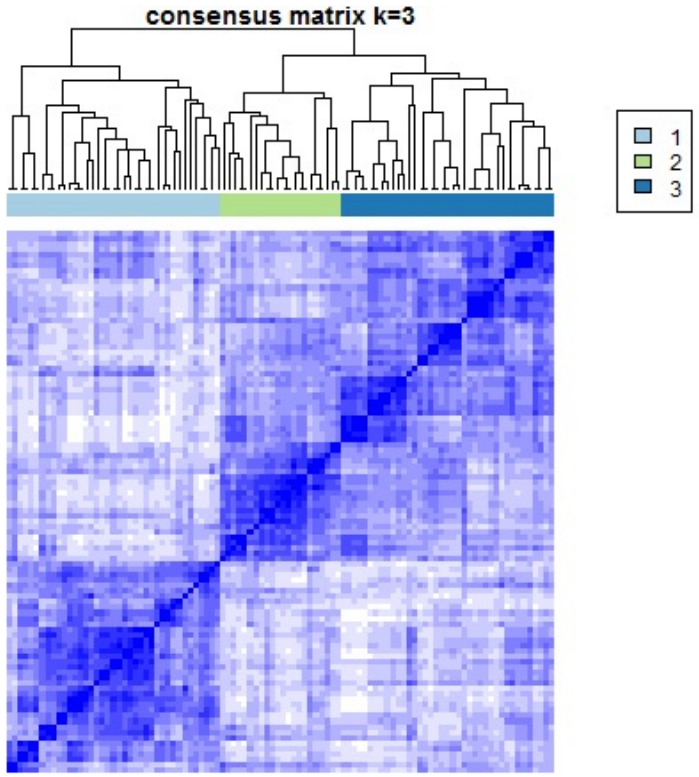
- ***Matice konsenzusu M*** je definovaná jako:

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

Myšlenka konsenzuálního shlukování

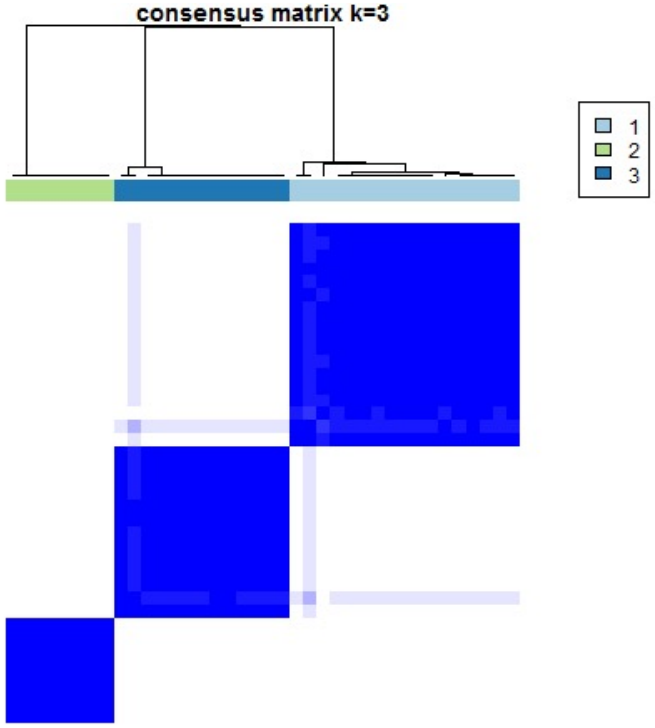
- Pokud se dva vzorky v jednotlivých výběrech nacházejí často spolu ve shluku, jsou důvěryhodnějšími členy shluku než ty, které se ve shluku nacházejí méně často

A.



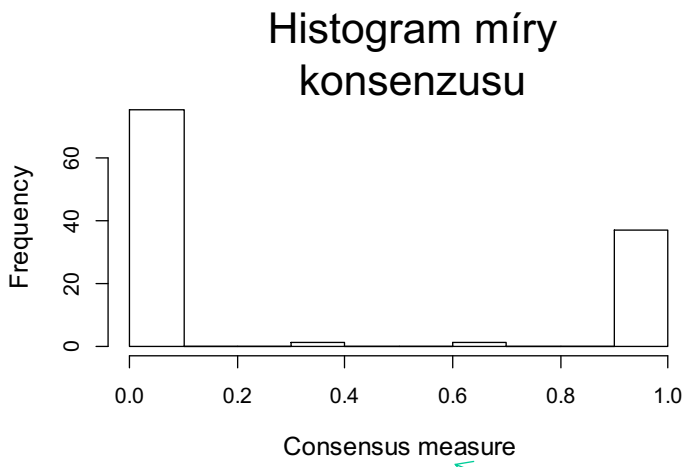
Data bez struktury (náhodný výběr z normálního rozložení)

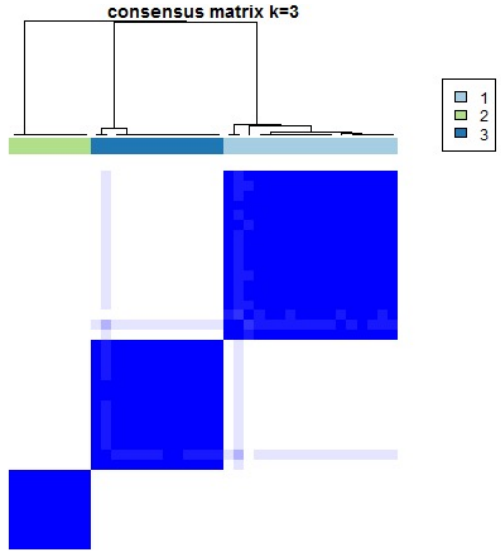
B.



Data se třemi skupinami

Odhad počtu shluků I



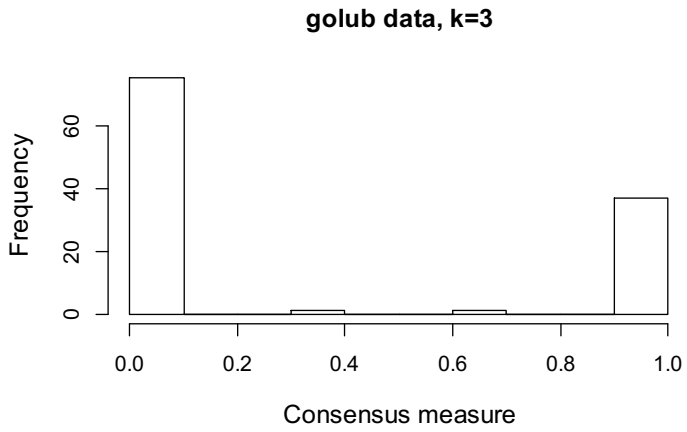


Konsenzus mezi dvěma vzorky

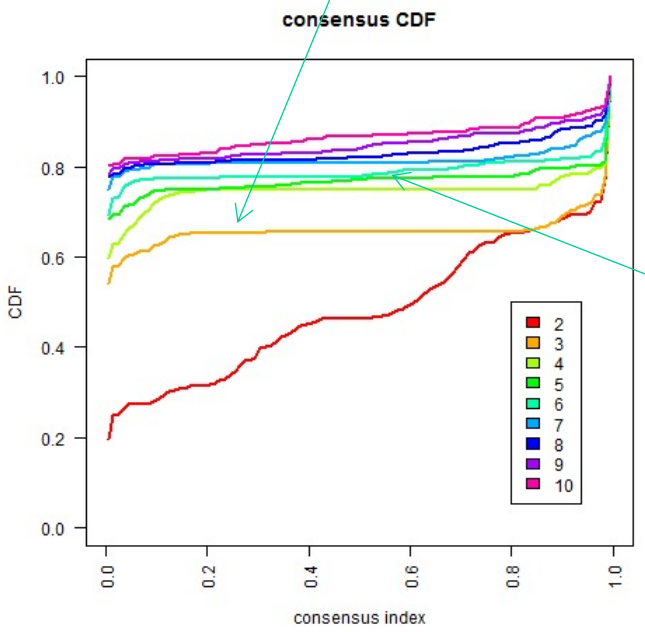
$$CDF^{x} = \frac{\sum_{i<j} 1\{M_{ij} \leq x\}}{N(N-1)/2}$$

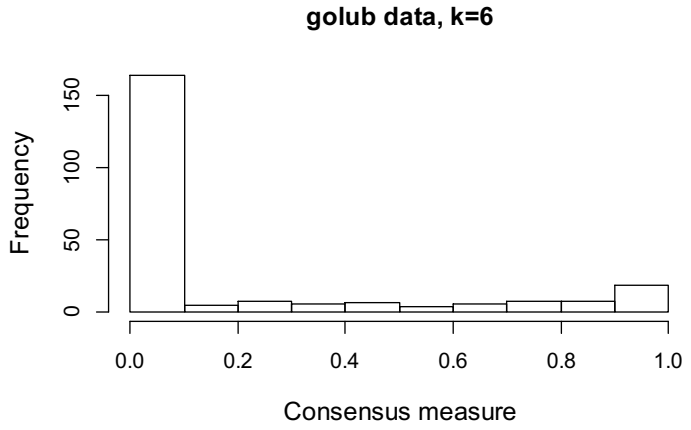
Kumulativní distribuční funkce

Odhad počtu shluků II



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





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

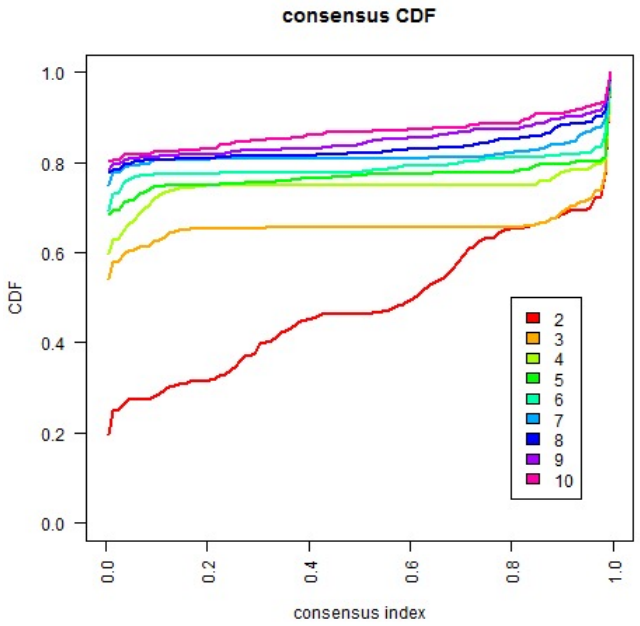
Struktura se 3 shluky naopak vypadá jako optimum

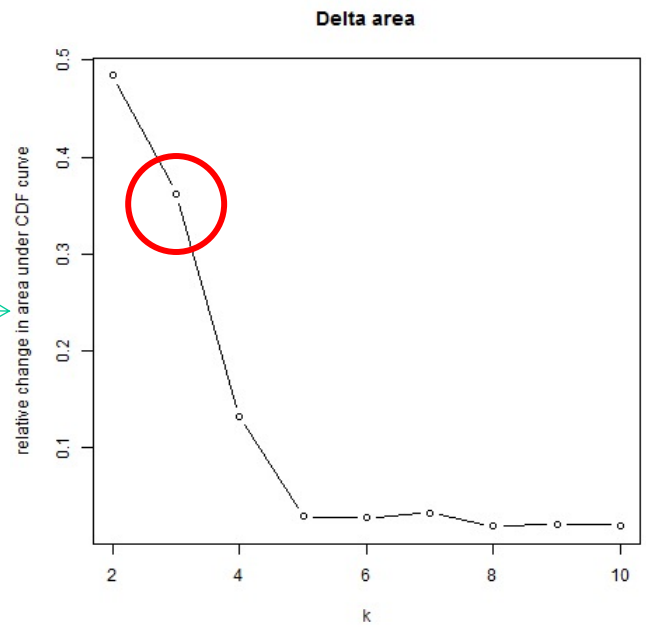
Jako rozhodovací pravidlo – **rozdíl v plochách pod CDF křivkami**

MUNI | RECETOX

Odhad počtu shluků III

Delta = relativní změna plochy
pod CDF křivkou mezi dvěma k





Další metriky konsenzuálního shlukování

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 k-tém shluku

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

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

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

Konsenzuální shlukování – 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 vzorcích (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
    - statistické kritéria pro výběr počtu skupin

Modely Gaussových směsí

- Skupina G genů pochází ze smíšeného rozdělení K skupin (populací): $C_1, \ldots, C_k$. Každý gen má marginální pravděpodobnost π_k ( ) príslušnosti ku skupine C_k.

- V závislosti na skupině, do které patří, genový/proteinový profil X_g genu g má smíšené rozdělení $\Phi(.; \theta_k)$:

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

kde parametr θ_k je specifický pro skupinu C_k

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

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

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

Pokud objekt patří do více skupin shluků

- Většina shlukovacích technik vytváří disjunktní shluky: každý objekt je součástí jediného shluku

- Toto zvlášť 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 zaráz řádky i sloupce

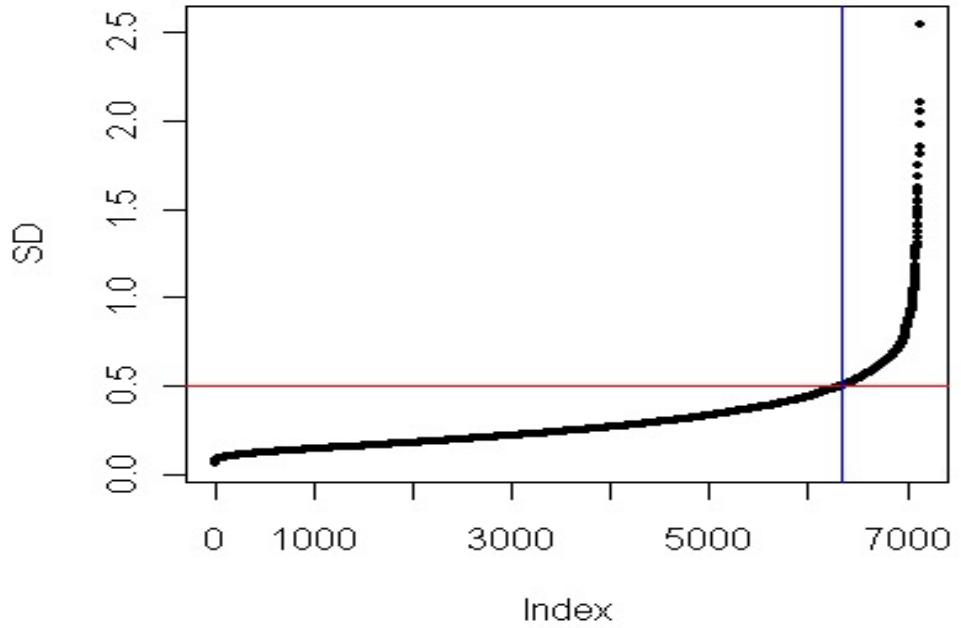
Jak shlukovat efektivně

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

<u>Je vhodné zredukovat dimenzi dat:</u>

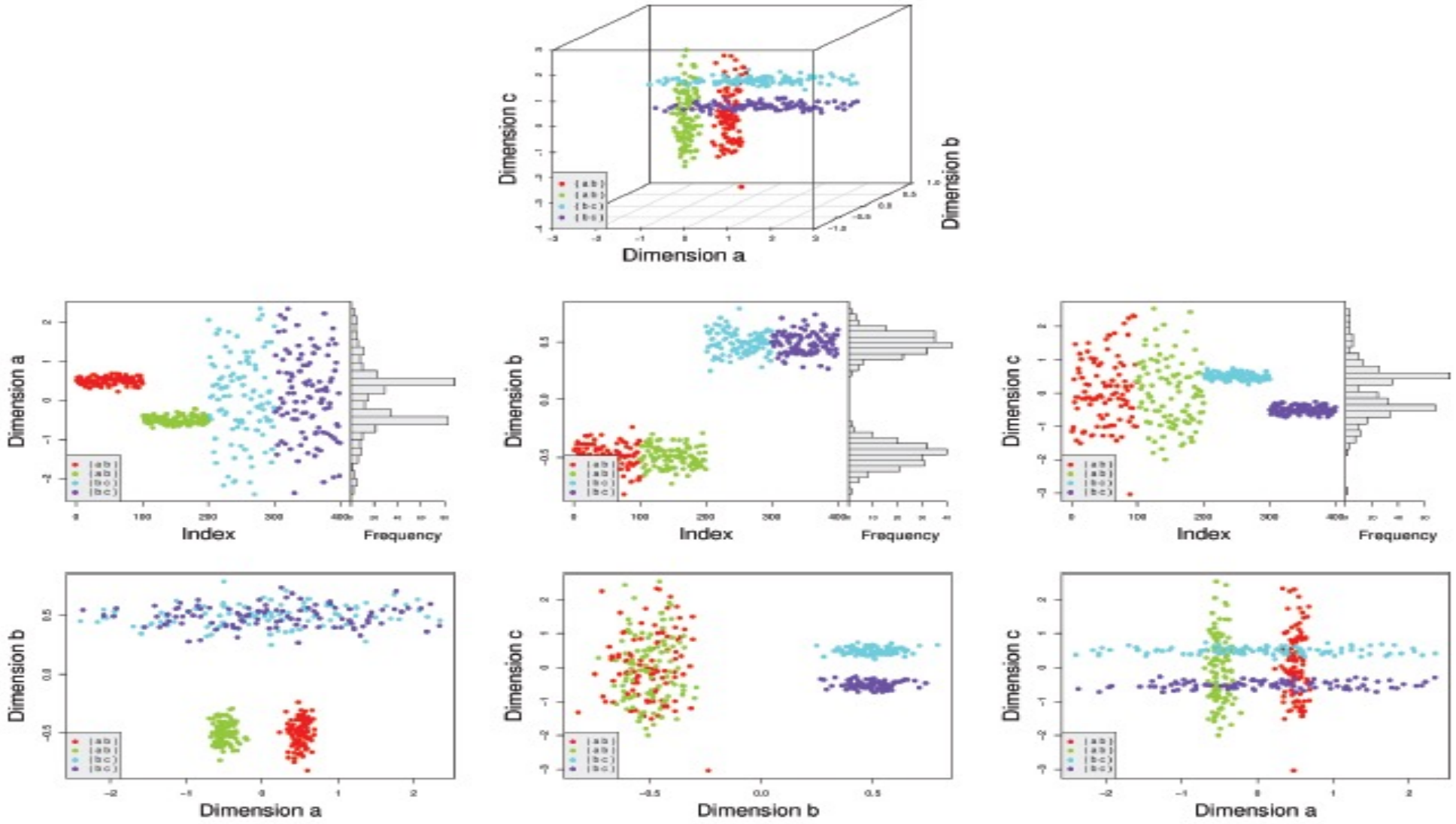
- PCA, gene-shaving, … - dokáží extrahovat informaci o genech/proteinech s podobnými charakteristikami, stačí potom ve shlukování reprezentovat charakteristikami těchto skupin
- Redukce na základě SD anebo CV



TOX

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 shlukovaní

Hledá shluky ve všech podprostorech

Počet podprostorů je 2^d, kde d je počet dimenzí (počet genů/proteinů)

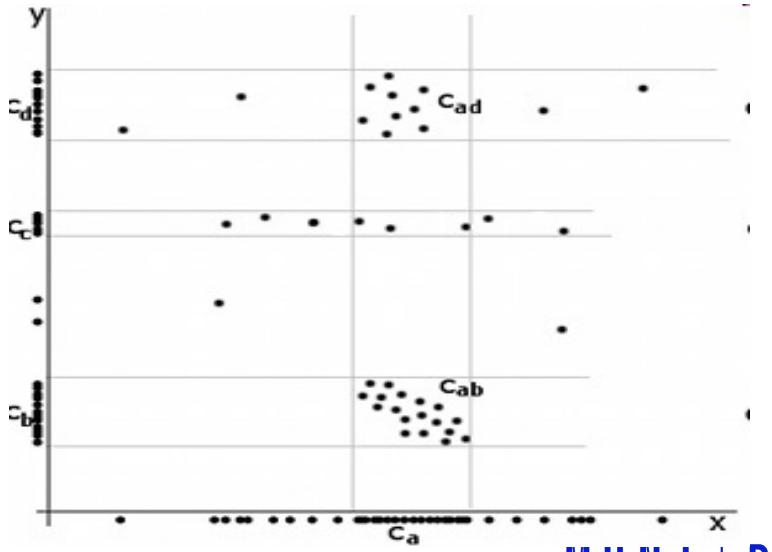
Typy algoritmů:

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)

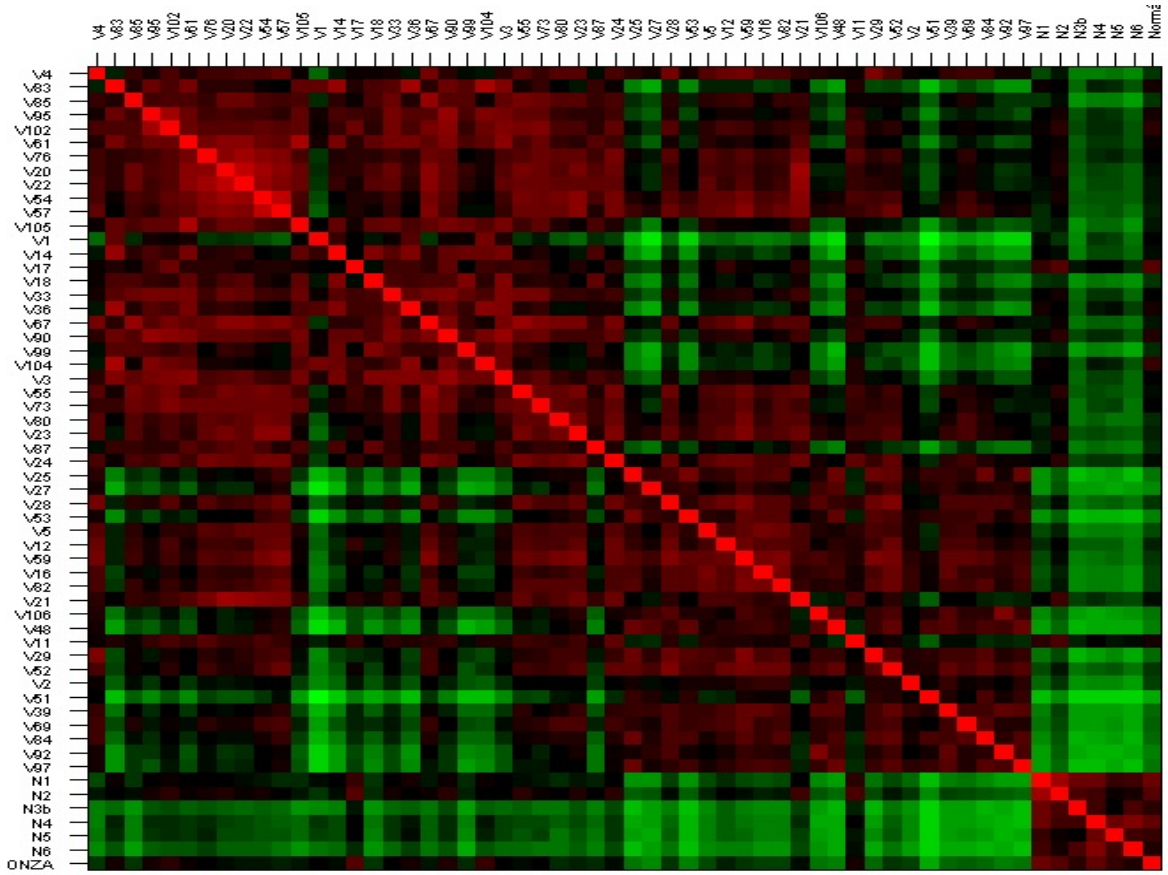
> library(orclus)



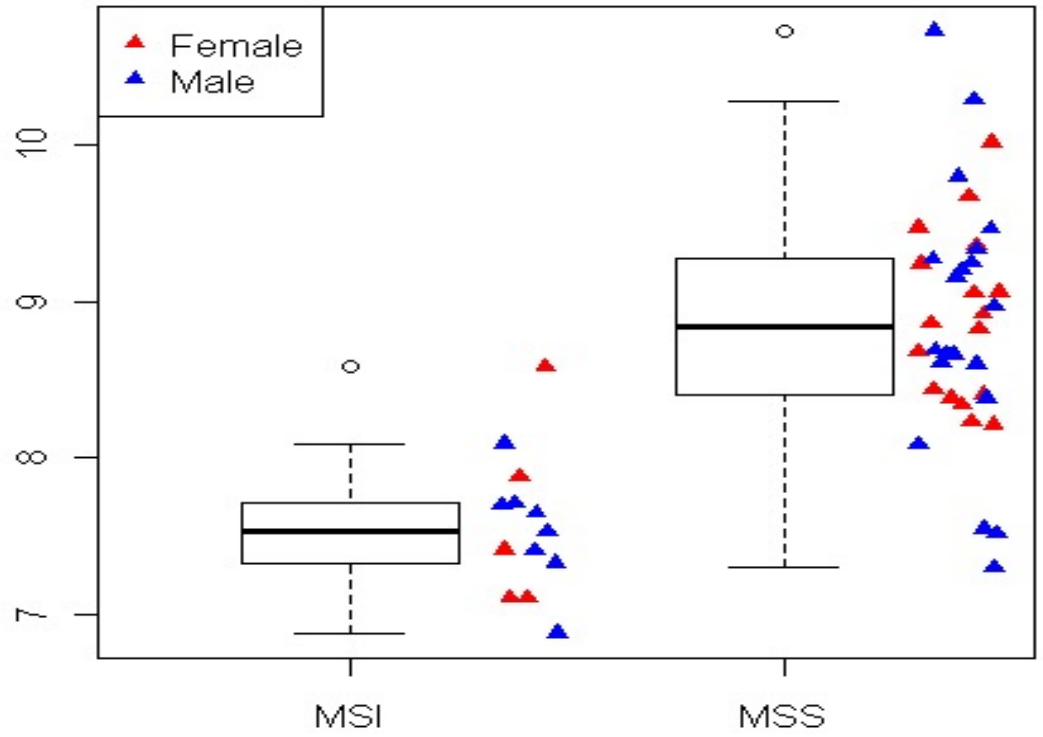
Vizualizace výsledků

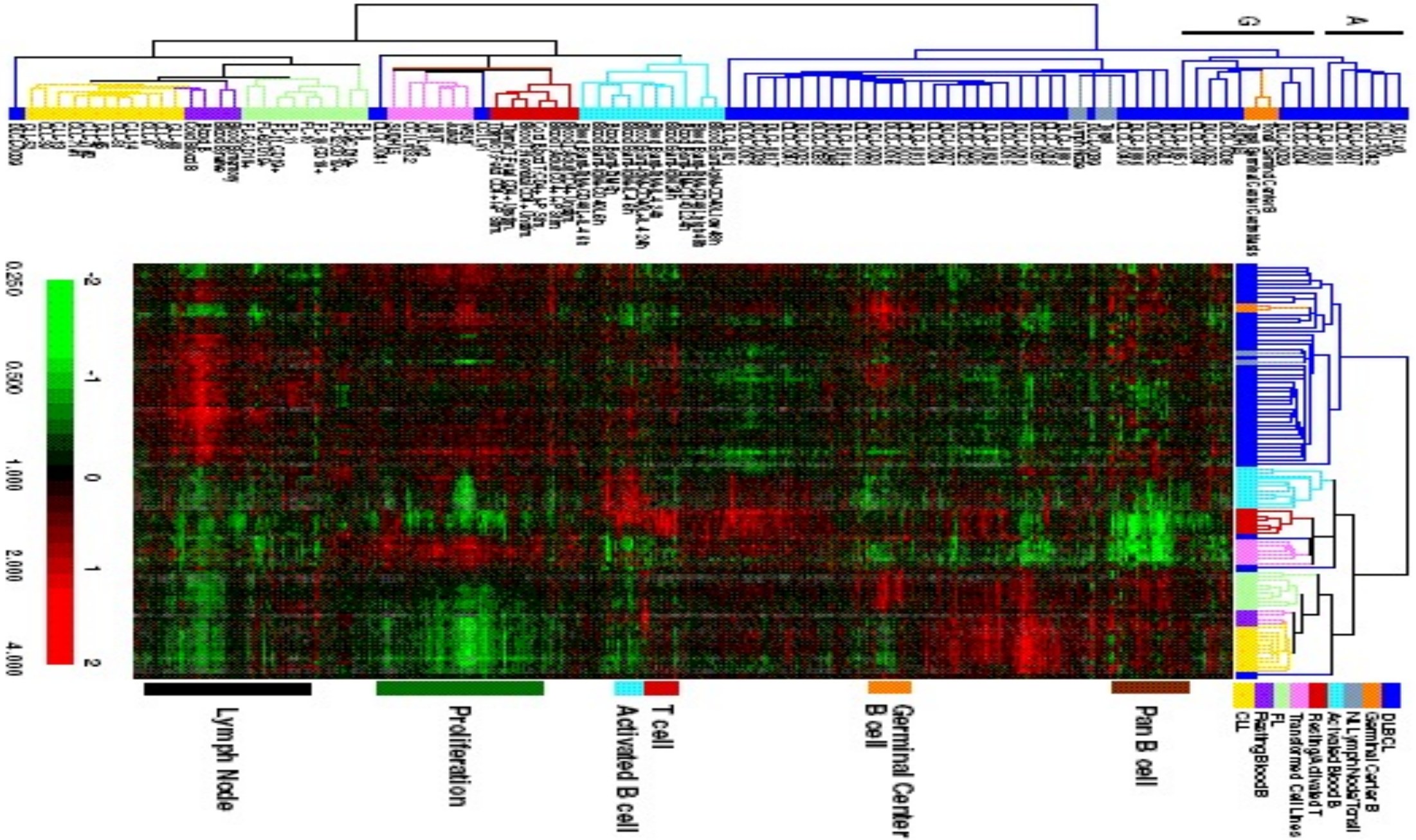
- Správná vizualizace výsledků je nejdůležitější součást analýzy!

Vizualizace korelací mezi vzorky



Boxploty exprese genů





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

Validace ve shlukování molekulárních dat

Jak validovat, když neznáme pravdu???

Validace algoritmu a parametrů modelu na testovacím souboru

(Když zopakujeme celou proceduru na dalším souboru, dostaneme stejný výsledek?)

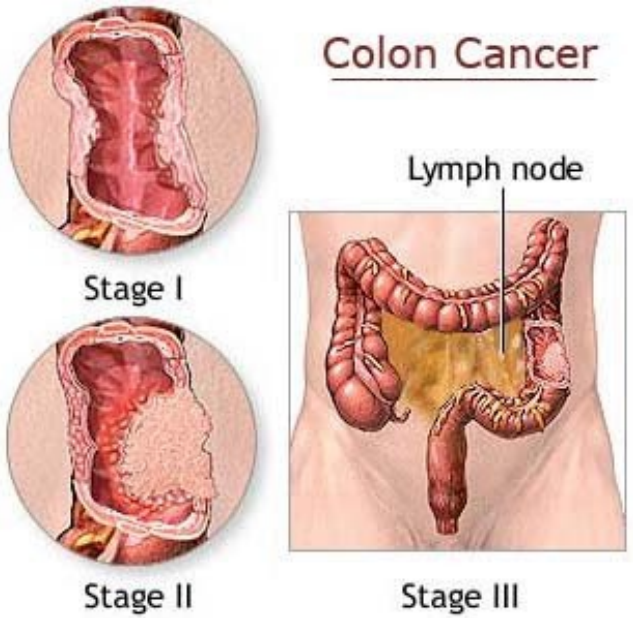
Jak validovat, když neznáme pravdu???

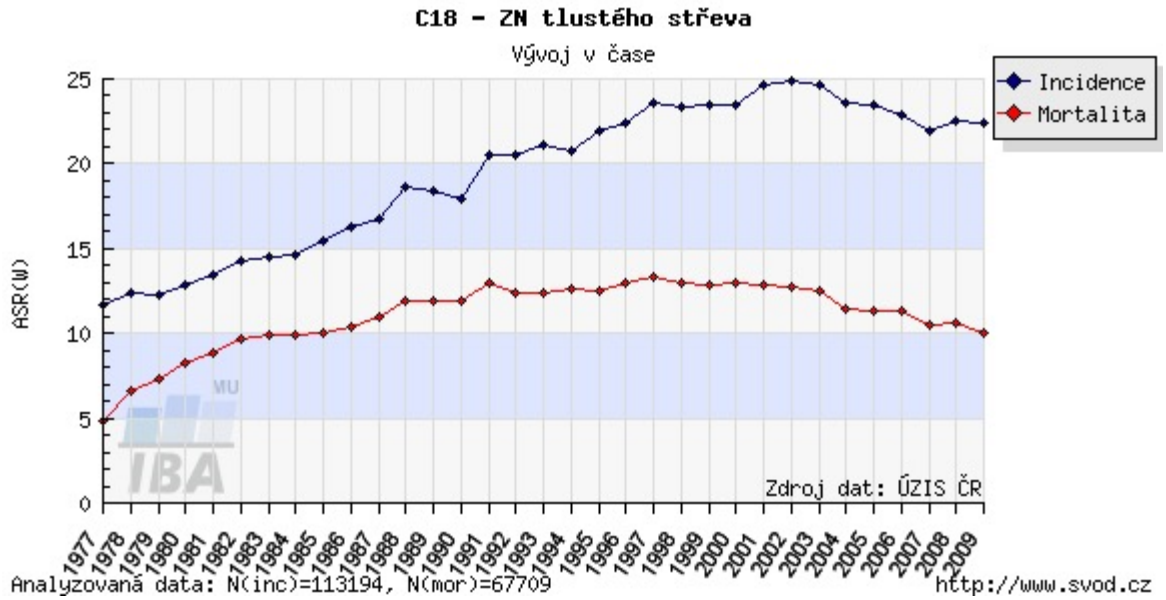
Validace konceptu pomocí klinických, molekulárních a histologických charakteristik objevených skupin

(Mají objevené skupiny biologickou podstatu / odrážejí známé vědecké poznatky?)

(Je rozložení těchto charakteristik mezi podtypy srovnatelné ve validačním souboru?)

Příklad shlukování – objevování skupin kolorektálního karcinomu





Kolorektální karcinom - Heterogenní onemocnění s rozdílnou odpovědí na terapii

Pouze několik klinicky používaných molekulárních markerů:

- *BRAF/KRAS mutace – pro kvalifikaci k antiEGFR terapii (u stadia s metastázemi)*
- *MSI – mikrosatelitová nestabilita – obecně považovaná za dobrý marker*

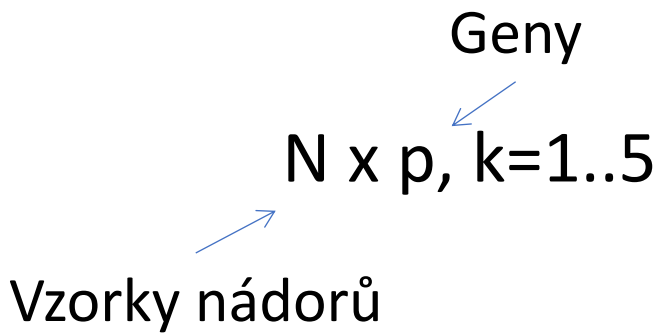
Cíl:

Nalézt skupiny nádorů kolorekta s podobnou expresí genů (podobným genovým profilem) ~ podtypy

Charakterizovat tyto podtypy pomocí klinických a dalších molekulárních parametrů.

Datové soubory:

1. Matice obsahující kvantitativní expresi genové aktivity nádorů



2. Matice klinických a molekulárních parametrů ke každé nádorové vzorce, včetně přežití pacienta

VÝBĚR PROMĚNNÝCH / Zmenšení dimenze dat

Nezávislý výběr – z 25 000 genů, výběr podmnožiny 3025 genů s nejvyšší variabilitou v souboru

Redukce dimenzionality - práce s genovými moduly

genový modul - sada genů s podobnou genovou expresí (na základě korelace)

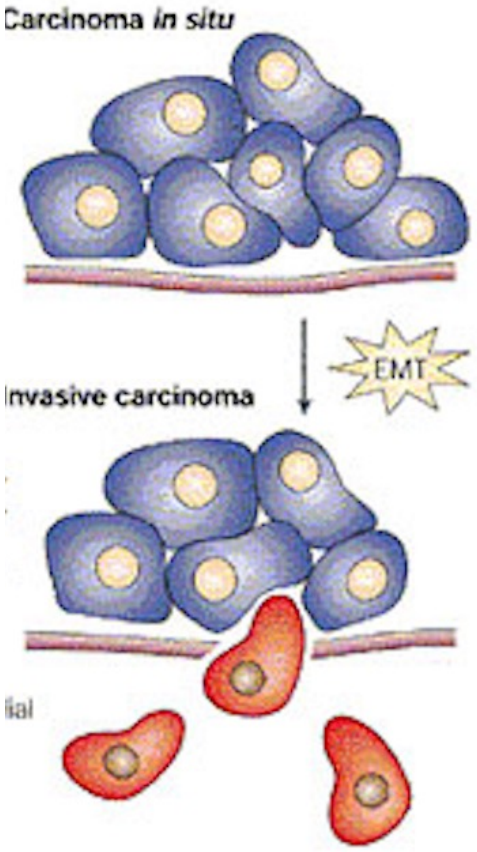
Jedná se o jistou formu **vážení efektu biologických motivů**

předpoklad:

sada korelovaných genů ~ biologický motiv

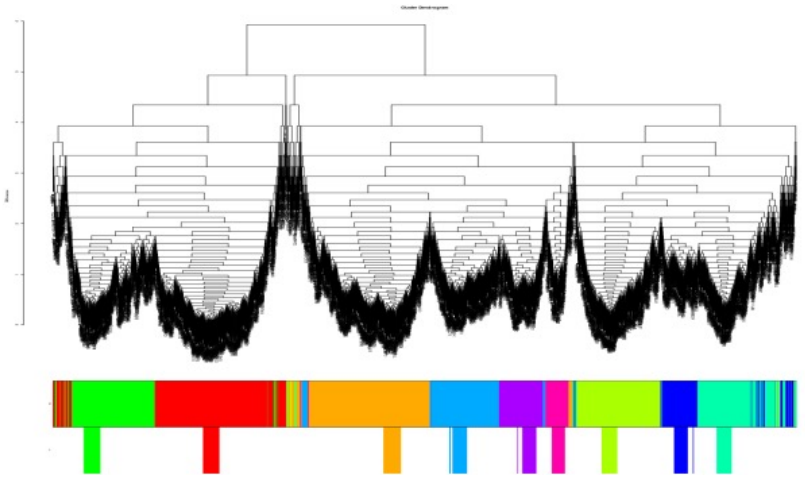
Příklad EMT

- EMT - epiteliálně mesenchymální přechod
- Genová exprese podobná zdravé mezenchymální tkáni
- Obvykle reprezentován změnou exprese stovek genů
- Identifikace modulu EMT a jeho reprezentace jedinou hodnotou (průměrem) zmenší jeho efekt v shlukování a dá šanci dalším důležitým procesem reprezentovaným menším množstvím genů



3025 genů
Datový soubor 1

Pearsonova korelace,
Hierarchické shlukování
Complete linkage
řezání dendrogramu



150 genových modulů

Validace jejich korelací v
4 dalších datových
souborech

Výběr 54 modulů
(625 genů)

Medián exprese genů v
modulu

54 meta-genů

8 hlavních shluků
~ Hlavní biologické
motivy

Pearsonova korelace,
Hierarchické shlukování
complete linkage
řezání dendrogramu

Identifikace biologických motivů

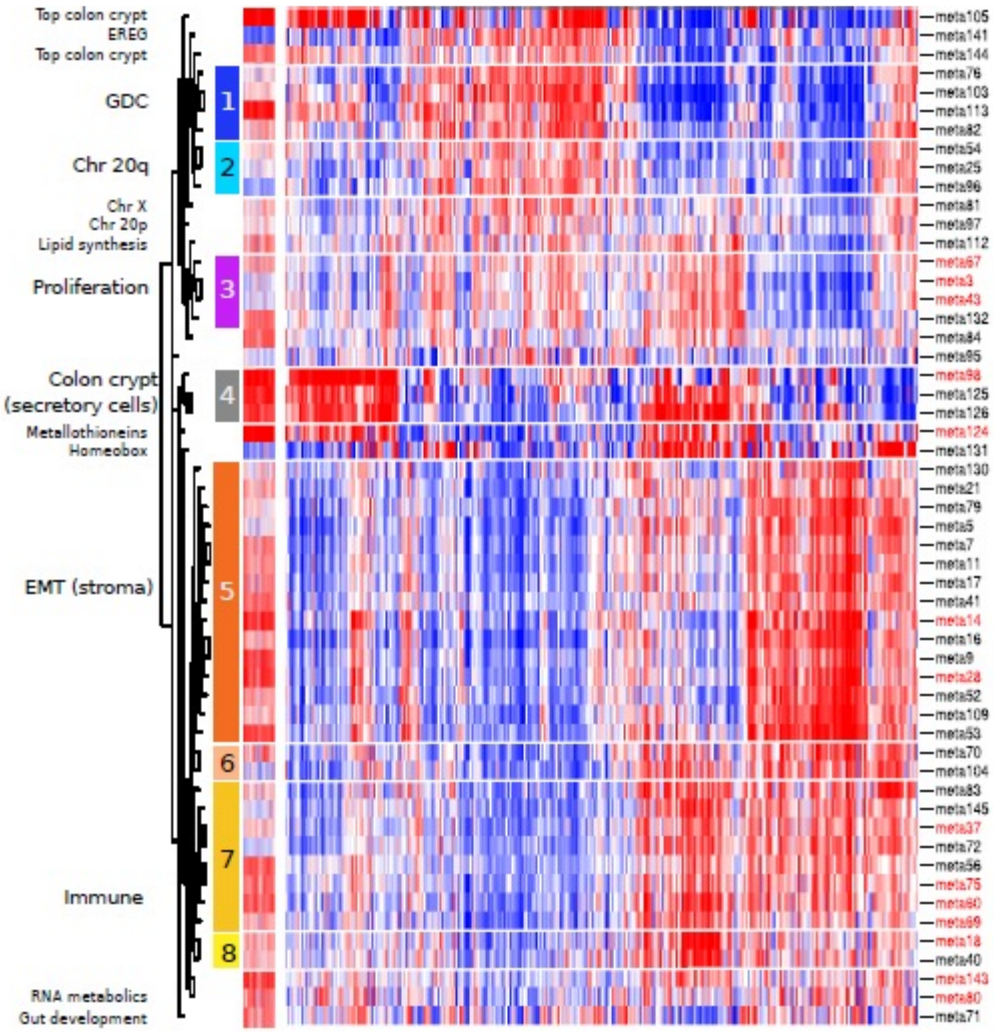
- Analýza genových sad

	V genomu	V biol motivu
V genomu	a	b
V modulu	c	d

Biol motiv:

EMT, proliferace,

Chromozomm 20q...



Cíl:

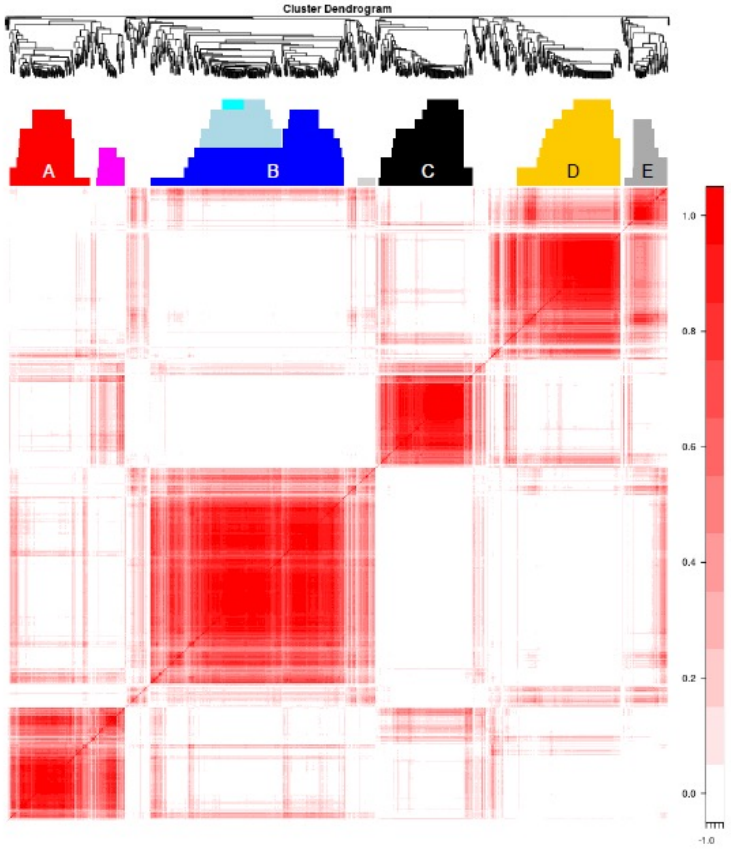
Najít skupiny nádorů kolorekta s podobnou expresí genů (podobným genovým profilem) ~ podtypy

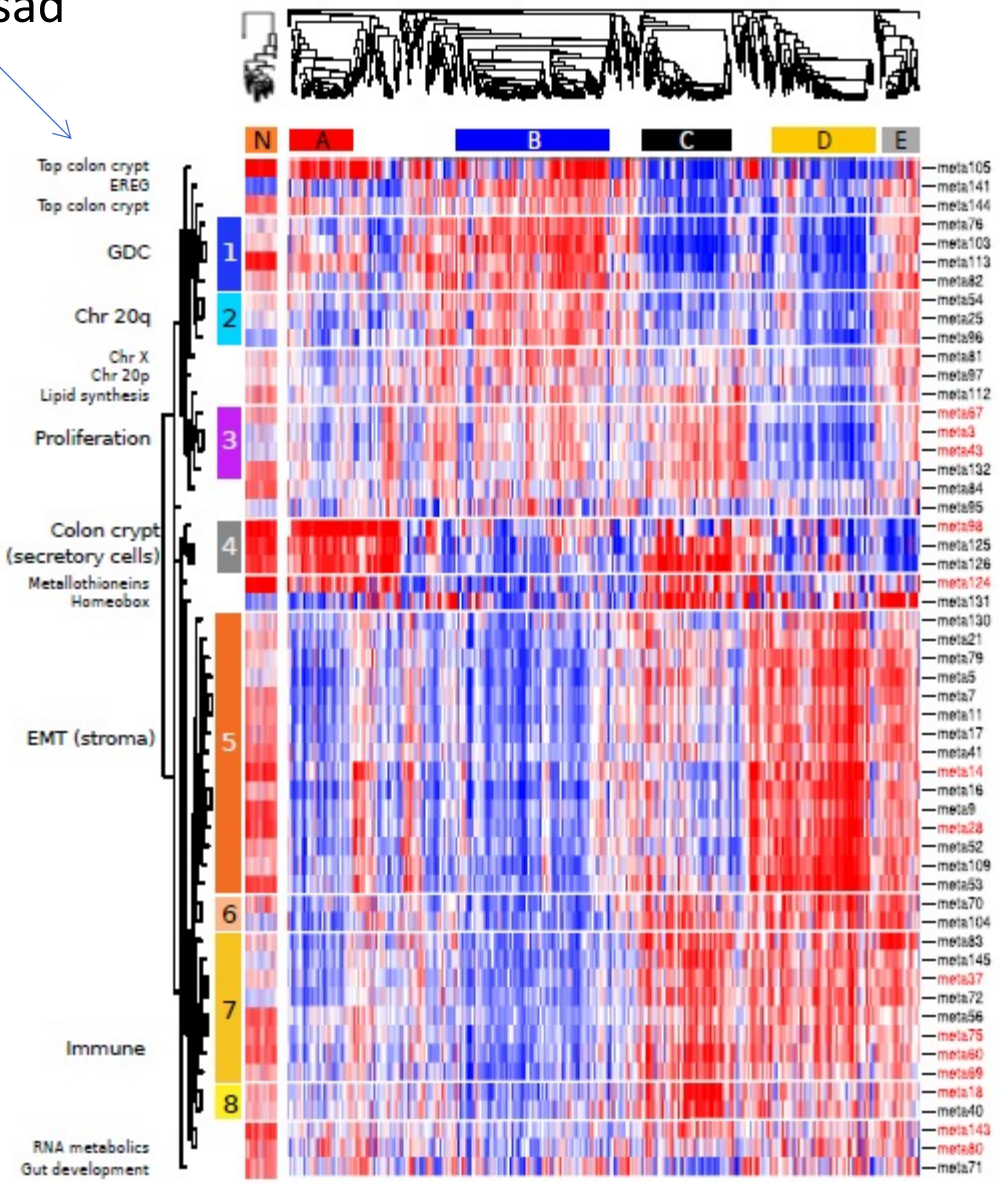
Charakterizovat tyto podtypy pomocí klinických a známých molekulárních parametrů.

Motivy genové expresee v podtypech

Analýza genových sad

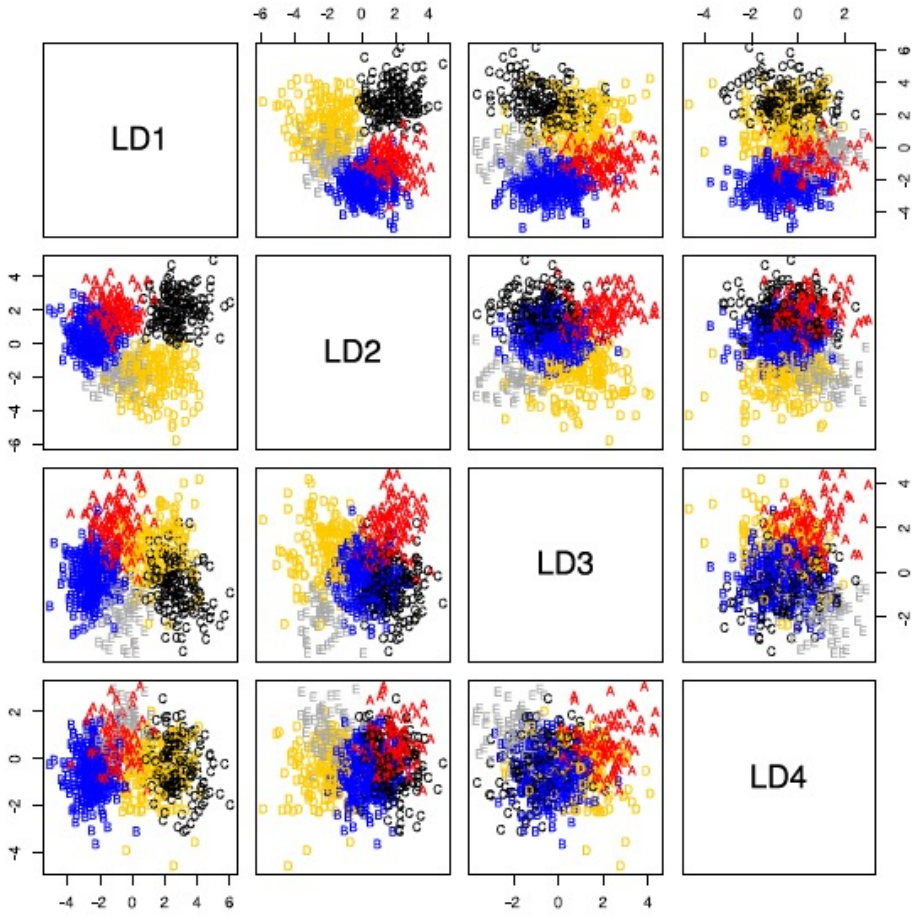
Hierarchické shlukování na matici konsensusu





Expresní profily pro jednotlivé podtypy

- Klasifikátor LDA (linear discriminant analysis)



Minimální genová sada

- Selekce genů pomocí ***elastic net*** - počítá s korelovanými geny (neodstraňuje jen jeden gen, ale všechny korelované)
- Klasifikátor vytvořený na selektovaných genech (Shrunken centroids)

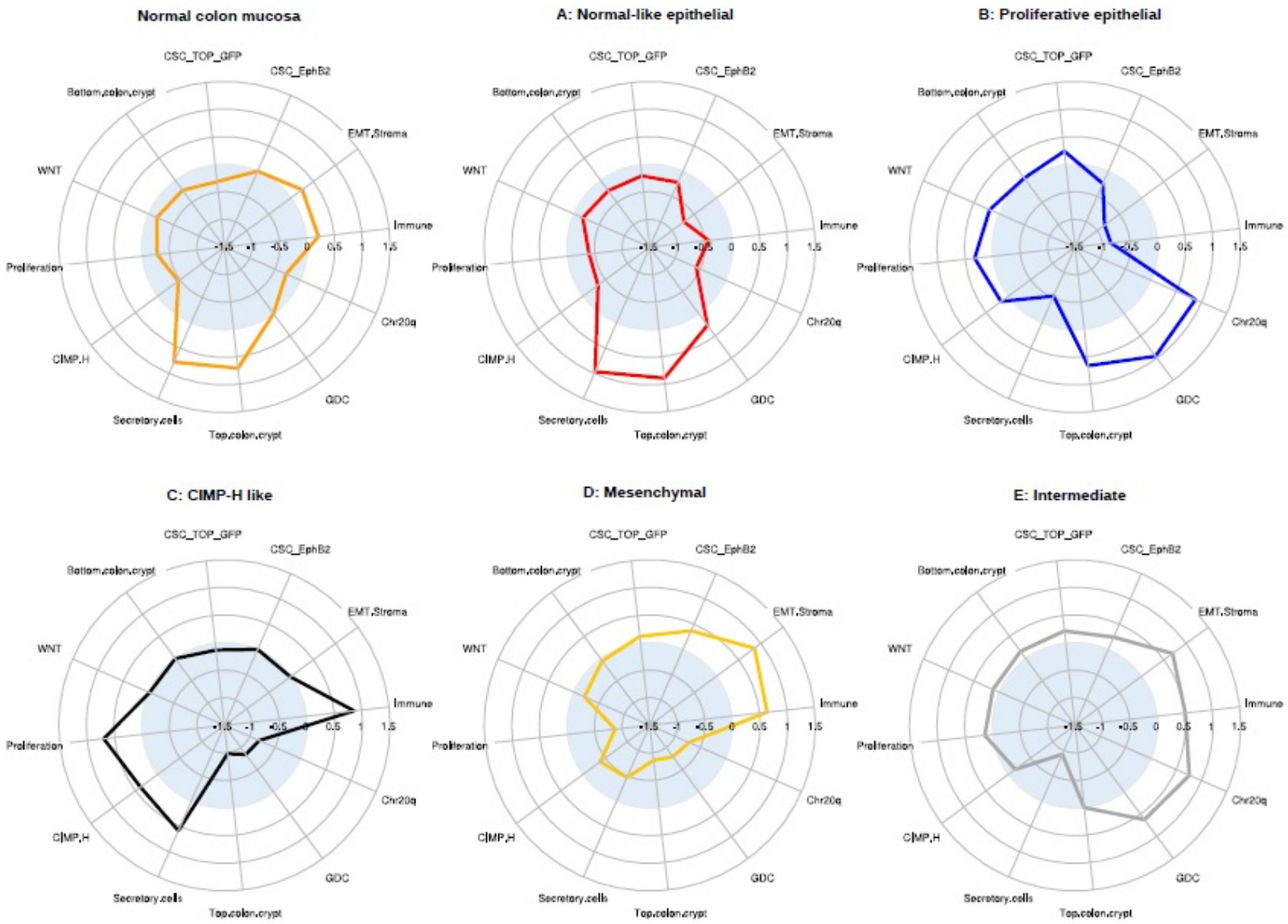
Podtyp	Minimální genová sada
A	CLCA1, PADI2, ADTRP, RETNLB, TIMP3, MUC2, FNDC1, NR3C2, SULF1, B3GNT7, STYK1, CHI3L1
B	FARP1, ALOX5, FSCN1, HNF4A, RARRES3, MYRIP, GPSM2, TSPAN6, CCDC113, CDHR1, KCTD12, SGK1, BASP1, MT1E, GPX8, RPS6KA3, SOCS3, SLC5A6, PRR15, PLAGL2, IHH, CREB3L1, TP53RK, YAE1D1, EPB41L3, QPRT, KCNK5, RNF43, VAV3, CXCR4, ITPRIP, GRM8, GFPT2, KCNMA1, KIAA0226L,RNASE1
C	TFAP2A, ATP9A, RAB27B, ANP32E, CXCL14, IDO1,RARRES3, EGLN3, KIAA0226L,C10orf99,RPL22L1, PLK2
D	PRICKLE1, RBM47, TAGLN, BOC, HOOK1, C7, ANK2, DCHS1, DDR2, CRYAB, GEM
E	REG4, IL6, CXCL5, RAB27B, CEACAM6, PI15, MRPS31, RAP2A, UQCC, AGR3, HSD11B1, IL1B

Cíl:

Najít skupiny nádorů kolorekta s podobnou expresí genů (podobným genovým profilem) ~ podtypy

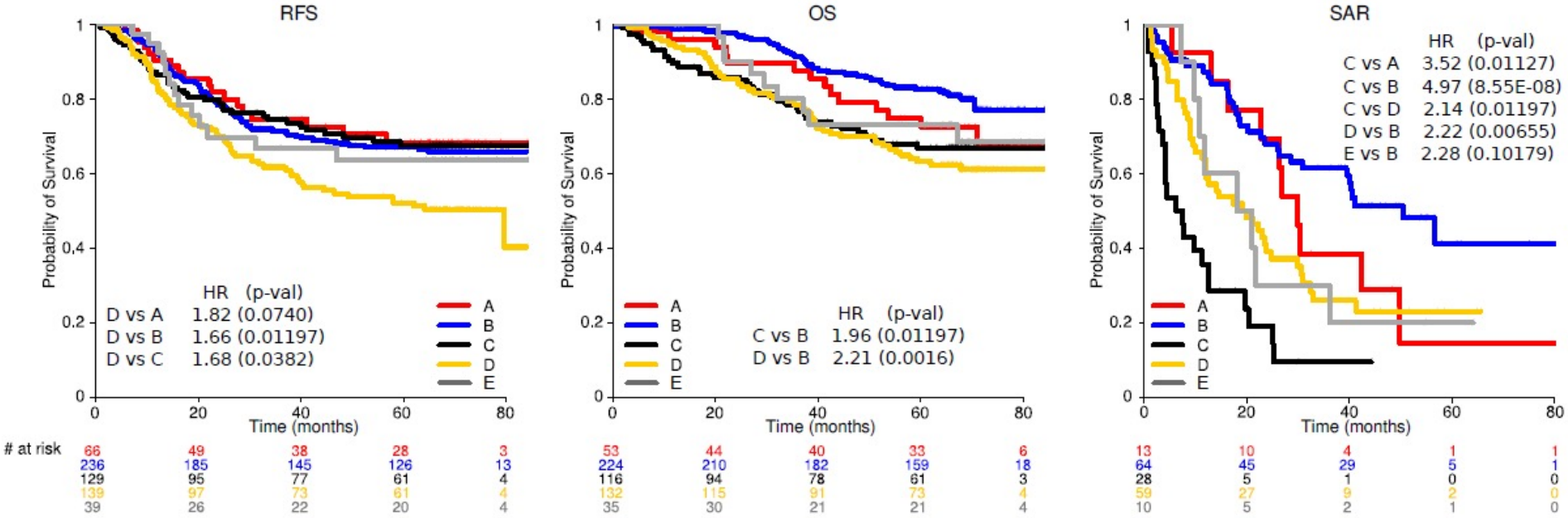
Charakterizovat tyto podtypy pomocí klinických a známých molekulárních parametrů.

Vzor exprese biologických motivů



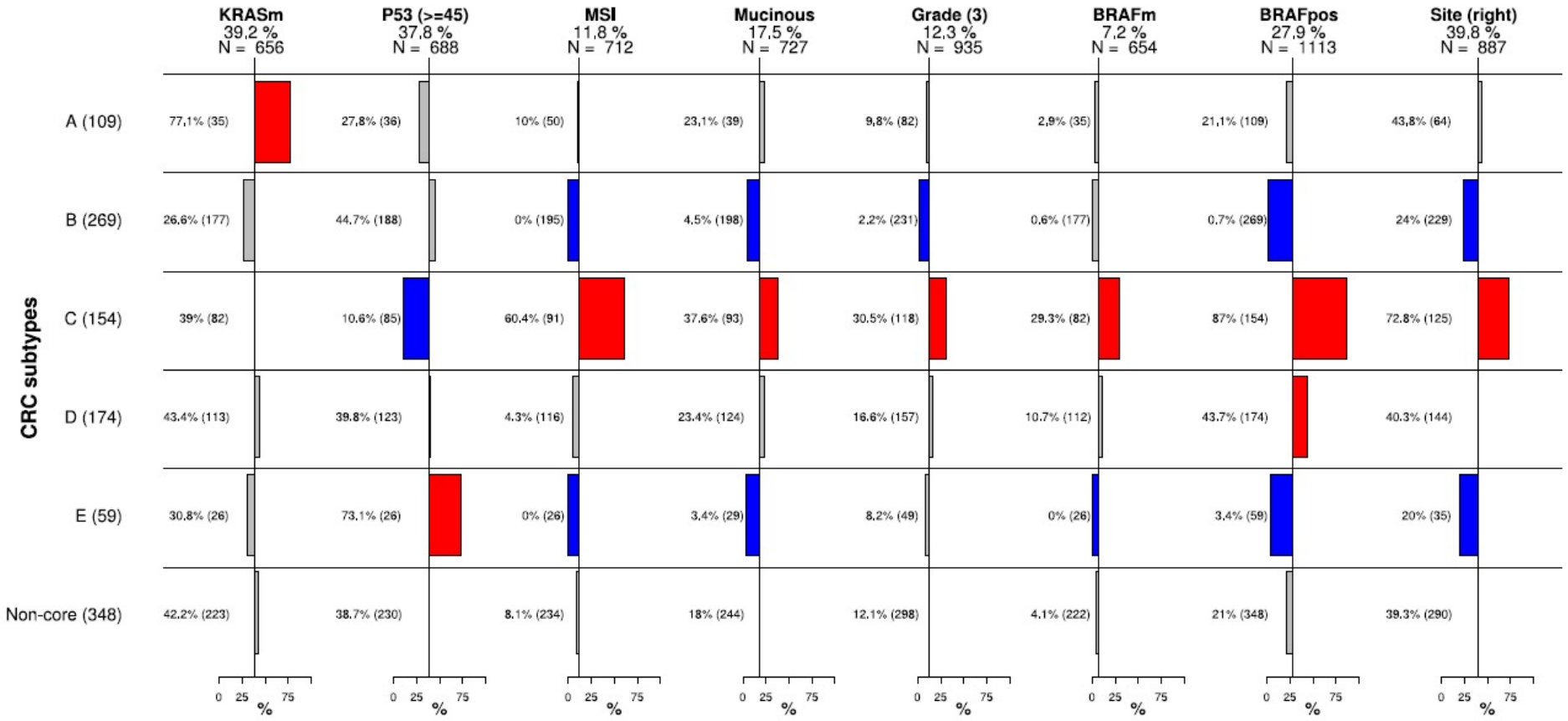
- Analýza genových sad pomocí KS testu

Rozdíly v prežití



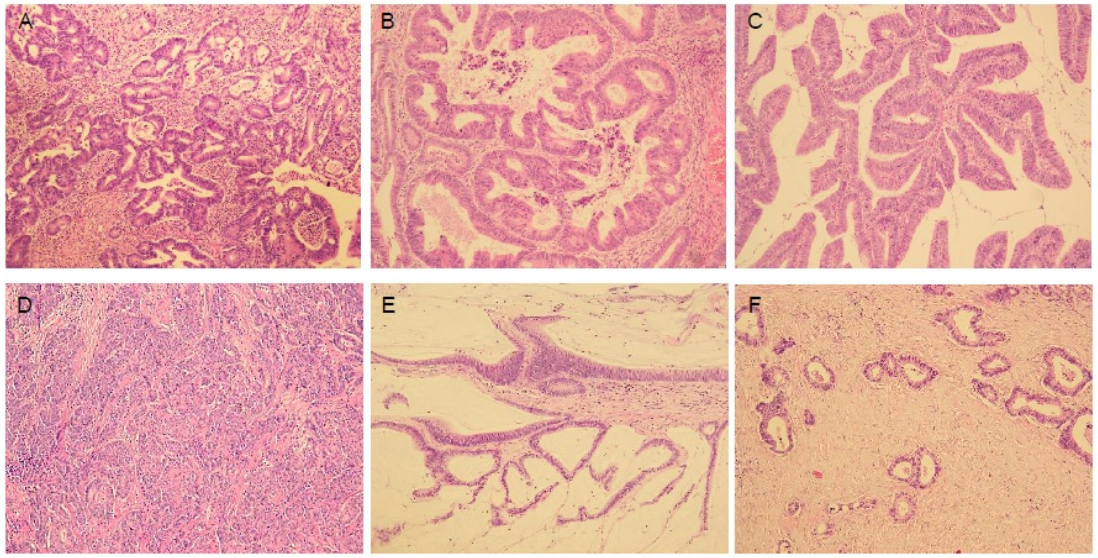
- Kaplan-Meierovy křivky přežití
- Coxův model proporcionálních rizik (efekt stádia vs podtypů)

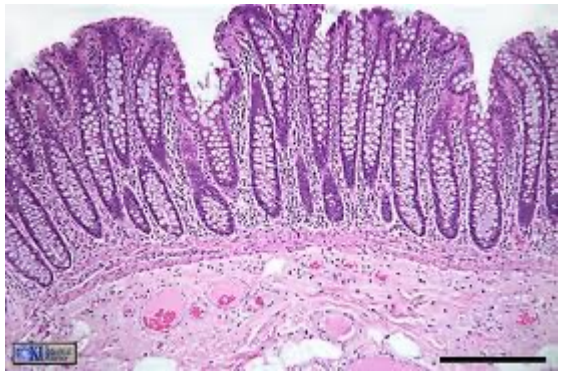
Charakterizace podtypů klinickými a molekulárními proměnnými



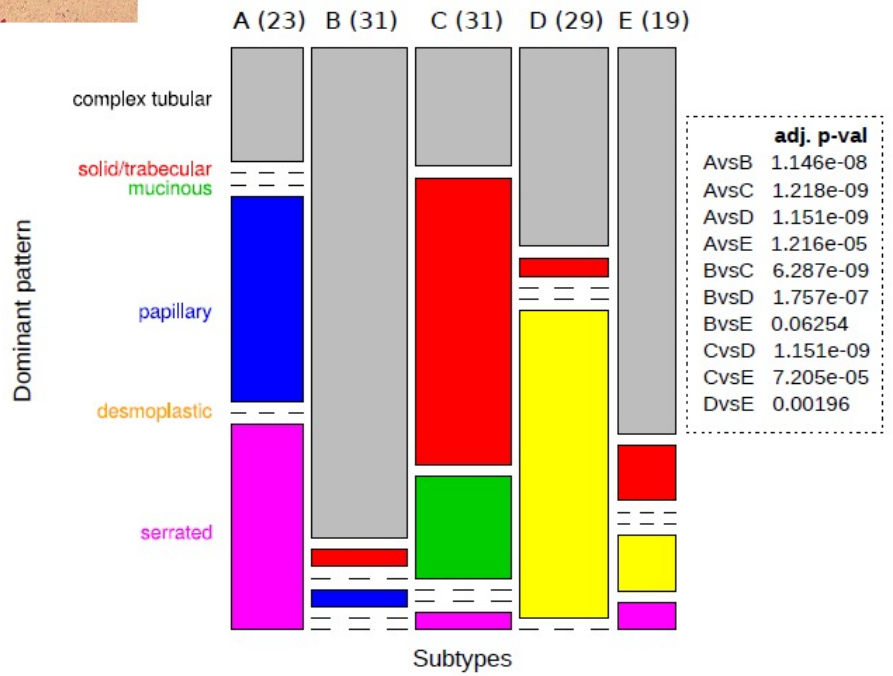
- Rozdíl od populační baseline u každého podtypu pomocí Fisherova exaktního testu, FDR úprava p-hodnot

Histologické rozdíy

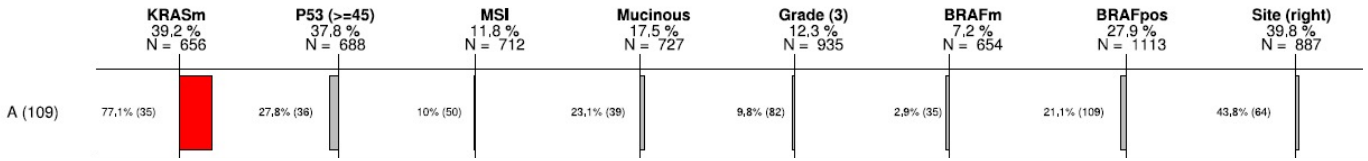


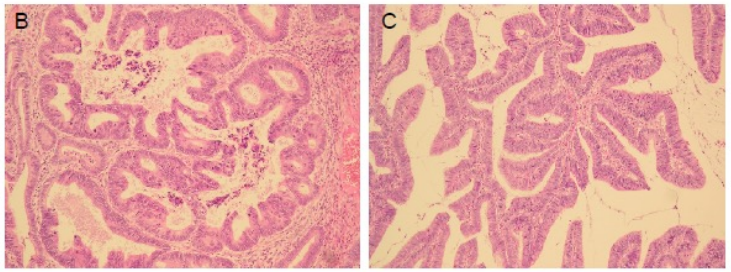


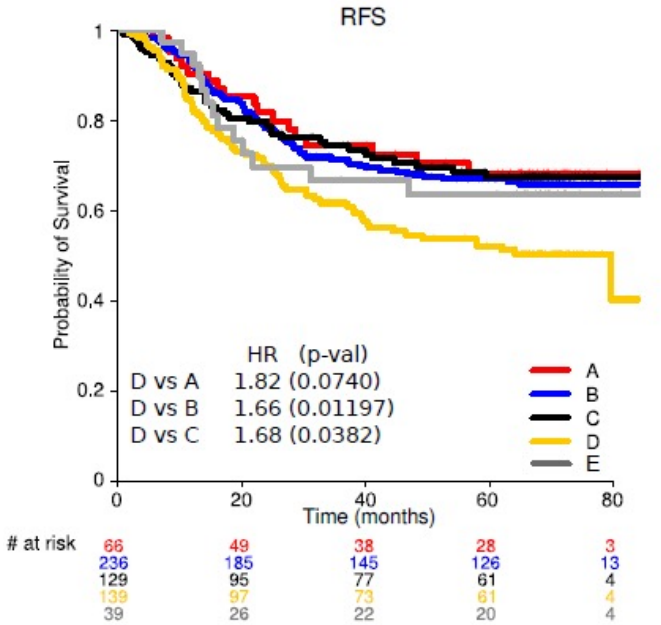
- Fisherův test, úprava na FDR

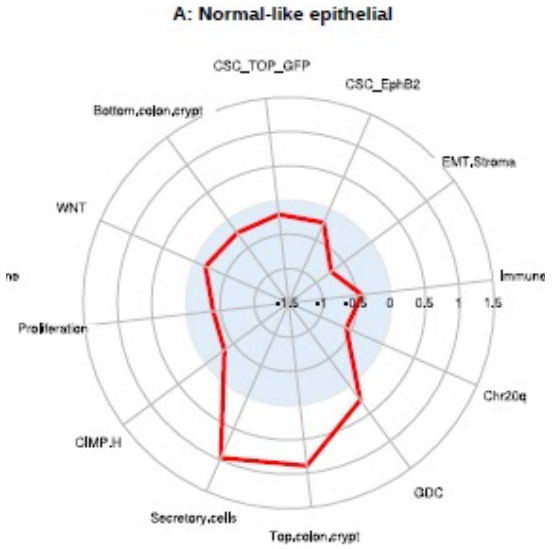


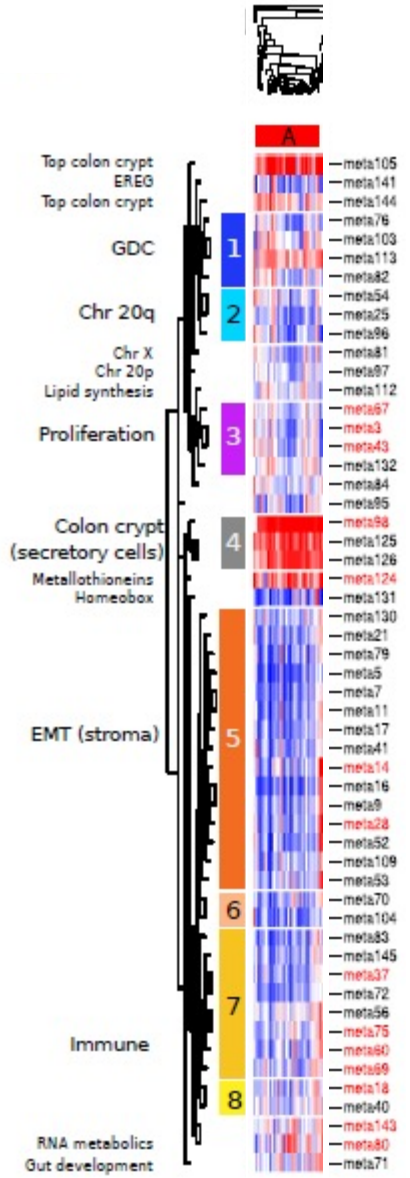
Podtyp A - Surface crypt like - KRAS mutanti, papillaris a Serrated morfotyp, nejvíce diferencovaný, bez aktivní Wnt signální dráhy. Dobré OS a RFS.



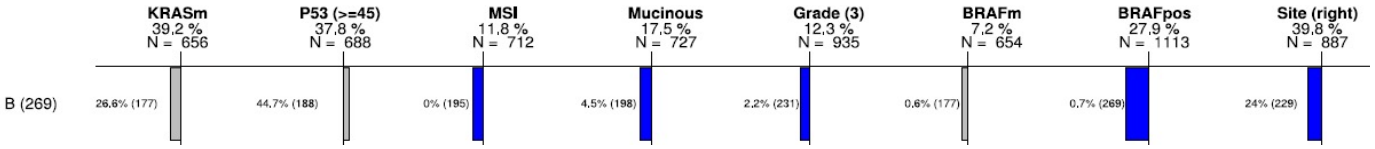


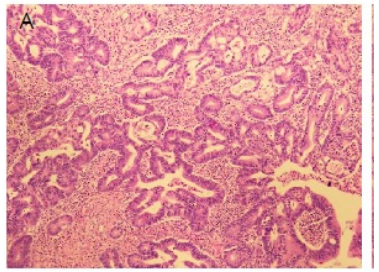


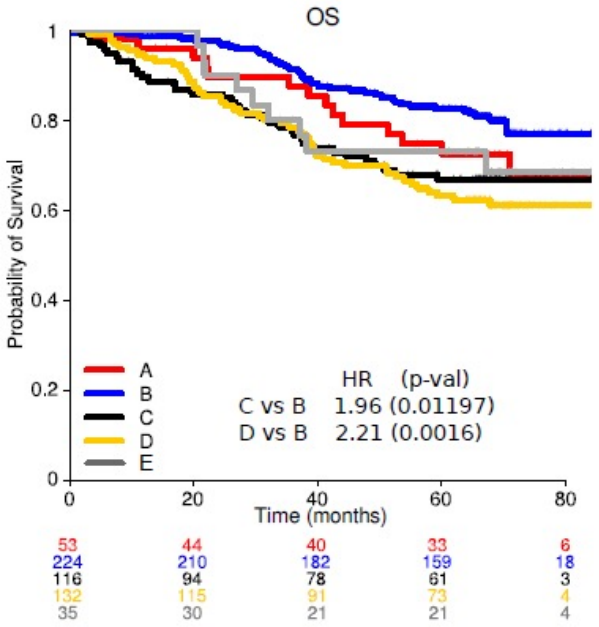


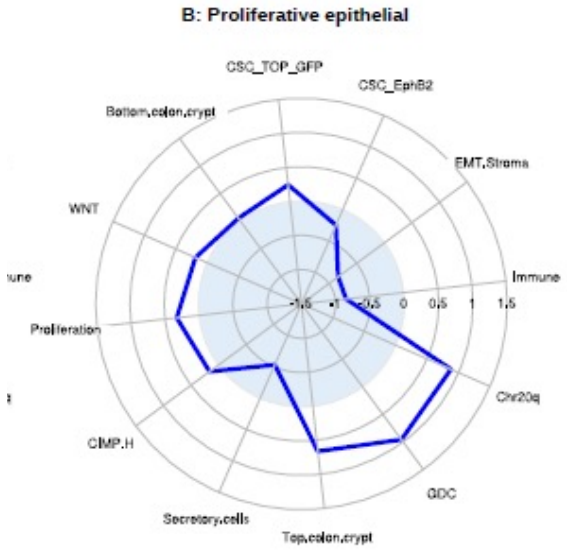


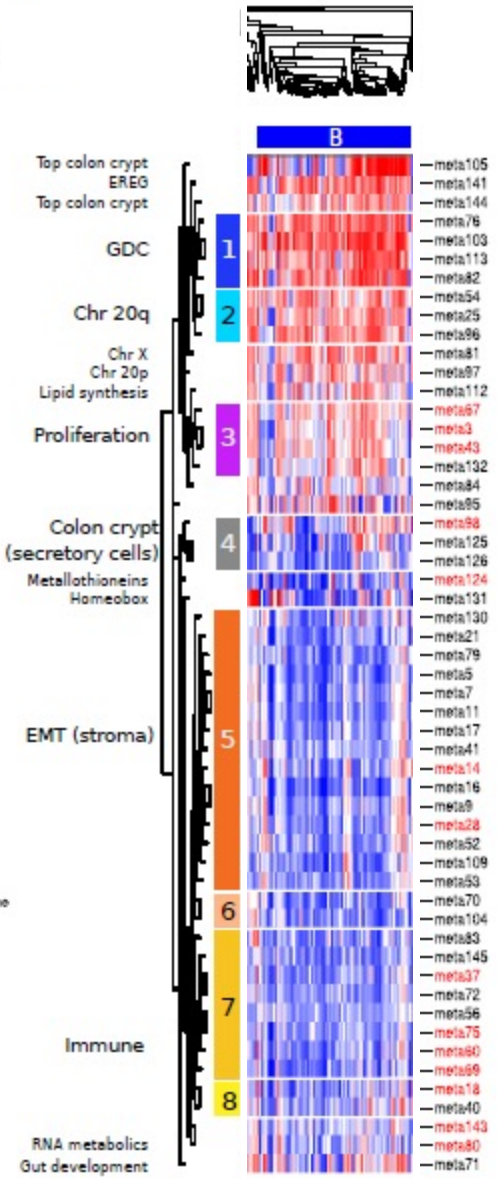
Podtyp B - Lower crypt like - diferencované ale bez sekrečních buněk, proliferující, a aktivní Wnt signální dráhou. Komplexní tubulární morfotyp. Časo MSS, BRAFwt, nižšího grady, dobré přežití v OS, RFS i SAR.



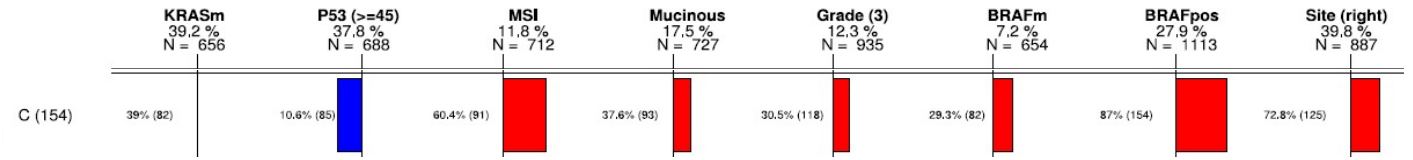


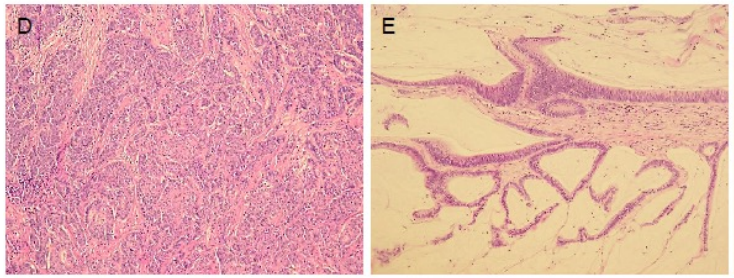


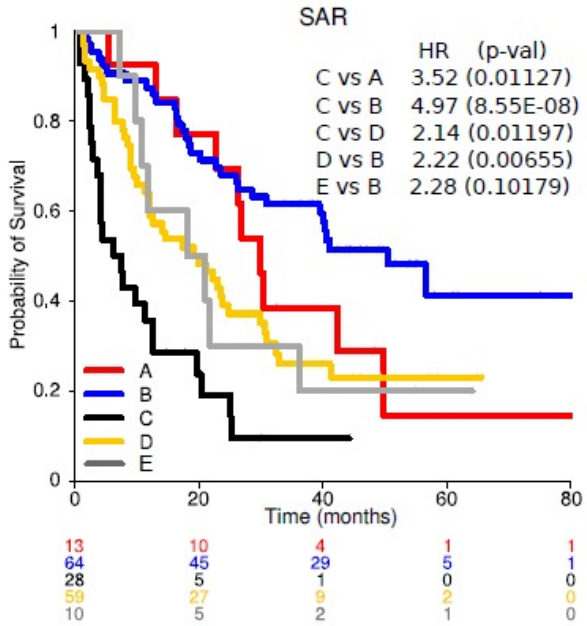


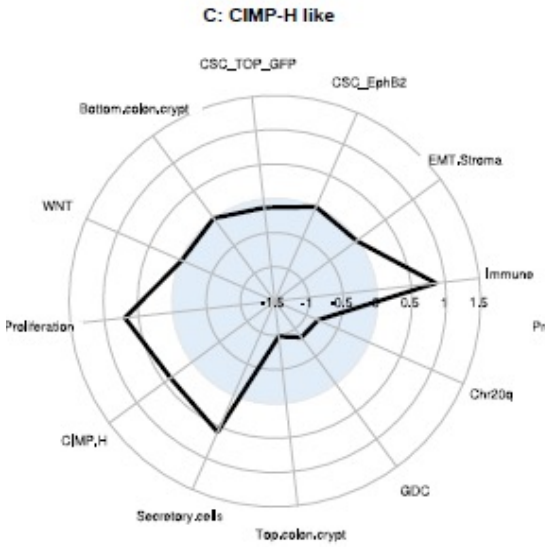


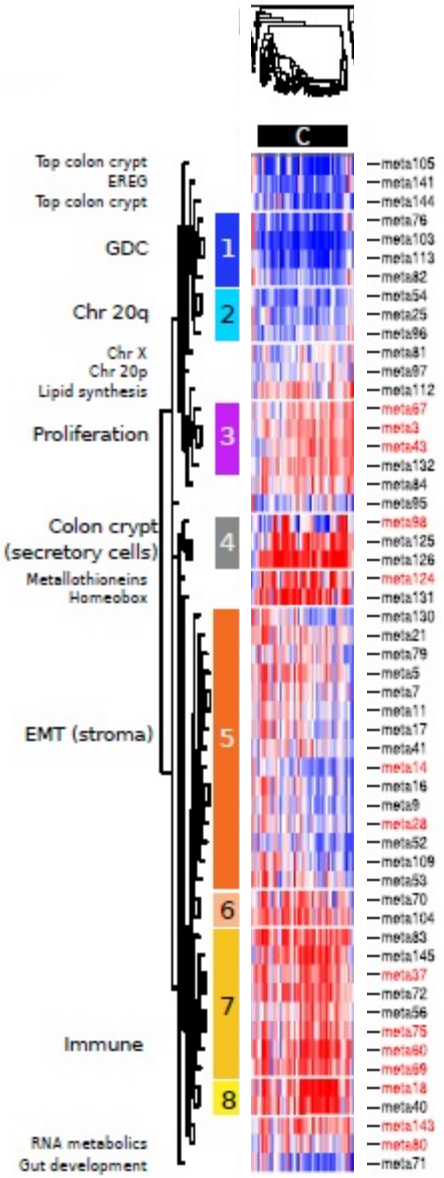
Podtyp C – CIMP-H like - často MSI, BRAF-mutantní, hypermutované, z pravé části tlustého střeva. Histologicky - horší diferencované, solidního-trámčité s mucínovéým morfotyp. Aktivní proliferujícími a mají silnou imunitní reakci. Dobrý RFS, ale špatný OS and SAR.



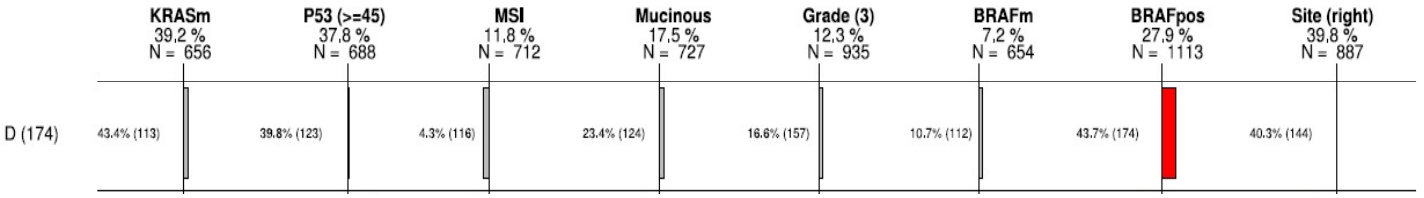


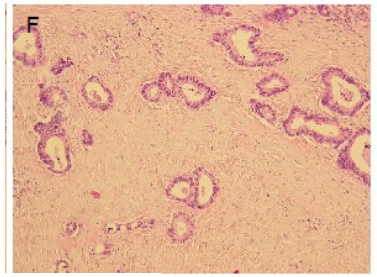


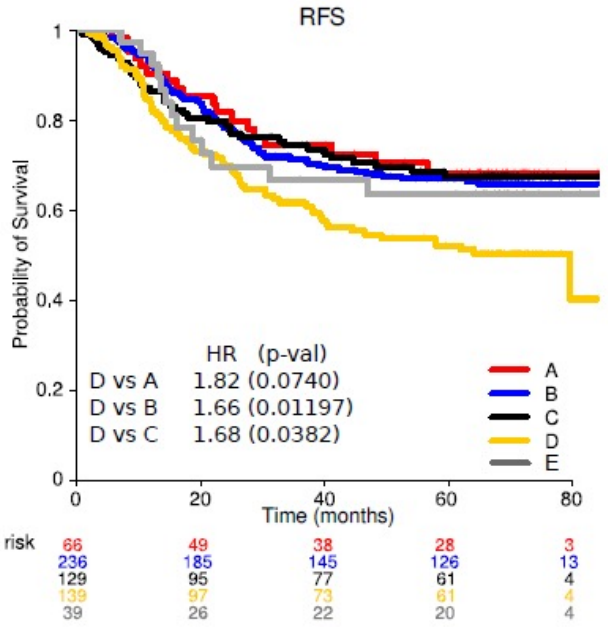


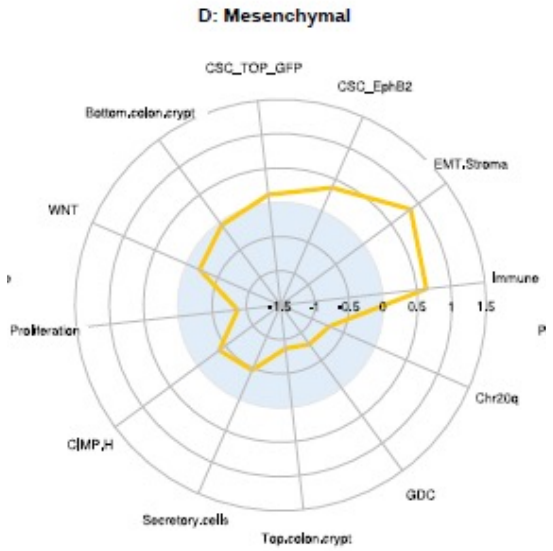


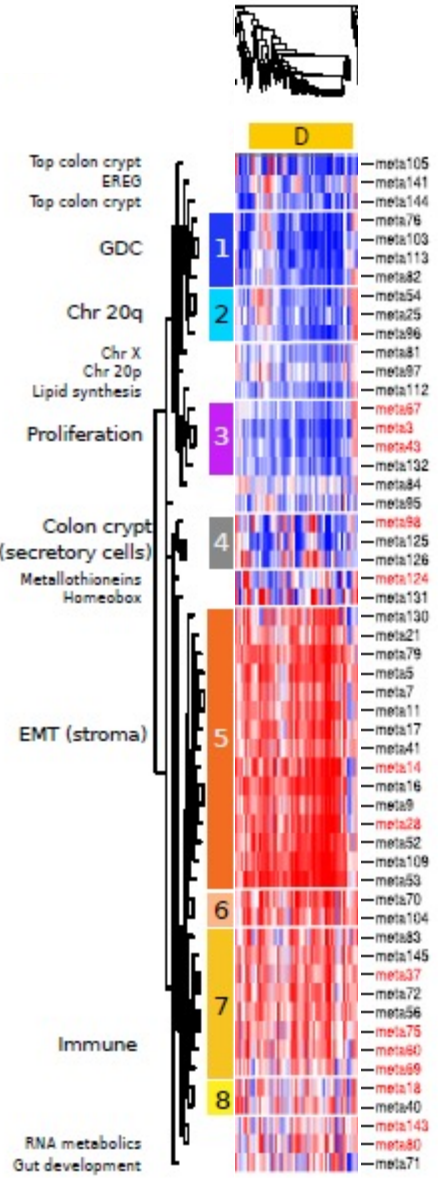
Podtyp D – Mesenchymal – markery kmenových buněk, mnoho mezenchymálních buněk, které se projevují expresí EMT genů. Wnt signální dráha je neaktivní a proliferace nízká. Klinické a mutační charakteristiky se neliší od populační baseline. Mají nejkratší přežití do relapsu, špatný OS a SAR.



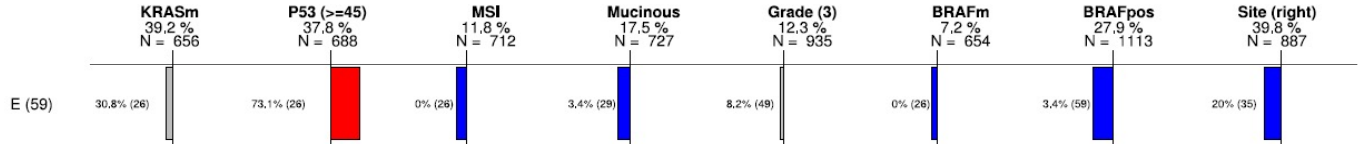


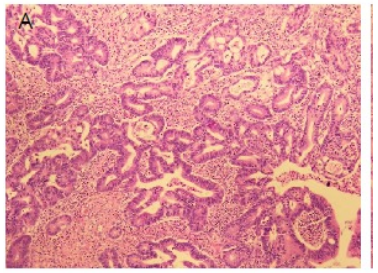


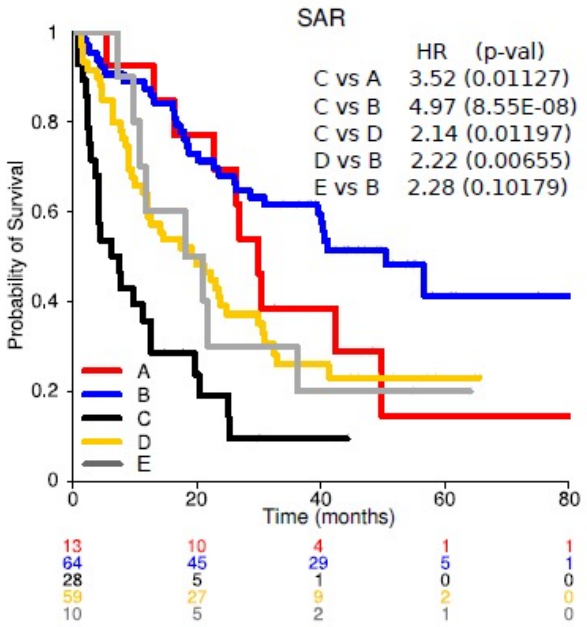


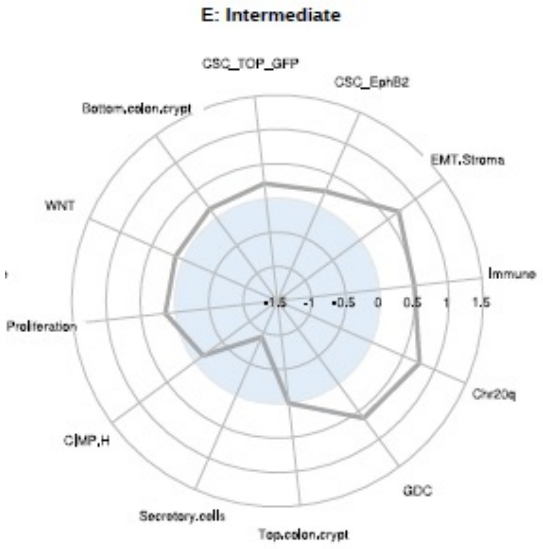


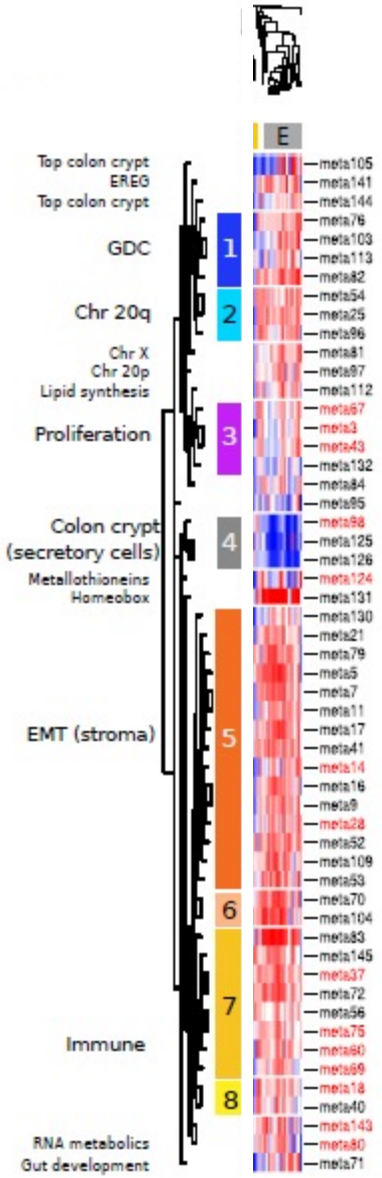
Podtyp E – **Mixed** – často MSS, BRAFwt, z levé strany tlustého střeva. Podobně jako podtyp D exprimuje geny kmenových buněk a EMT procesu, avšak podobně s B má vysoce aktivní kanonické Wnt dráhy a vypadá více diferenciovaný. Je podobný B - komplexní tubulární, jen častěji obsahuje mutaci p53.







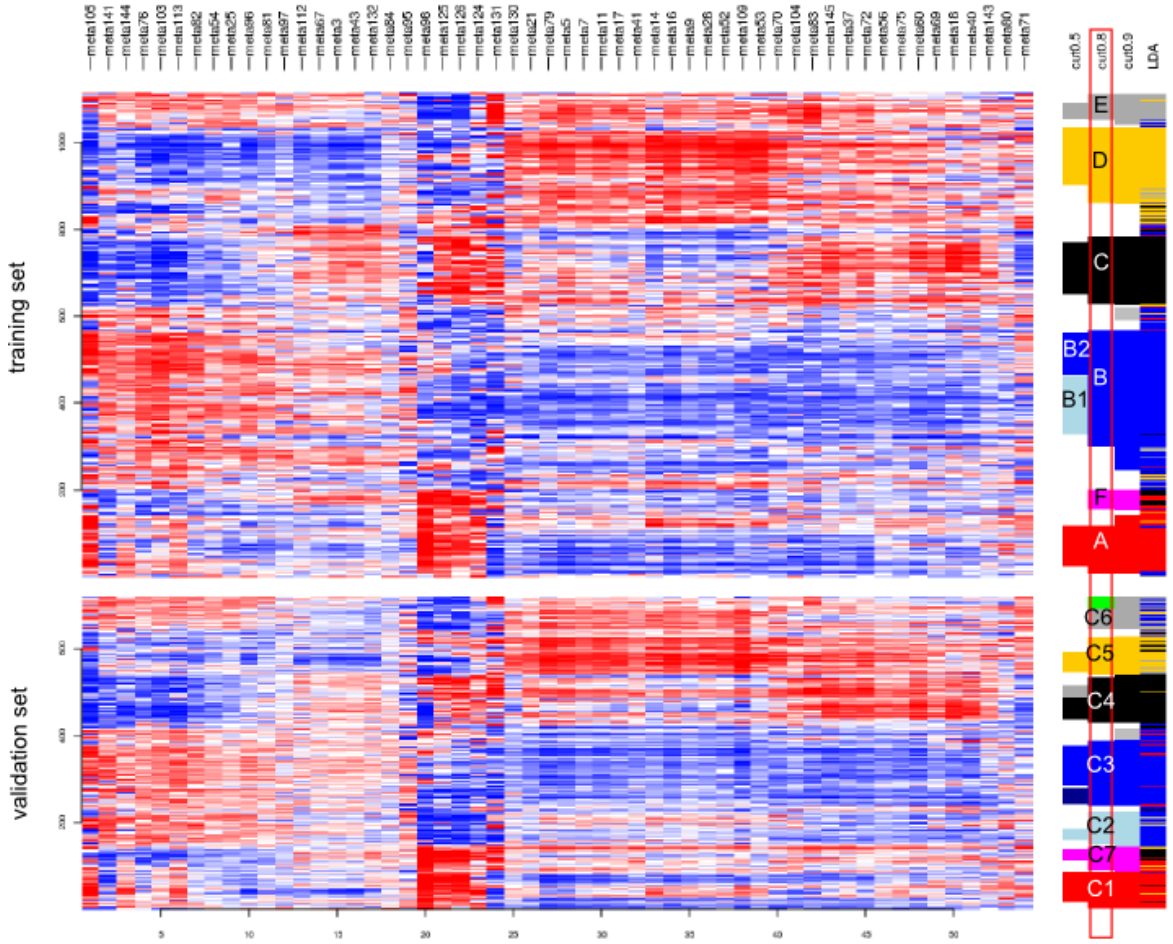




Validace podtypů kolorektálního karcinomu

Validace algoritmu a parametrů modelu na testovacím souboru

Když zopakuji celou proceduru na jiném souboru, dostanu podobné skupiny?



Cluster/subtype in validation set	LDA assignment					SUM
	A	B	C	D	E	
C1 / A	**74**	4	3	3	0	84
C2 / B1	1	**58**	0	2	13	74
C3 / B2	12	**134**	1	0	1	148
C4 / C	1	2	**99**	4	0	106
C5 / D	0	3	12	**64**	7	86
C6 / E	1	**17**	0	**17**	**13**	48
C7 / F	**23**	1	**22**	9	1	56
Non-core	21	53	18	8	18	118
SUM	133	272	155	107	53	720

Cluster/subtype in validation set	Subtypes from training set most correlated to validation subtypes					
	First subtype			Second subtype		
	Subtype	Cor	P-val	Subtype	Cor	P-val
C1 / A	A	0.85	p<1.0E-15	F	0.41	p<1.0E-15
C2 / B1	B	0.71	p<1.0E-15	E	0.47	p<1.0E-15
C3 / B2	B	0.91	p<1.0E-15	A	0.36	p<1.0E-15
C4 / C	C	0.89	p<1.0E-15	F	0.29	p<1.0E-15
C5 / D	D	0.93	p<1.0E-15	E	0.37	p<1.0E-15
C6 / E	E	0.63	p<1.0E-15	D	0.58	p<1.0E-15
C7 / F	F	0.61	p<1.0E-15	C	0.55	p<1.0E-15

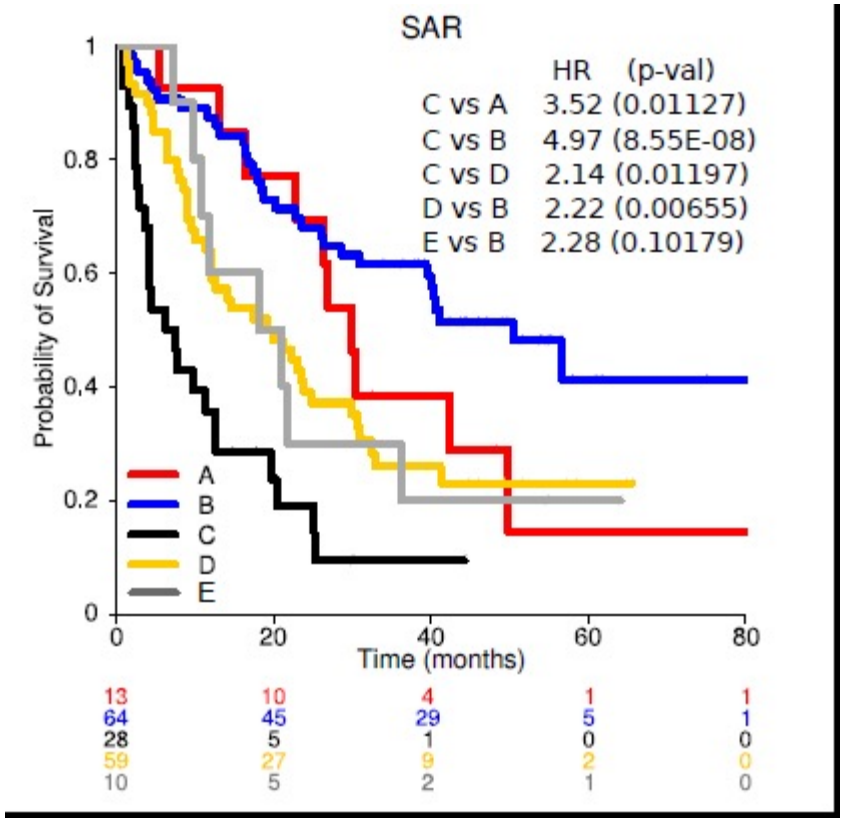
Validace konceptu pomocí klinických, molekulárních a histologických charakteristik objevených skupin

Mají objevené skupiny biologickou podstatu / odrážejí známé vědecké poznatky?

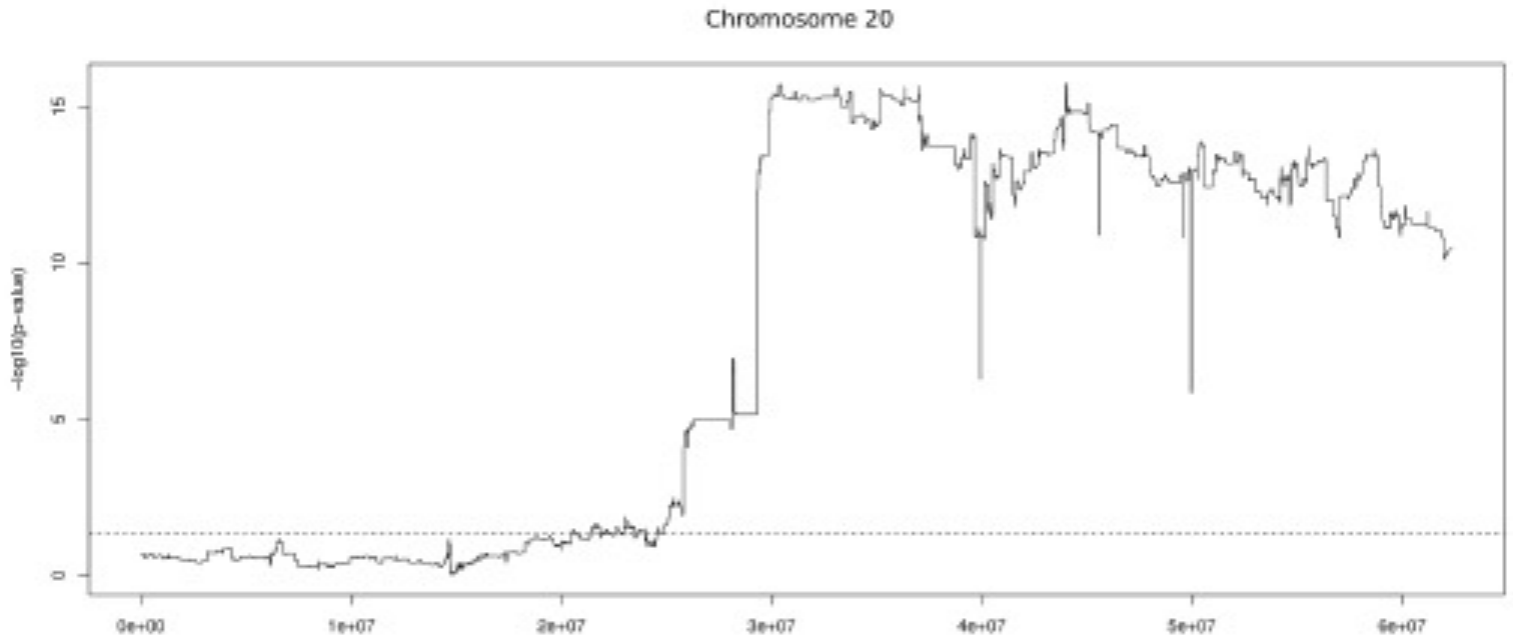
Je rozložení těchto charakteristik mezi podtypy srovnatelné ve validačním souboru?

Mají objevené skupiny biologickou podstatu / odrážejí známé vědecké poznatky?

Podtyp C – pravostranné, BRAFm, MSI nádory, které jsou známé **špatným přežitím po relapsu**

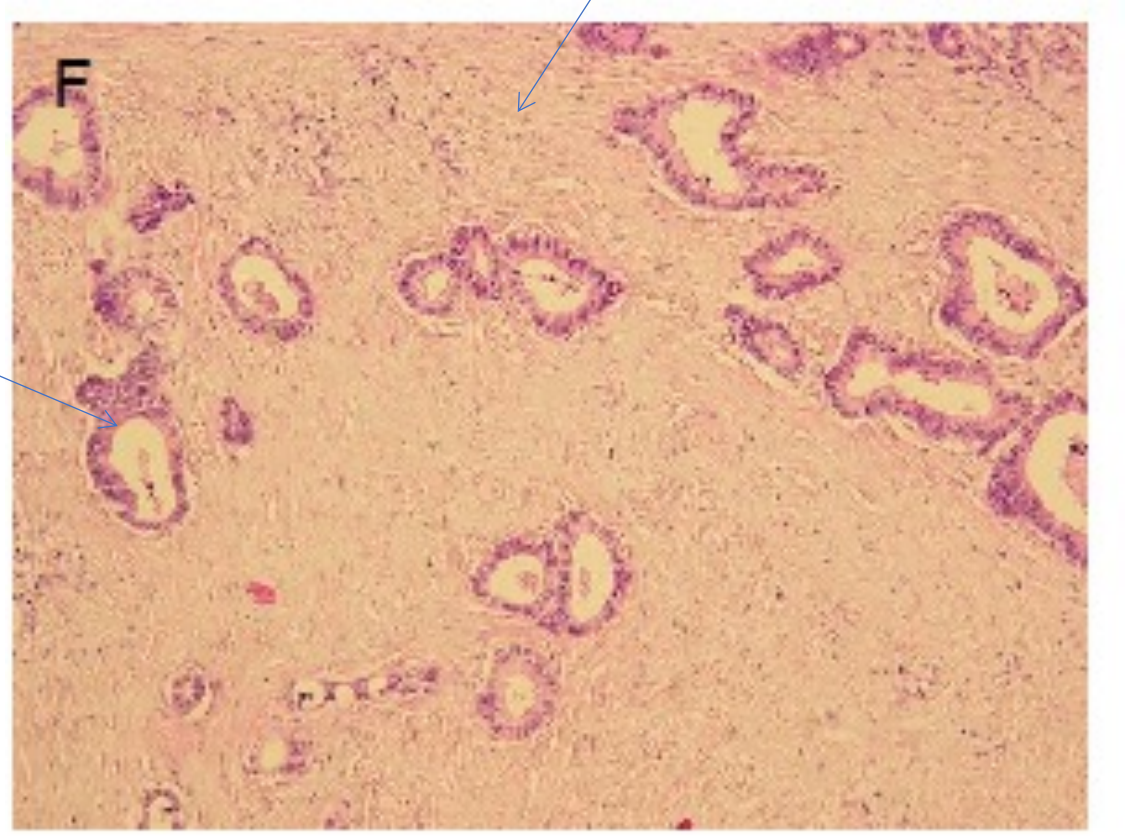


Zvýšená exprese genů chr. 20q v podtypu B by mohla znamenat amplifikaci chr20q regionu.

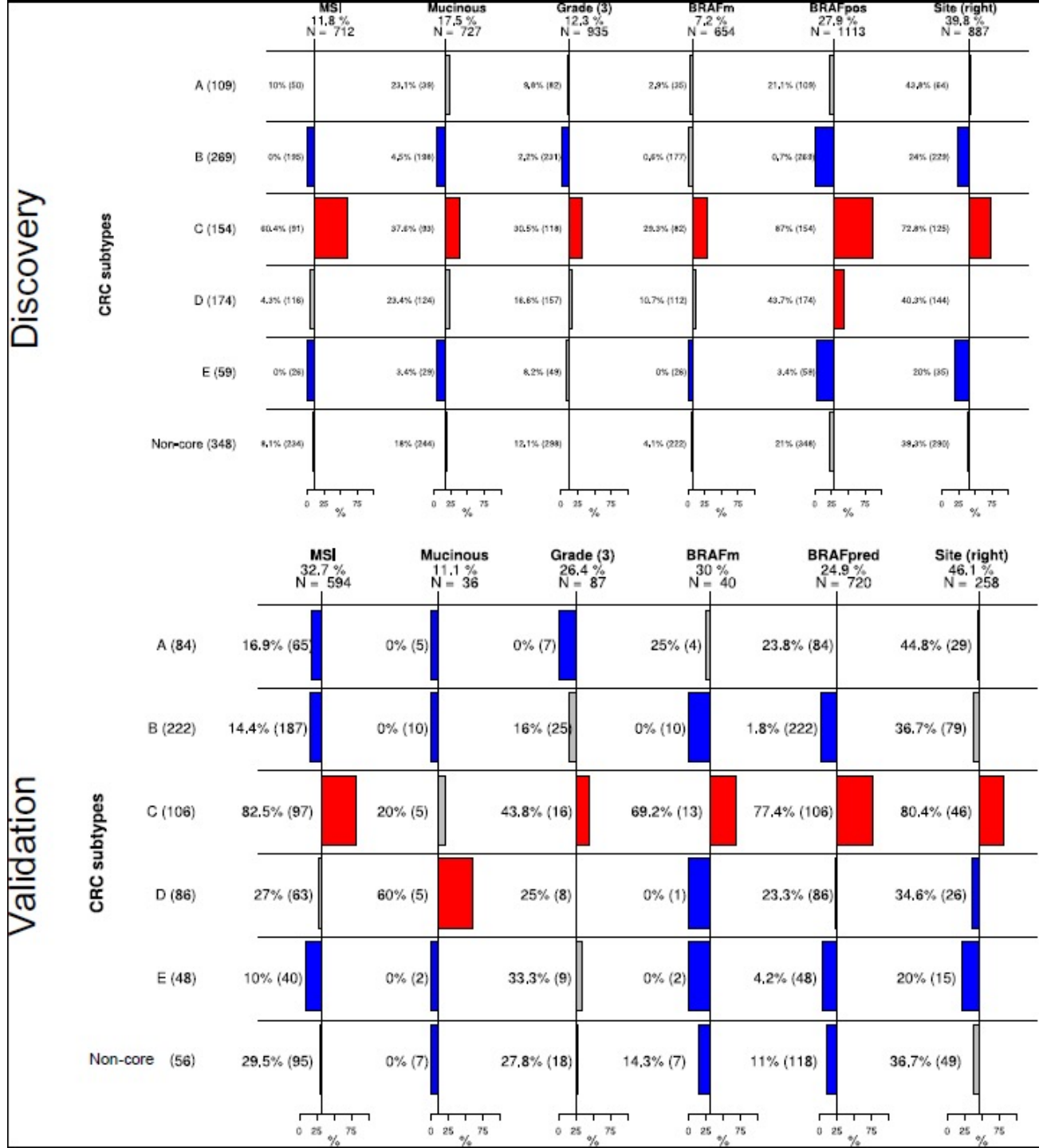


Podtyp D – mezenchymální – histologické vyhodnocení: v nádoru přítomna silná desmoplastická reakce (mezenchymální tkáň)

Nádor



Je rozložení klinických charakteristik mezi podtypy srovnatelné ve validačním souboru?


Discovery
MSI
11.8 %
N = 712
Mucinous
17.5 %
N = 727
Grade (3)
12.3 %
N = 935
BRAFm
7.2 %
N = 654
BRAFpos
27.9 %
N = 1113
Site (right)
39.8 %
N = 887
CRC subtypes
A (109)
B (269)
C (154)
D (174)
E (59)
Non-core (348)
10% (50)
23.1% (38)
9.8% (82)
2.9% (35)
21.1% (109)
43.8% (64)
0% (195)
4.5% (198)
2.2% (231)
0.6% (177)
0.7% (269)
24% (229)
60.4% (91)
37.6% (93)
30.5% (118)
25.3% (82)
87% (154)
72.8% (125)
4.3% (116)
23.4% (124)
16.6% (157)
10.7% (112)
43.7% (174)
40.3% (144)
0% (26)
3.4% (29)
8.2% (49)
0% (26)
3.4% (58)
20% (35)
8.1% (234)
18% (244)
12.1% (298)
4.1% (222)
21% (348)
39.3% (290)
0 25 75
%
Validation
MSI
32.7 %
N = 594
Mucinous
11.1 %
N = 36
Grade (3)
26.4 %
N = 87
BRAFm
30 %
N = 40
BRAFpred
24.9 %
N = 720
Site (right)
46.1 %
N = 258
CRC subtypes
A (84)
B (222)
C (106)
D (86)
E (48)
Non-core (56)
16.9% (65)
0% (5)
0% (7)
25% (4)
23.8% (84)
44.8% (29)
14.4% (187)
0% (10)
16% (25)
0% (10)
1.8% (222)
36.7% (79)
82.5% (97)
20% (5)
43.8% (16)
69.2% (13)
77.4% (106)
80.4% (46)
27% (63)
60% (5)
25% (8)
0% (1)
23.3% (86)
34.6% (26)
10% (40)
0% (2)
33.3% (9)
0% (2)
4.2% (48)
20% (15)
29.5% (95)
0% (7)
27.8% (18)
14.3% (7)
11% (118)
36.7% (49)
0 25 75
%

Závěrem

- Zkoušejte více metod v rámci jedné studie
- Nezapomeňte na robustní shlukování
- Pokud je potřeba (hierarchické shlukování), použijte dynamické řezání stromu
- Propojte výsledky s biologickými a klinickými proměnnými, interpretuje nálezy
- Pokud je to možné, validujte výsledky na testovacím souboru! Pokud ne, vaše interpretace a závěry jsou pouze spekulativní.