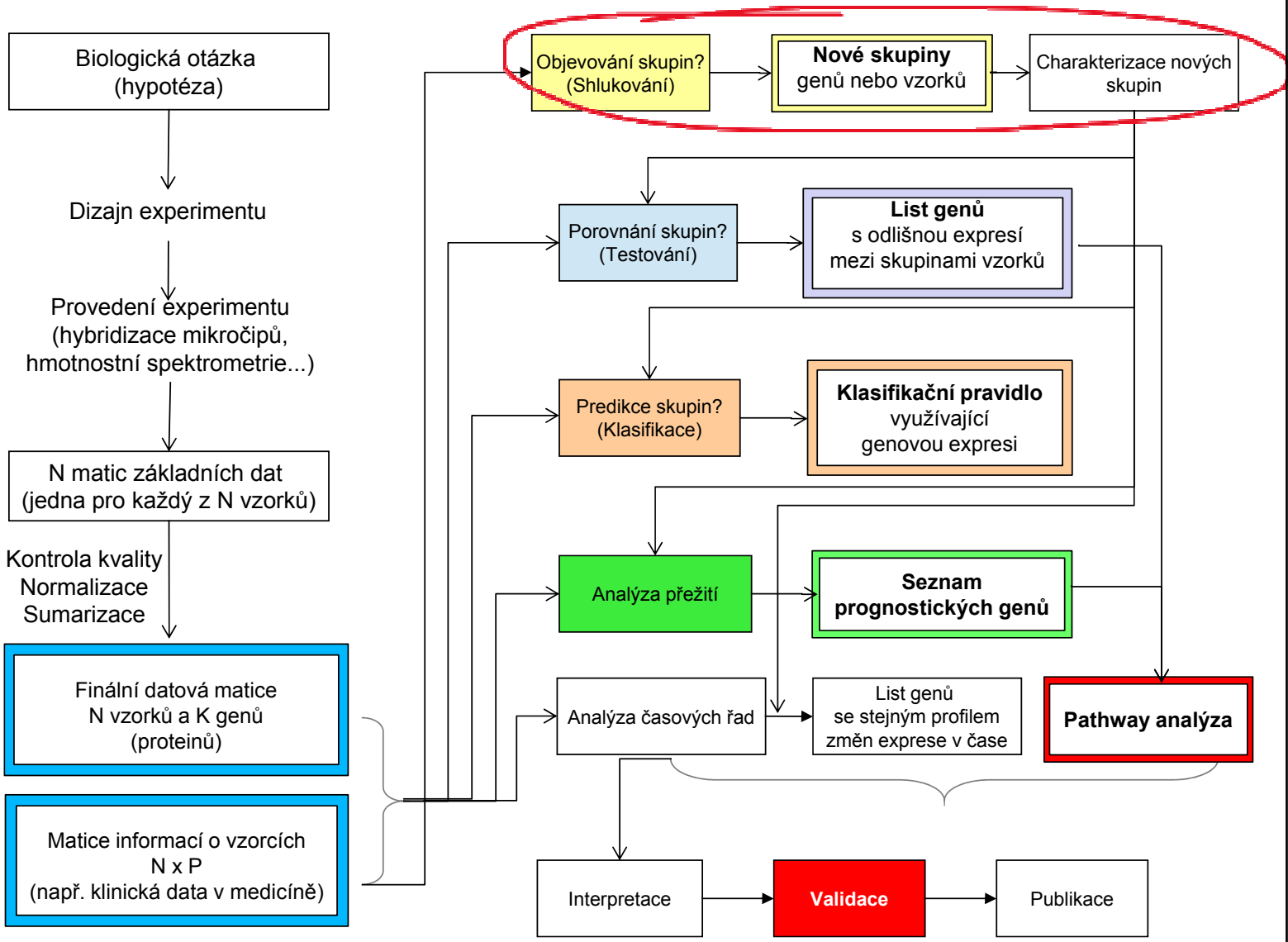
Detekce biomarkerů z omics experimentů

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

- Podzim 2019

Objeování neznámých skupin

Jak se hledá potenciální biomarker v omics datech



Objevování skupin

- Snažíme se **vyvodit závěry** o **vzorcích** nebo **molekulách 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 vzájemně **tak odlišné, jak jen je to možné**

Odpovídáme na otázky:

Existují skupiny vzorků s podobným molekulárním profilem?

Existují skupiny molekul s podobnými vlastnostmi?

Kdy provádíme shlukování v molekulární biologii

- **Shlukování proměnných**
  - Chceme identifikovat skupiny ko-regulovaných genů/proteinů/metabolitů…
  - Chceme zredukovat dimenzi dat na základě funkčních genových/proteinových/… skupin
- **Shlukování vzorků**
  - Kontrolujeme kvalitu vzorků
  - Chceme najít nové skupiny vzorků (například podtypy)
  - Chceme zkontrolovat diskriminační schopnost proměnných (genů etc.…) vybraných při porovnávání známých skupin

Příklady objevování skupin

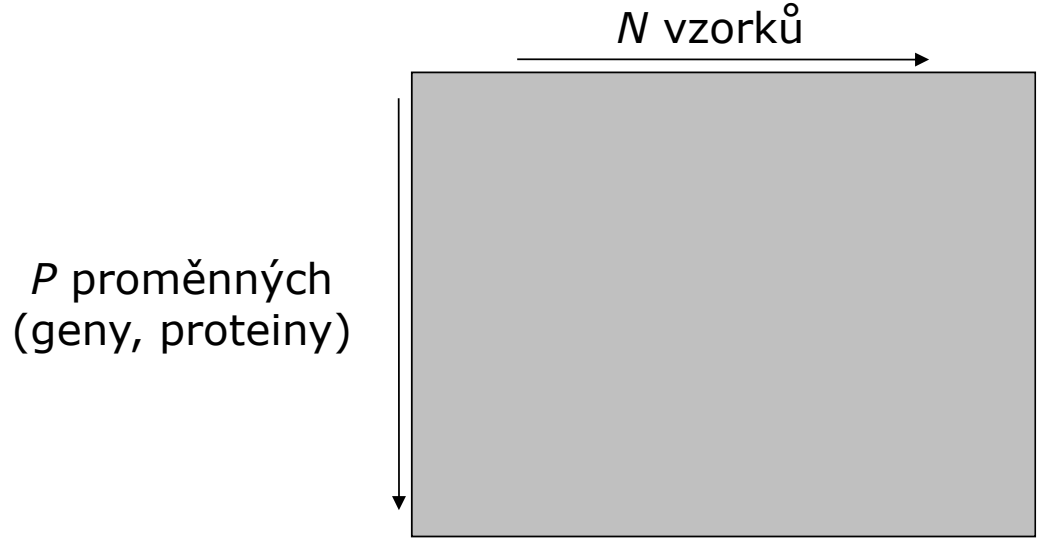
- Molekulární podtypy karcinomu prsu
- Bakterie se stejnými geny rezistence
- Proteiny s podobnou abundancí u zdravých vs nemocných
- Geny se stejnou genovou expresí v čase
- Skupiny bakterií společně se vyskytuících s konkrétními metabolity

Omezení

- Neznámá pravda – nutné co nejobjektivněji popsat skutečnost
- (Ne) reprezentatívnost populace v datech
- Technické možnosti
- Odchylky v datech můžou podstatně ovlivnit výsledek
- Typ dat určuje uhol pohledu (génová exprese vs proteinová abundance, metabolický profil, epidemiologie, histologie, DNA mutace...)

Princip shlukování

- Máme datovu matici ***X*** velikosti ***N x P***
  - *N* – počet objektů (vzorků)
  - *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:
  - **Metody založené na vzdálenosti**
    - neparametrické
    - nejčastěji používané, intuitivní
    - hierarchické a nehierarchické shlukování
  - **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

Princip:

- Zvolíme **metriku vzdálenosti** (jak vzdálenost měříme?)
- Vypočteme matici vzdáleností mezi objekty (každý s každým)
- Vybereme shlukovací algoritmus
- Stanovíme počet shluků – jen u některých metod
- Aplikujeme shlukovací algoritmus na matici vzdálenosti
  → získáme shluky

Shlukovací algoritmy:

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

Metriky vzdálenosti

Euklideovská vzdálenost

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

- Euklideovská vzdálenost:

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

- Standardizovaná Euklideovská vzdálenost:

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

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

Manhattanova vzdálenost

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

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

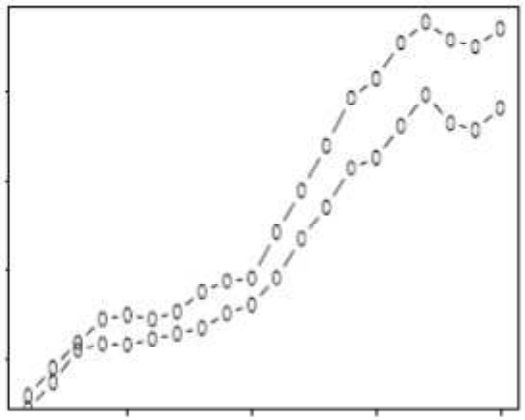
Robustnější vůči odlehlým hodnotám.

Metriky vzdálenosti založené na korelačním koeficientu

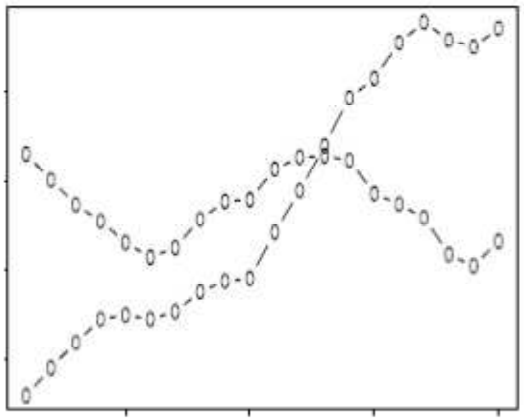
Obzvláště často využívané u dat z omics experimentů

Můžeme odvodit dvě různé metriky: $d_1(x,y) = [1 - r(x,y)]/2$

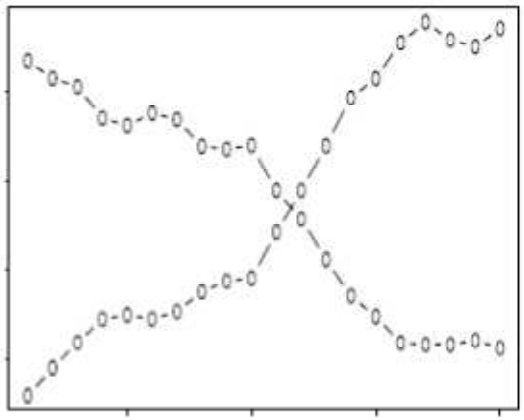
$d_2(x,y) = 1 - [r(x,y)]^2$



d_1=0.05, d_2=0.19



d_1=0.5, d_2=1



d_1=0.95, d_2=0.19

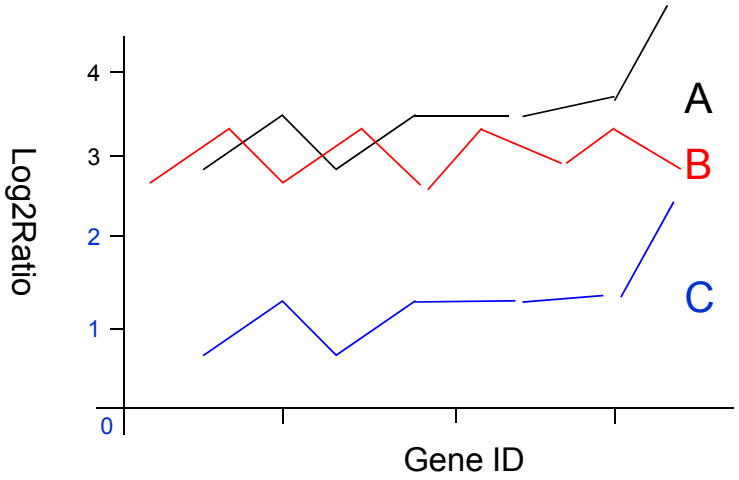
Pří použití ***d_1*** budou geny s opačnými profily patřit do odlišných shluků, zatímco při použití metriky ***d_2*** budou patřit do toho stejného shluku. Pokud chceme shluky interpretovat jako množiny genů ze stejné regulační sítě, použijeme raději d_2.

Výběr metriky

Výběr metriky záleží na tom, jaký typ podobnosti nás zajímá

Pokud nás zajímá **podobnost hodnot** (A a B jsou podobné), aplikujeme Euklidovskou vzdálenost

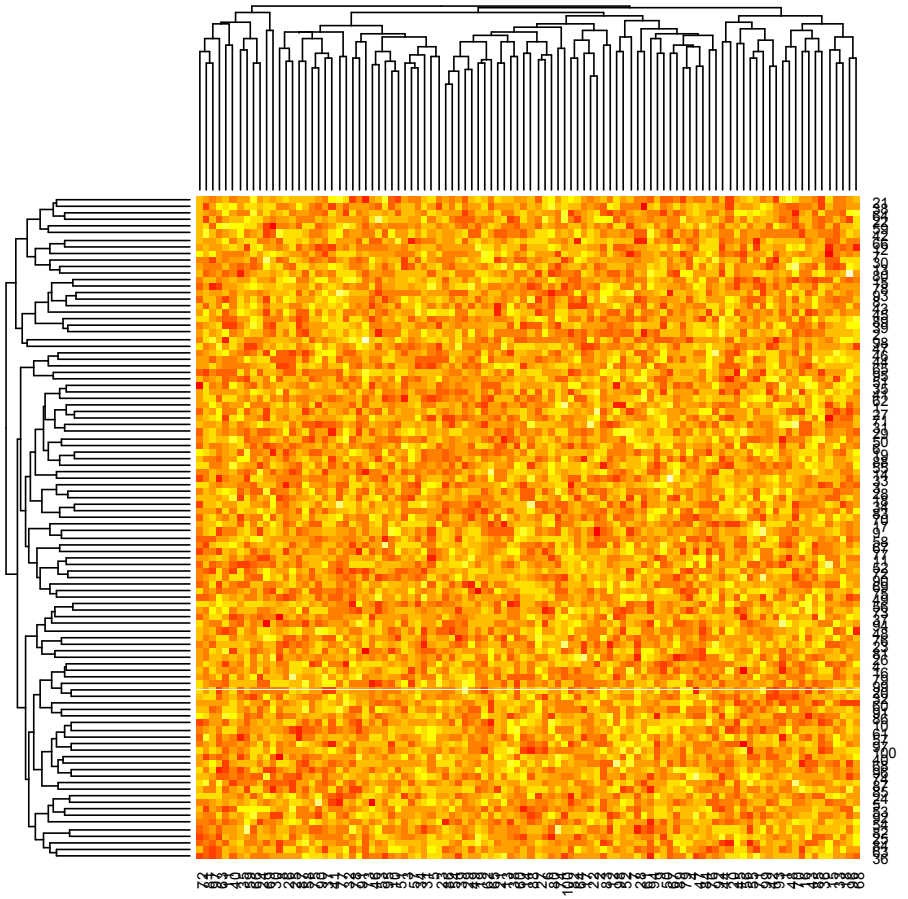
Pokud nás zajímá **vzor hodnot** (A a C jsou podobné), aplikujeme vzdálenost založenou na korelaci



Na co si dávat pozor I.

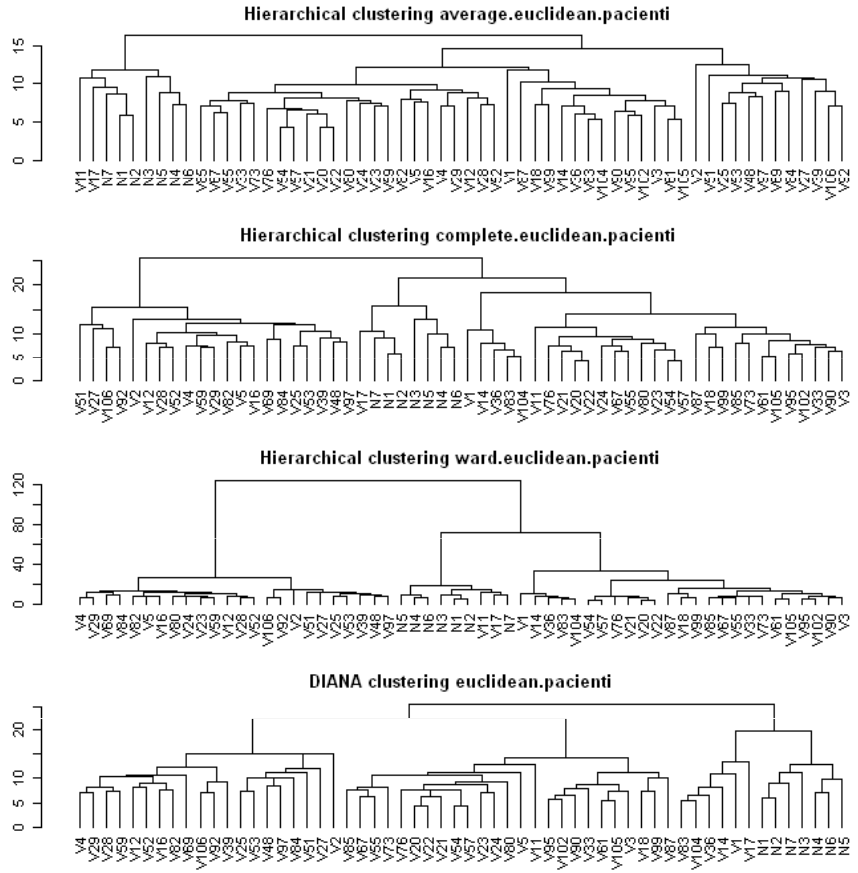
Mnoho shlukovacích technik najde shluky i v datech, ve kterých nejsou žádné přirozené shluky, jen proto, že byly pro tento účel vytvořeny.

Příkladem je právě hierarchické shlukování!



Na co si dávat pozor II.

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



Další problémy

Výběr shlukovacího algoritmu a metriky ovlivňuje konečné výsledky

Výsledky jsou závislé na samotných datech

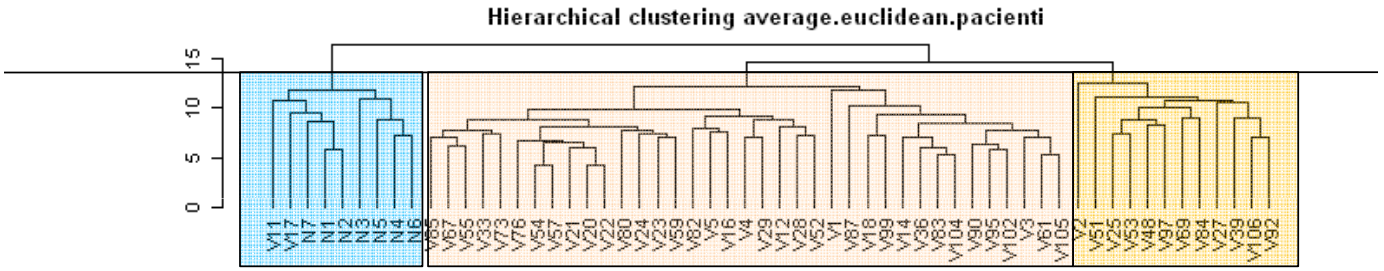
Kolik shluků?

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

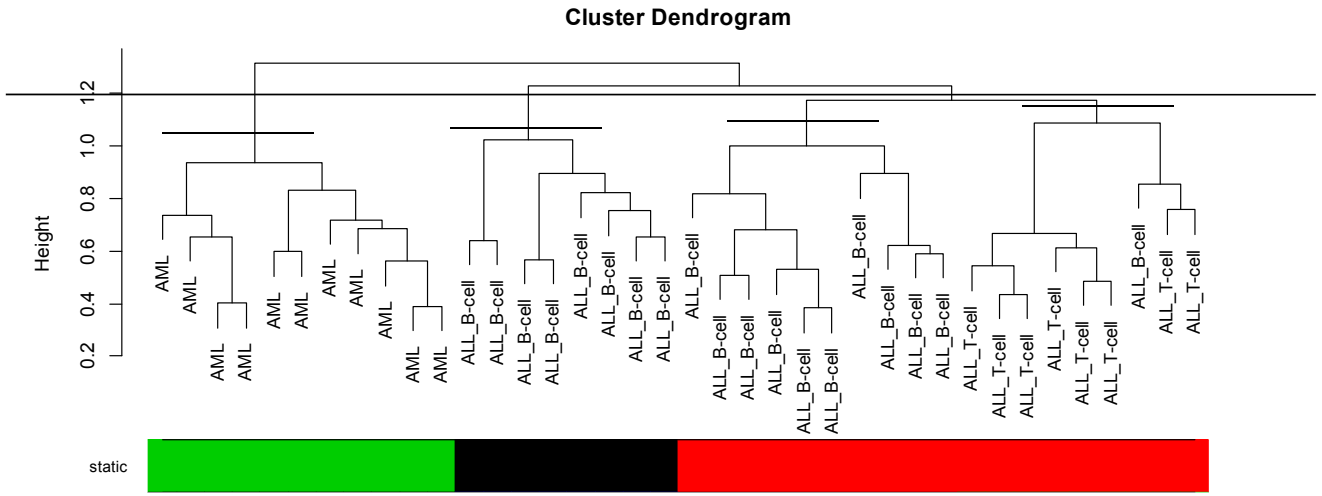


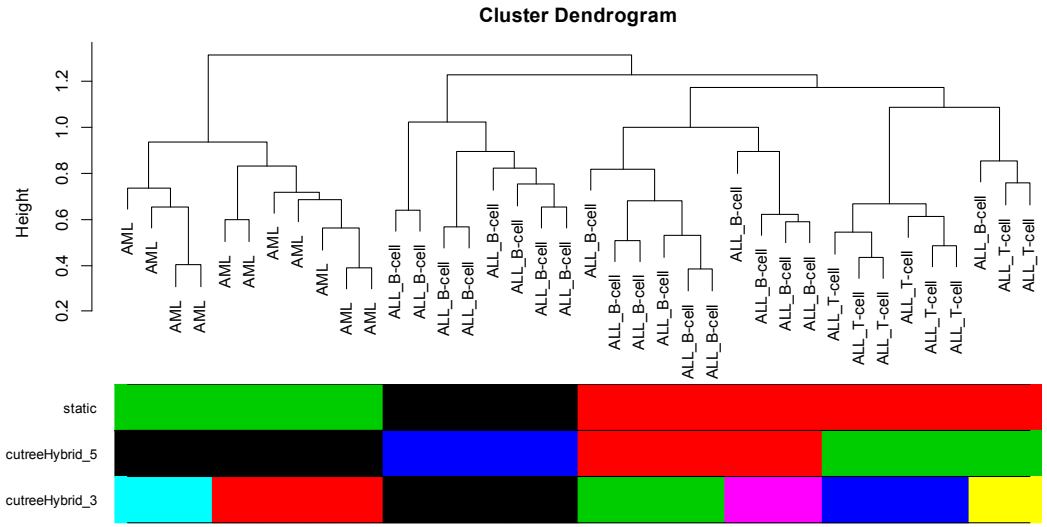
- Počet shluků je následně určený tak, aby heterogenita v rámci shluků byla co nejmenší a mezi shluky co největší
- Různé metriky heterogenity shluků – variabilita, Silhouette, ...

Řezání dendrogamu jeho problém

- U hiararchického shlukování se **stanovuje fixní výška řezu** dendrogramu

- Problém: u omicsových molekulárních dat se často vyskytují **shluky v různých výškách řezu**





- Metoda dynamického prořezávání dendrogramu (Langfelder et al, 2007)
- Dynamické řezání dendrogramu na základě minimální velikosti shluků, maximální výšky řezu a dalších parametrů
- Pokud vzorek nesplňuje kritéria, není zařazen do shluku!

Dynamic hybrid tree cut

Dynamic hybrid tree cut - základy

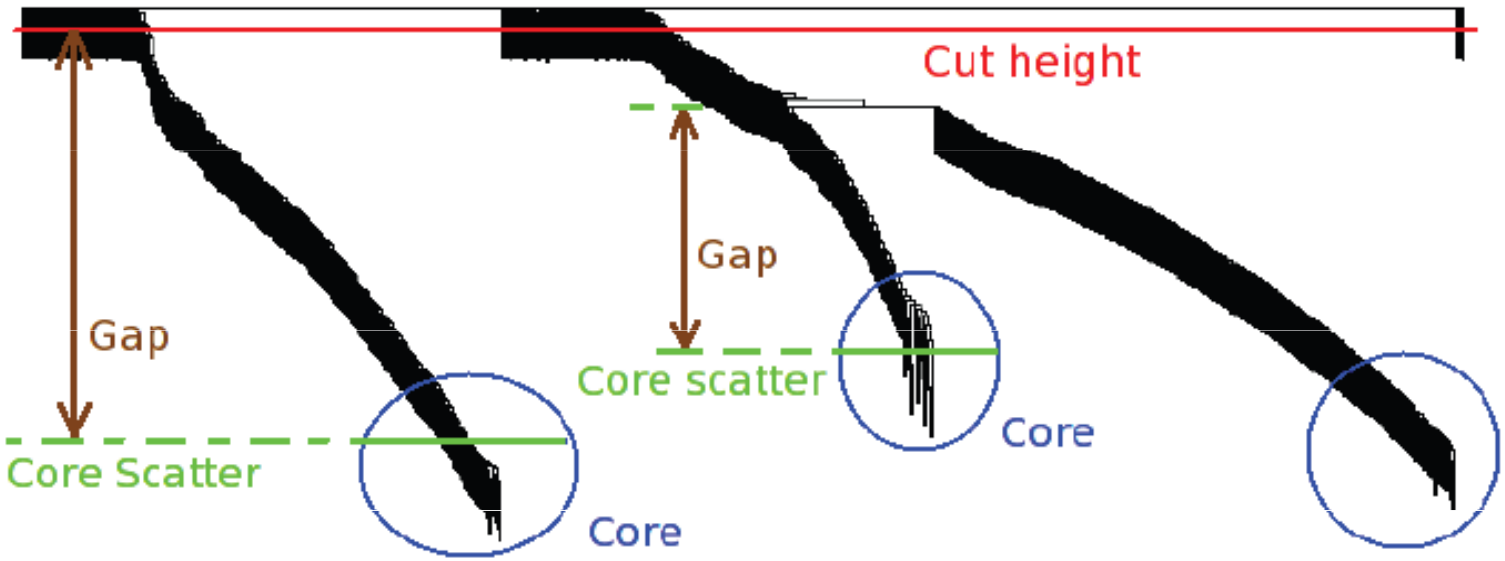
Shluk musí mít minimální počet objektů

Příliš vzdálené objekty jsou ze shluku vyloučené i v případě, že patří do stejného ramena dendrogramu

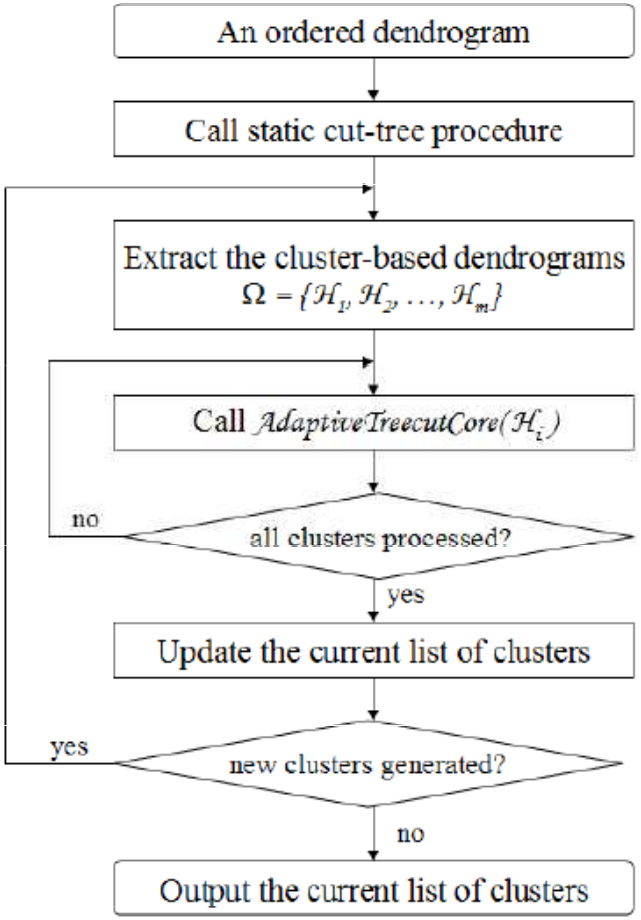
Každý shluk by měl být oddělený od okolí mezerou

Jádro (nejnižší propojené objekty) každého shluku musí být silně propojeno

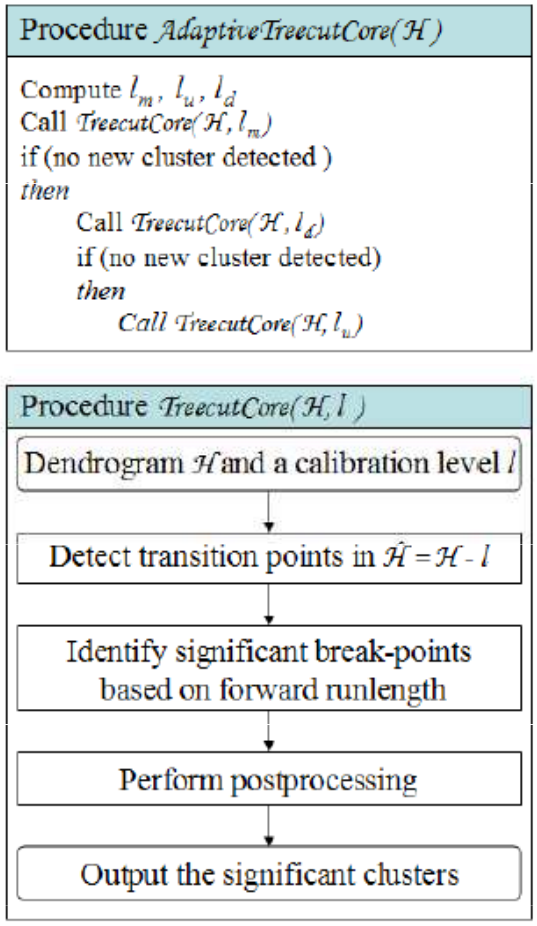
Parametry dynamic hybrid cut



Algoritmus dynamic hybrid tree cut



(a)



(b)

Robustní shlukování

V analýze vysocepokryvních molekulárních dat mají výše uvedené problémy větší váhu

Malý počet vzorků a vysoký počet proměnných (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.

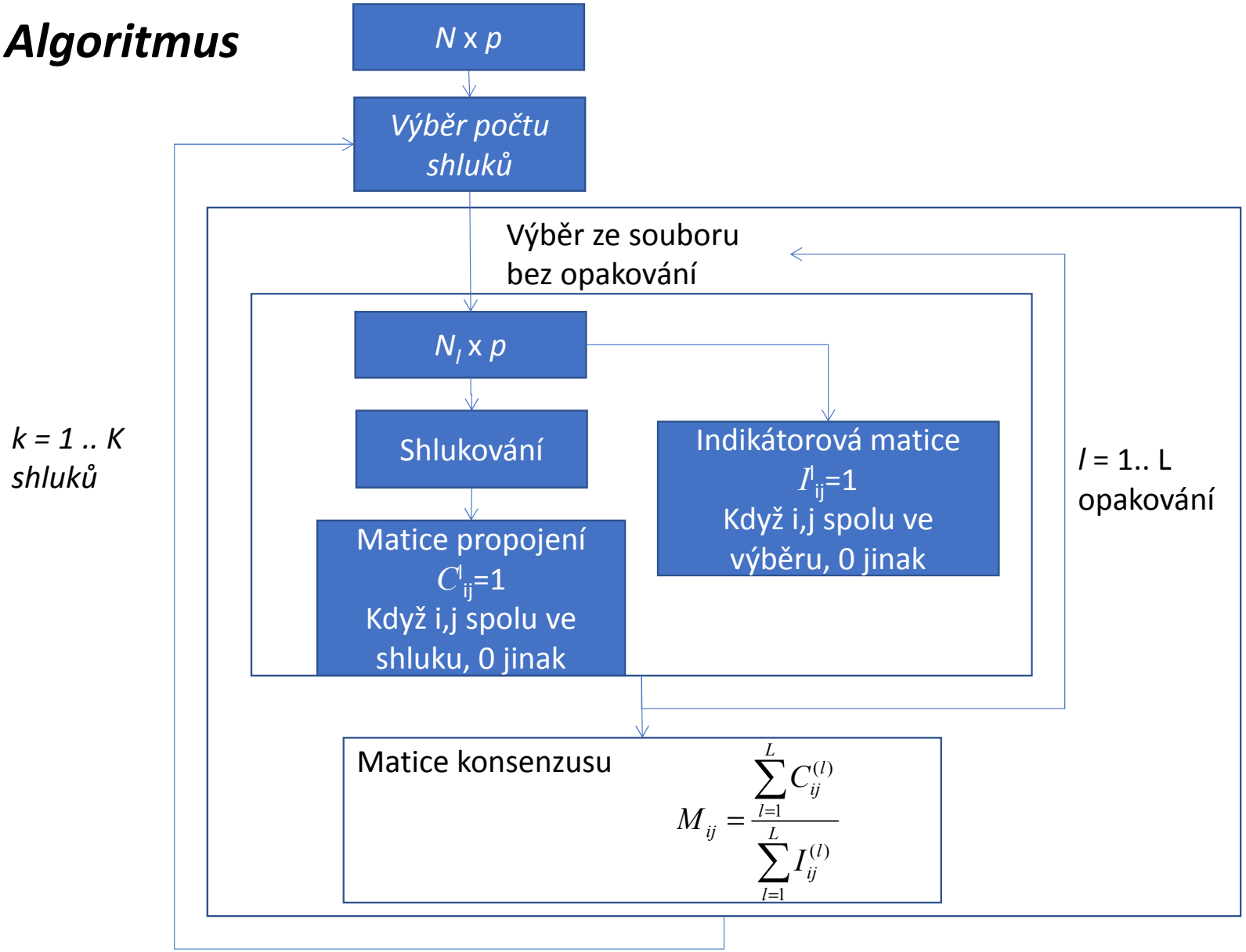
Základní princip konsenzuálního shlukování

(i) Rozrušení 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

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

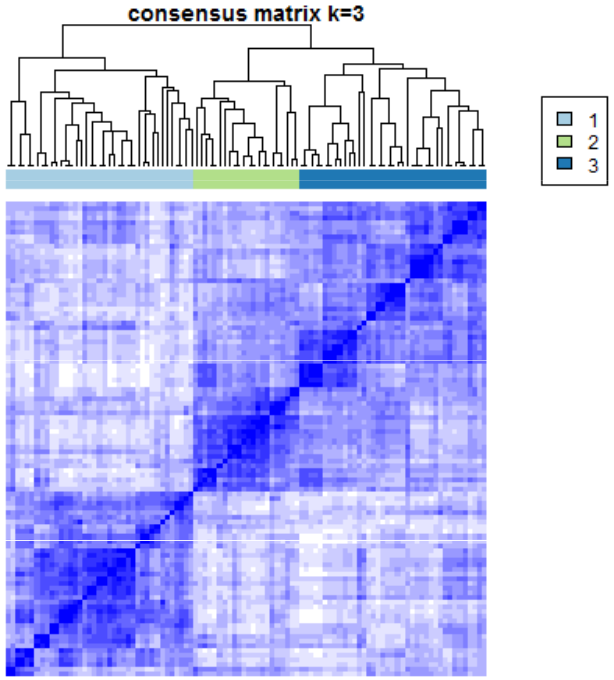
- ***Matice konsenzusu M*** je definovaná jako:

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

Myšlenka konsenzuálního shlukování

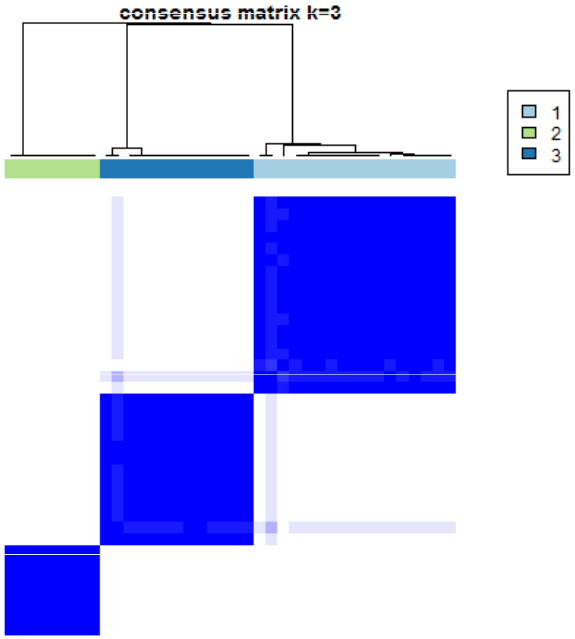
- Pokud se dva vzorky v jednotlivých výběrech nacházejí často spolu ve shluku, jsou důvěryhodnějšími členy shluku než ty, které se ve shluku nacházejí méně často

A.



Data bez struktury (náhodný výběr z normálního rozložení)

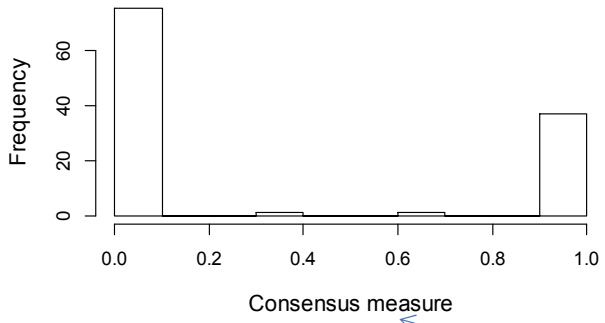
B.

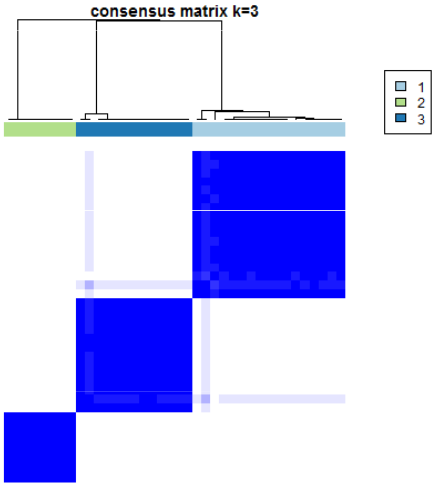


Data se třemi skupinami

Odhad počtu shluků I

Histogram míry konsenzusu



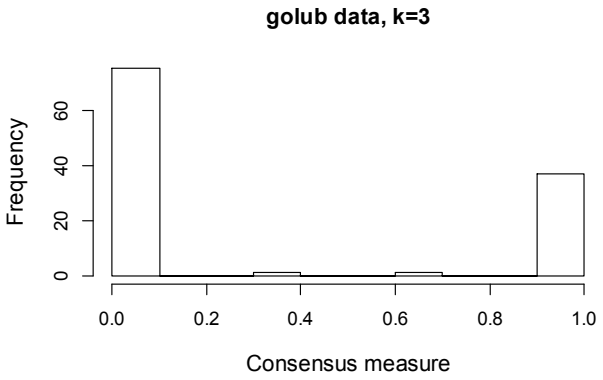


Konsenzus mezi dvěma vzorky

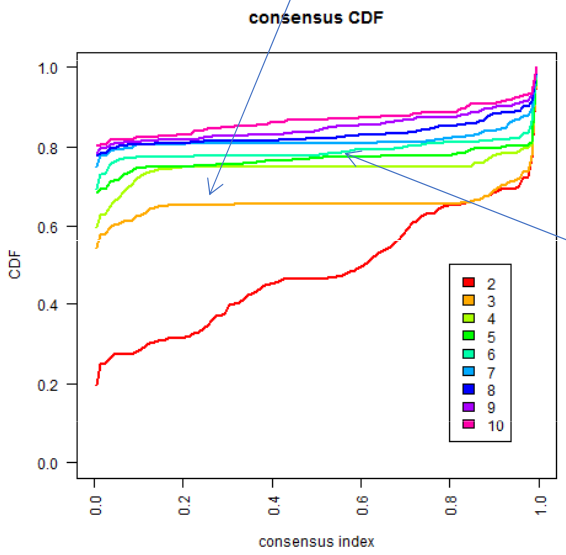
$$CDF^{x} = \frac{\sum_{i<j} 1\{M_{ij} \leq x\}}{N(N-1)/2}$$

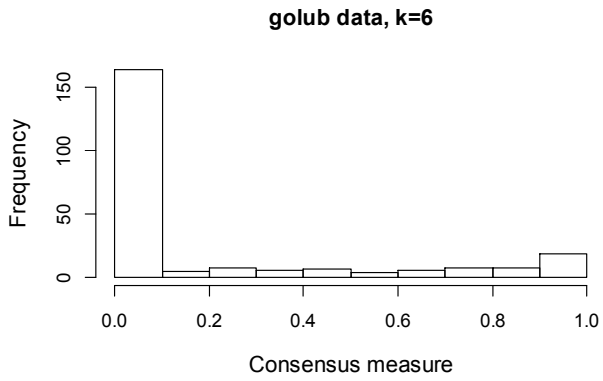
Kumulativní distribuční funkce

Odhad počtu shluků II



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





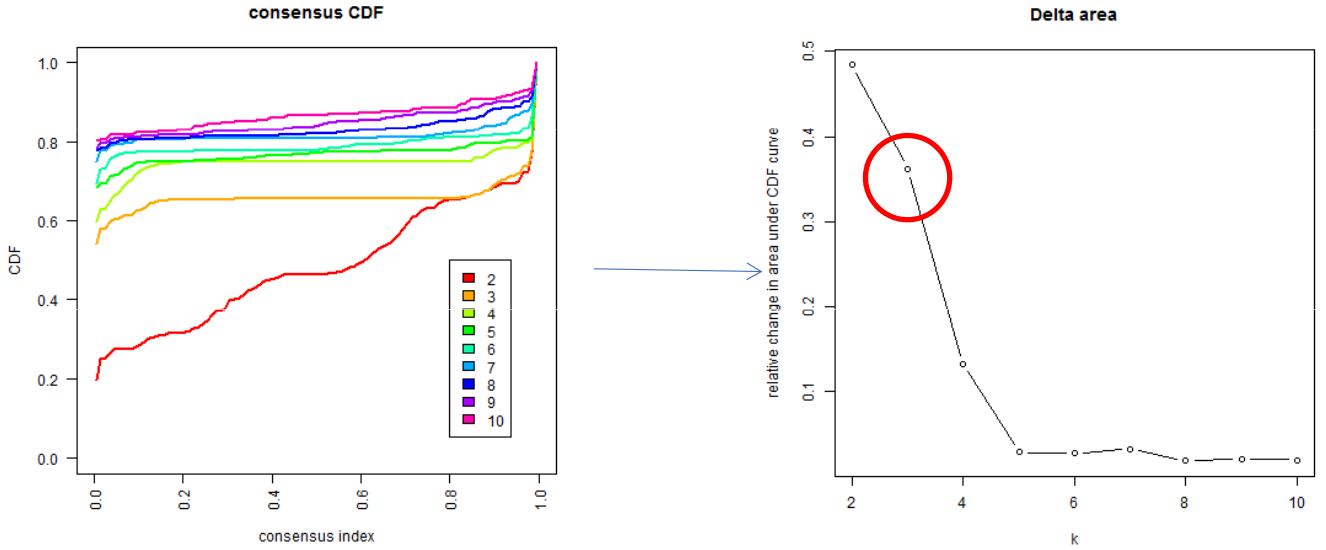
6 shluků má podstatně míň vzorků s konsenzusem 1 a tím pádem jsou tyto shluky míň důvěryhodné

Struktura se 3 shluky naopak vypadá jako optimum

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

Odhad počtu zhlukov 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)

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
    - 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 ($1\{s_i \in I_k\}$) 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 $(g=1,\ldots,n)$:

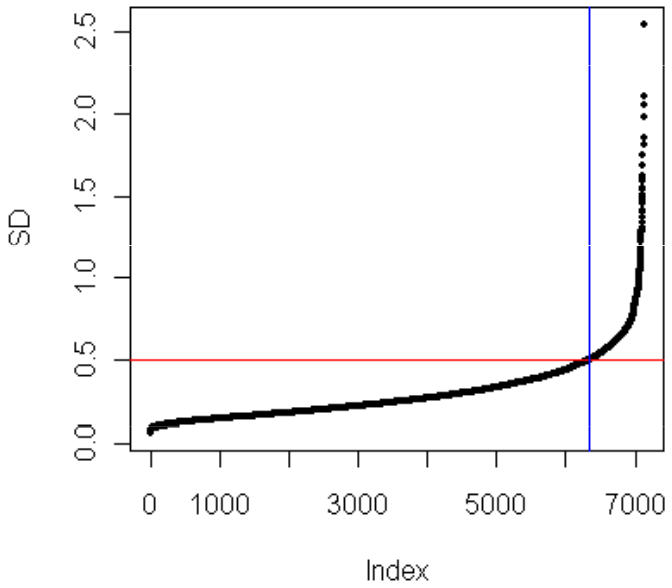
  $$\log \mathcal{L}(\{X_g\}; \{\pi_k, \theta_k\}) = \sum_g \log\left[\sum_k \pi_k \, \Phi(X_g, \theta_k)\right]$$

- 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

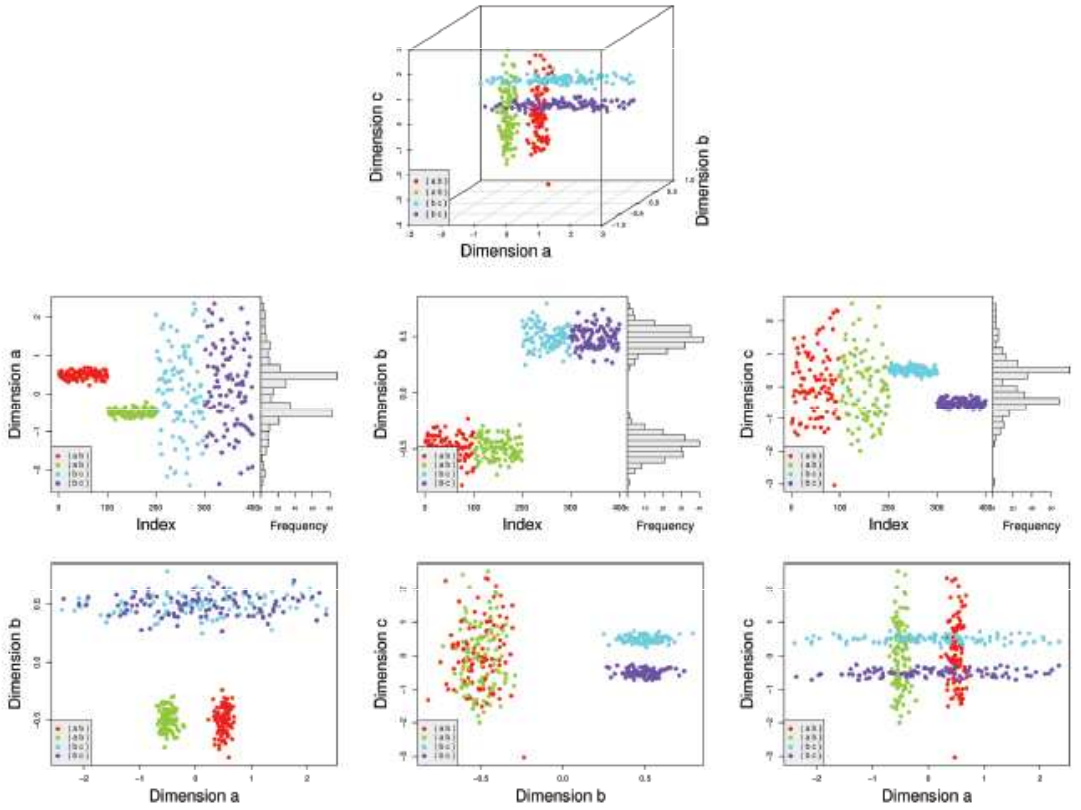
Problémy shlukování u omicsových molekulárních dat

- Šum v datech negativně ovlivňuje výsledky

- **Jedna veľká množina genů (proteinů, metabolitů) ze stejného biologického motivu ovplyvní zásadně shluky**

Kde hledat shluky

Data můžou vytvářet shluky v odlišných dimenzích



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

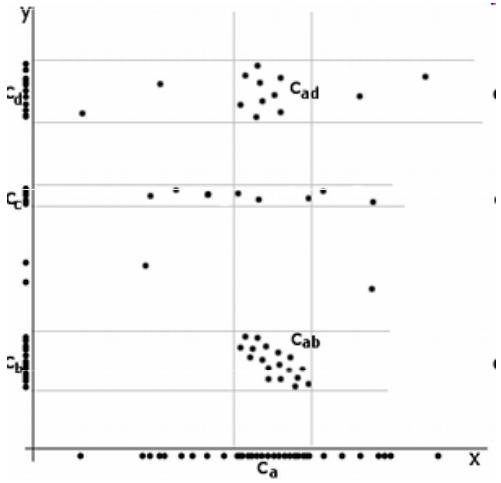
Podprostorové shlukovaní

- 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)

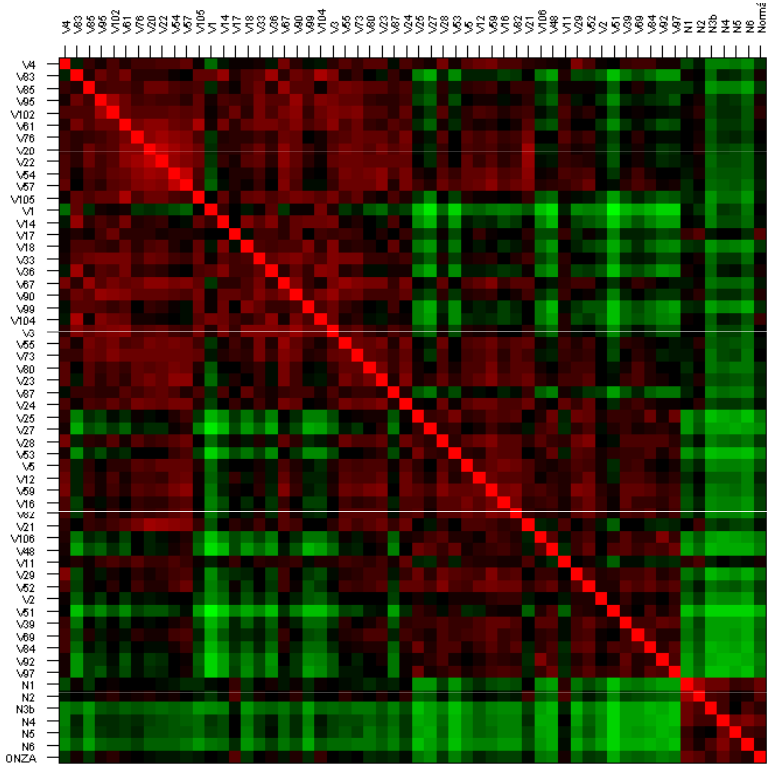
> library(orclus)



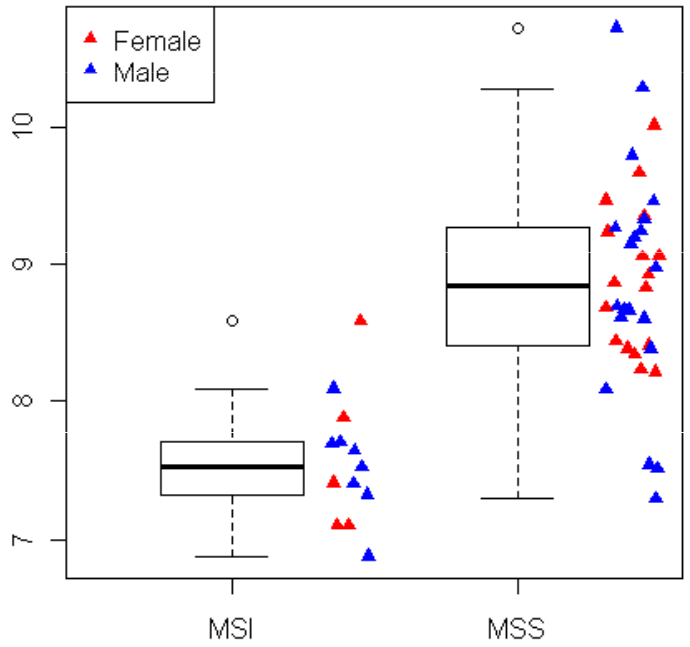
Vizualizace výsledků

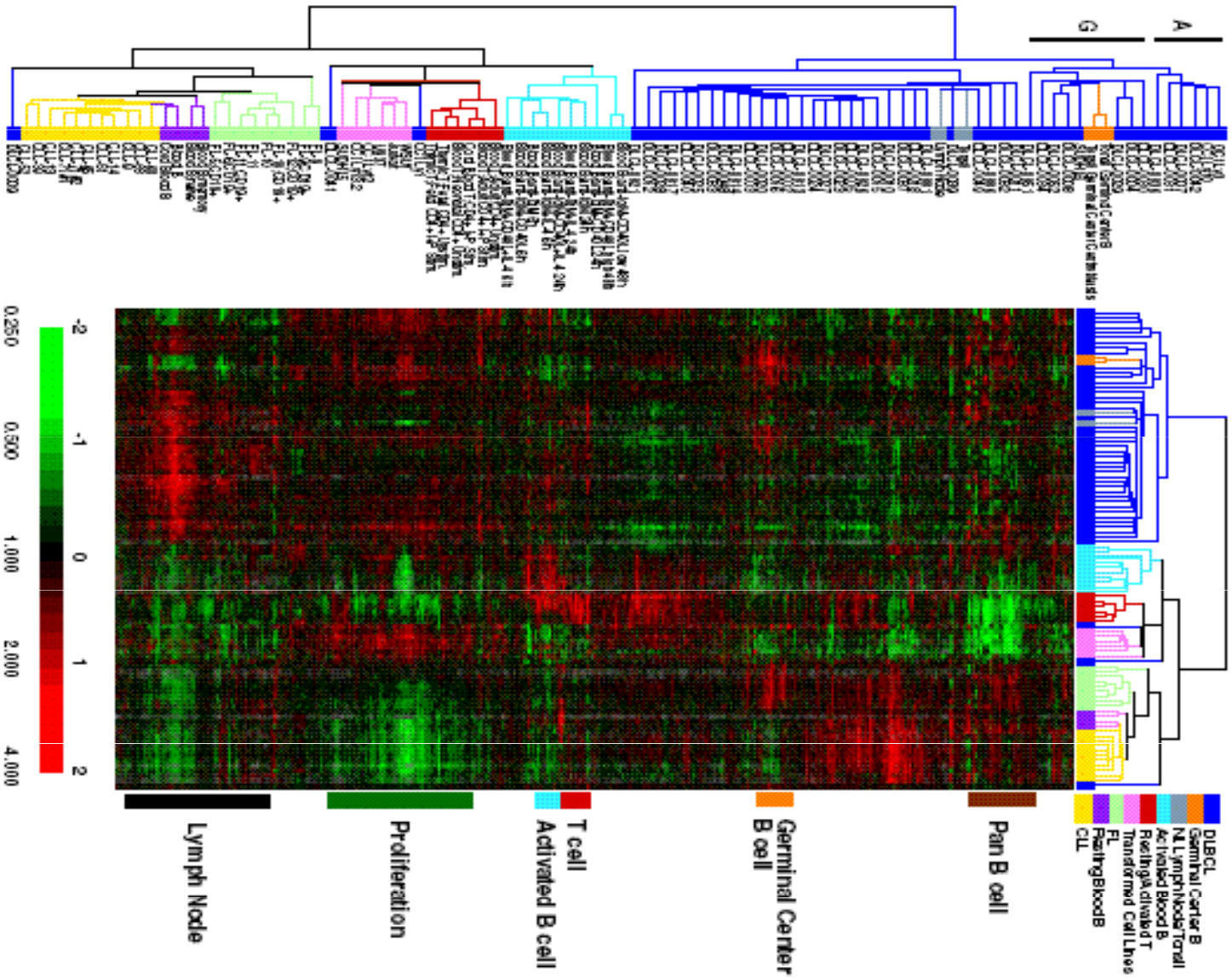
- Správná vizualizace výsledků je nejdůležitější součást analýzy!

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?)

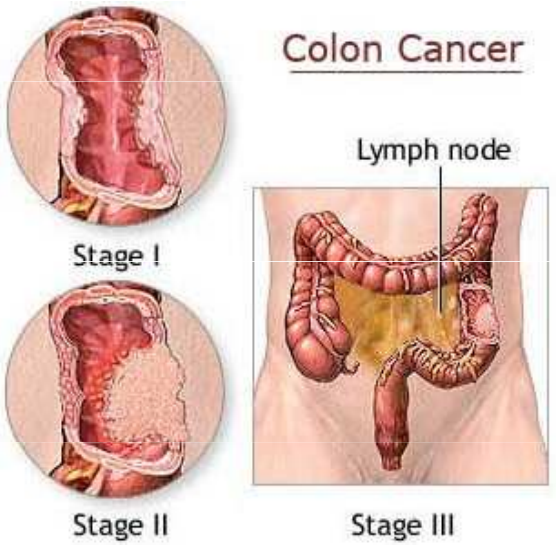
(Je rozložení těchto charakteristik mezi podtypy srovnatelné ve validačním souboru?)

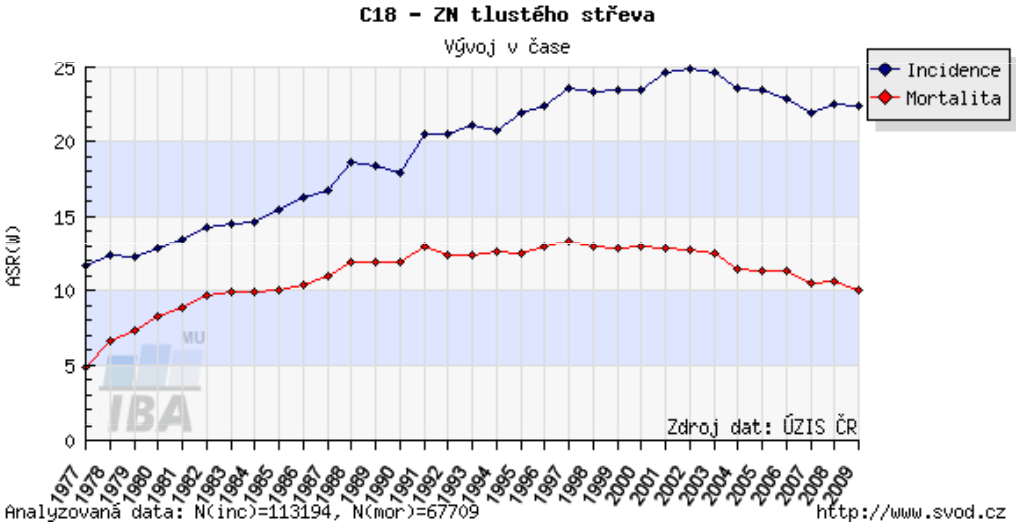
Shrnutí

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

Kolorektálny karcinóm

príklad





Heterogénne ochorenie s rozdielnou odpoveďou na terapiu.

Len niekoľko klinicky používaných markerov:

- *BRAF/KRAS mutácia – pre kvalifikáciu na antiEGFR terapiu (u štádia s metastázami)*
- *MSI – mikrosatelitová nestabilita – všeobecne považovaná za dobrý marker*

Cieľ:

Nájsť skupiny nádorov kolorekta s podobnou expresiou génov (podobným génovým profilom) ~ podtypy

Charakterizovať tieto podtypy pomocou klinických a známych molekulárnych parametrov.

Dátové súbory:

Matica obsahujúca kvantitatívnu expresiu génovej aktivity nádorov

Gény

N x p, k=1..5

Vzorky nádorov

Matica klinických a molekulárnych parametrov ku každej nádorovej vzorke, vrátane prežitia pacienta

VÝBER PREMENNÝCH / Zmenšenie dimenzie dát

Nezávislý výber – z 25 000 génov, výber podmnožiny 3025 génov s najvyššou variabilitou v súbore

Redukcia dimenzionality – práca s génovými modulmi

génový modul – sada génov s rovnakou génovou expresiou

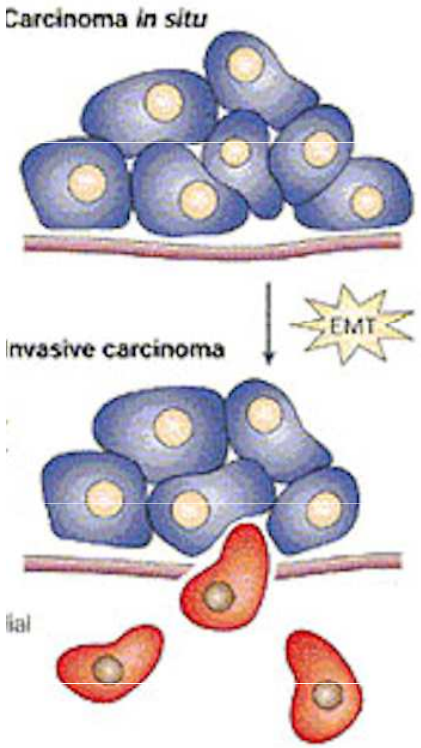
Jedná sa o istú formu váženia efektu biologických motívov

Predpoklad:

sada korelovaných génov ~ biologický motív

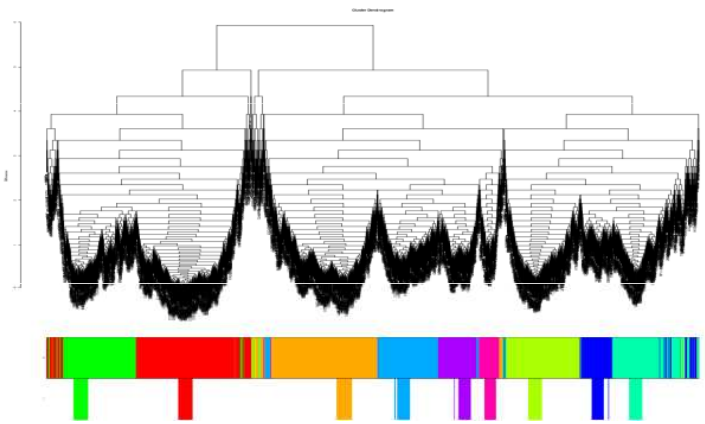
Príklad EMT

- EMT – epiteliálno mesenchymálny prechod
- Génová expresia podobná zdravému mezenchymálnemu tkanivu
- Obvykle reprezentovaný **zmenou v stovkách génov**
- Identifikácia modulu EMT a jeho reprezentácia jedinou hodnotou (priemerom) **zmenší jeho efekt v zhlukovaní** a dá šancu **ďalším dôležitým procesom** reprezentovaným menším množstvom génov



3025 génov
Dátový súbor 1

Pearsonova korelácia,
Hierarchické zhlukovanie
Complete linkage
Rezanie dendrogramu



150 génových modulov

Validácia ich korelácií v
4 ďalších dátových súboroch

Výber 54 modulov
(625 génov)

Medián génov v module

54 Meta-génov

8 hlavných zhlukov
~ hlavné biologické
motívy

Pearsonova korelácia,
Hierarchické zhlukovanie
Complete linkage
Rezanie dendrogramu

Identifikácia biologických motívov

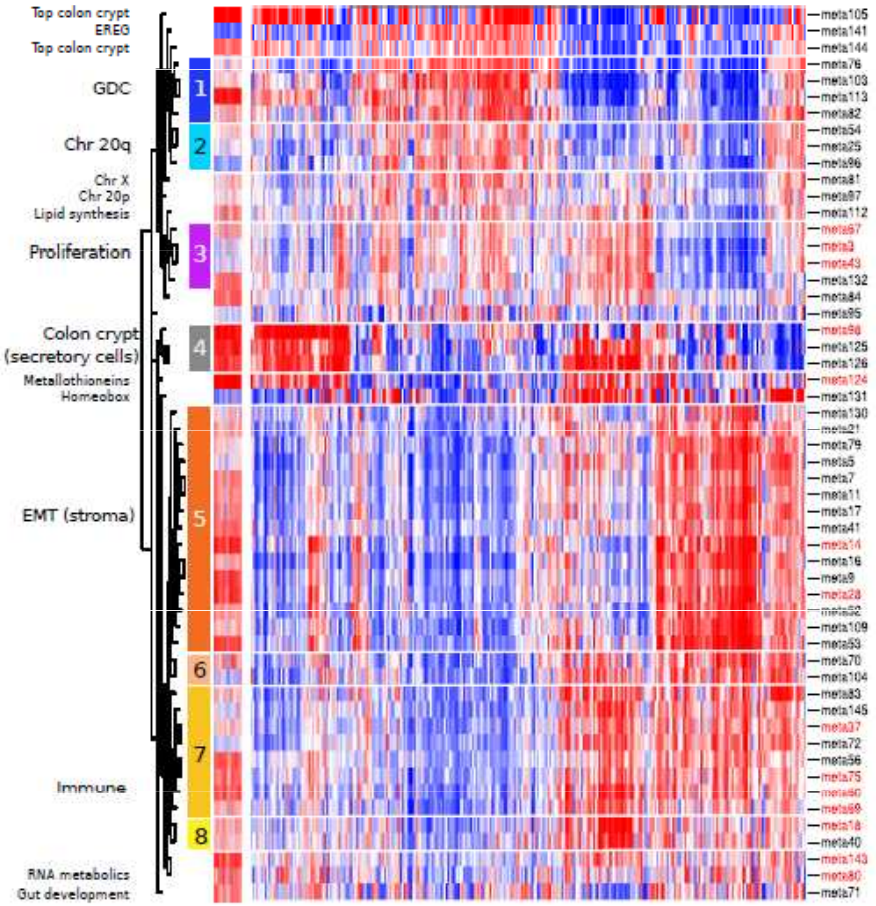
- Analýza génových sád

	V genóme	V biol motíve
V genóme	a	b
V module	c	d

Biol motív:

EMT, proliferácia,

Chromozóm 20q...



Cieľ:

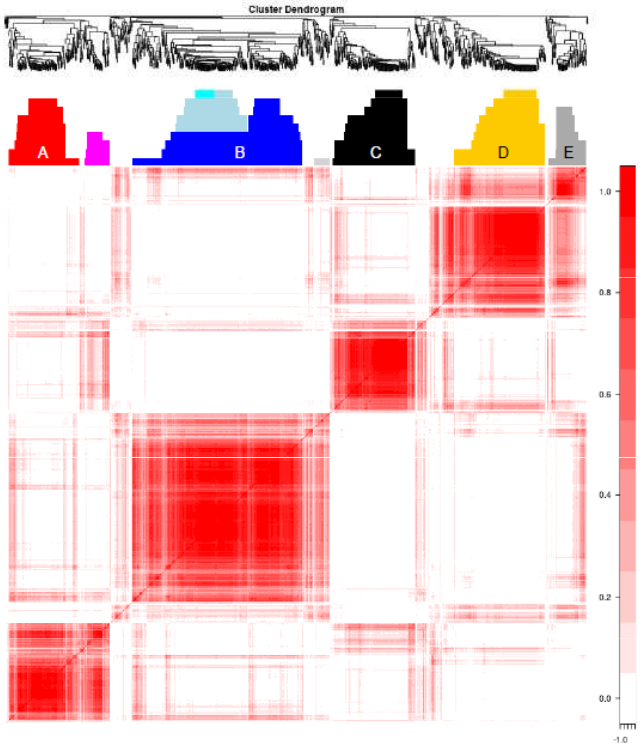
Nájsť skupiny nádorov kolorekta s podobnou expresiou génov (podobným génovým profilom) ~ podtypy

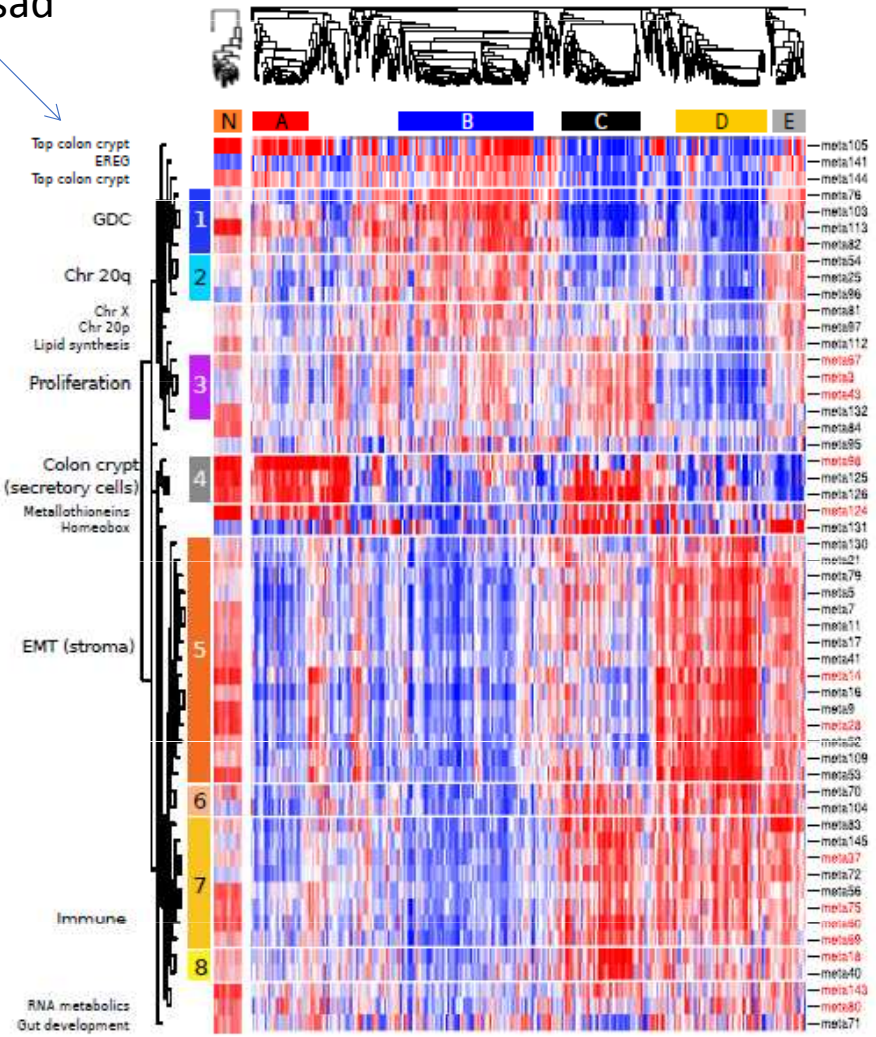
Charakterizovať tieto podtypy pomocou klinických a známych molekulárnych parametrov.

Motívy génovej expresie v podtypoch

Analýza génových sád

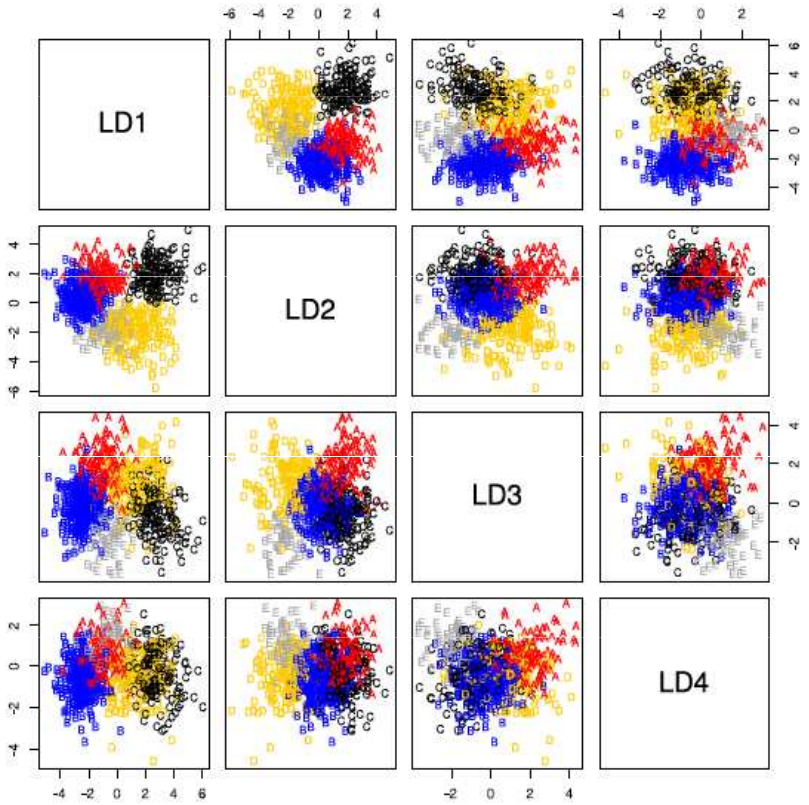
Hierarchické zhlukovanie na matici konsenzusu





Expresné profily pre jednotlivé podtypy

- Klasifikátor LDA (linear discriminant analysis)



Minimálna génová sada

- Selekcia génov pomocou ***elastic net –*** počíta s korelovanými génmi (neodstraňuje len jeden gén, ale všetky korelované)
- Klasifikátor vytvorený na selektovaných génoch (shrunken centroids)

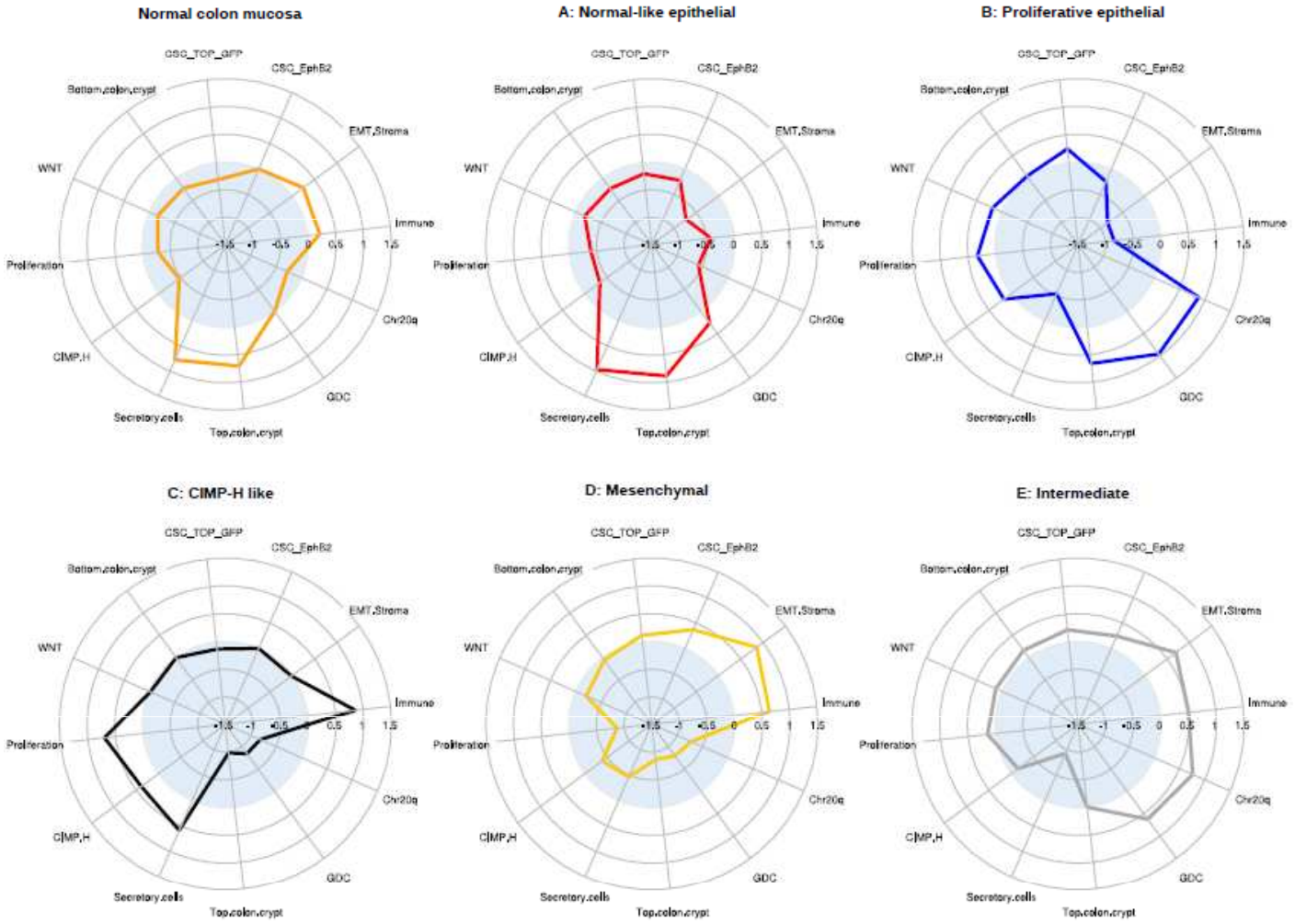
Subtype	Minimálna génová sada
A	CLCA1, PADI2, ADTRP, RETNLB, TIMP3, MUC2, FNDC1, NR3C2, SULF1, B3GNT7, STYK1, CHI3L1
B	FARP1, ALOX5, FSCN1, HNF4A, RARRES3, MYRIP, GPSM2, TSPAN6, CCDC113, CDHR1, KCTD12, SGK1, BASP1, MT1E, GPX8, RPS6KA3, SOCS3, SLC5A6, PRR15, PLAGL2, IHH, CREB3L1, TP53RK, YAE1D1, EPB41L3, 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

Cieľ:

Nájsť skupiny nádorov kolorekta s podobnou expresiou génov (podobným génovým profilom) ~ podtypy

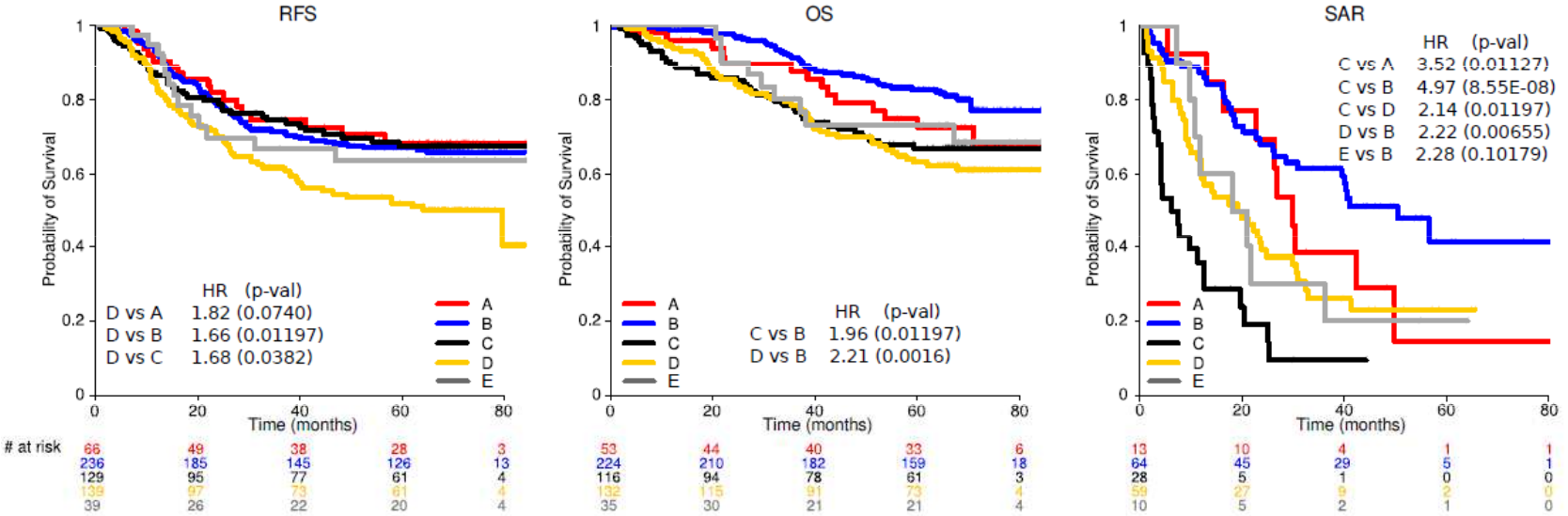
Charakterizovať tieto podtypy pomocou klinických a známych molekulárnych parametrov.

Vzor expresie biologických motívov



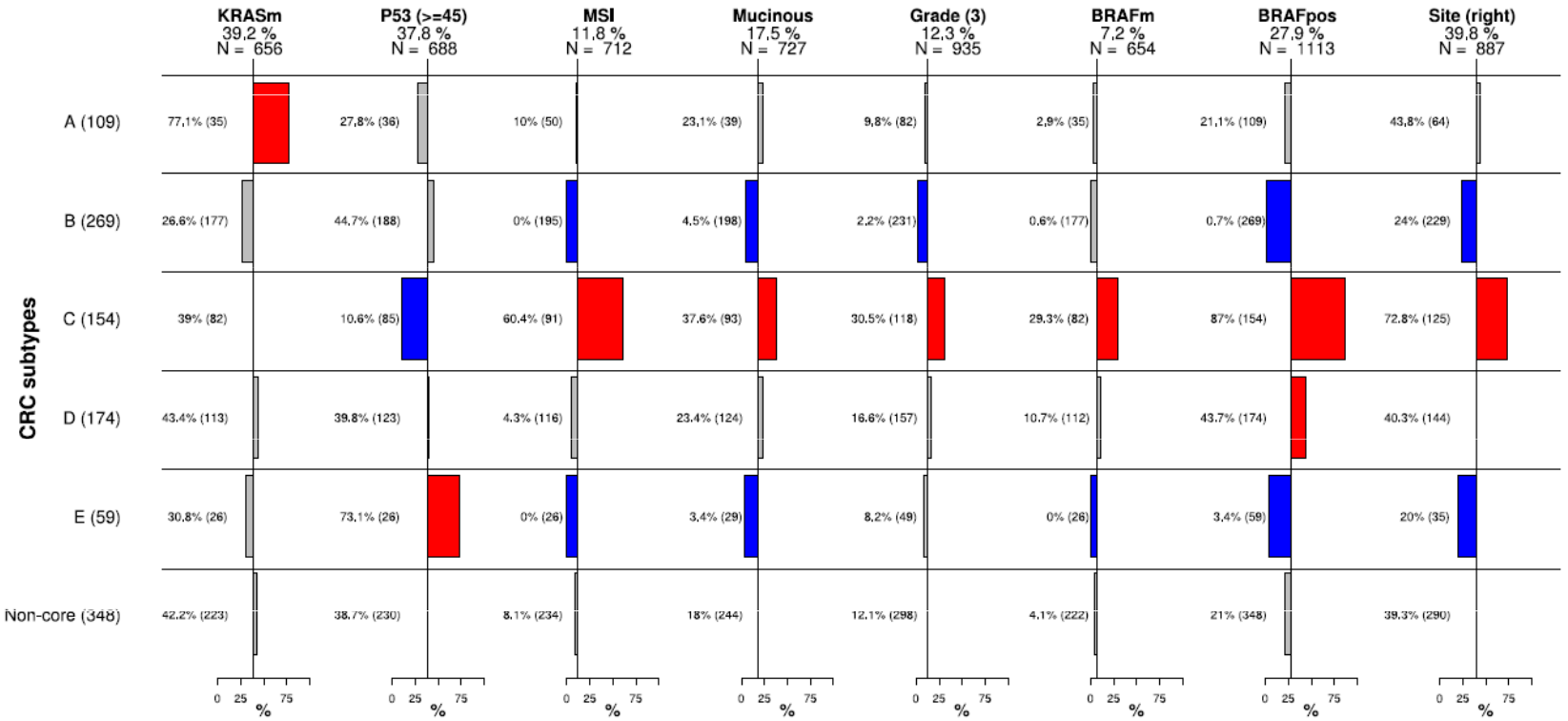
- Analýza génových sád pomocou KS testu

Rozdiely v prežití



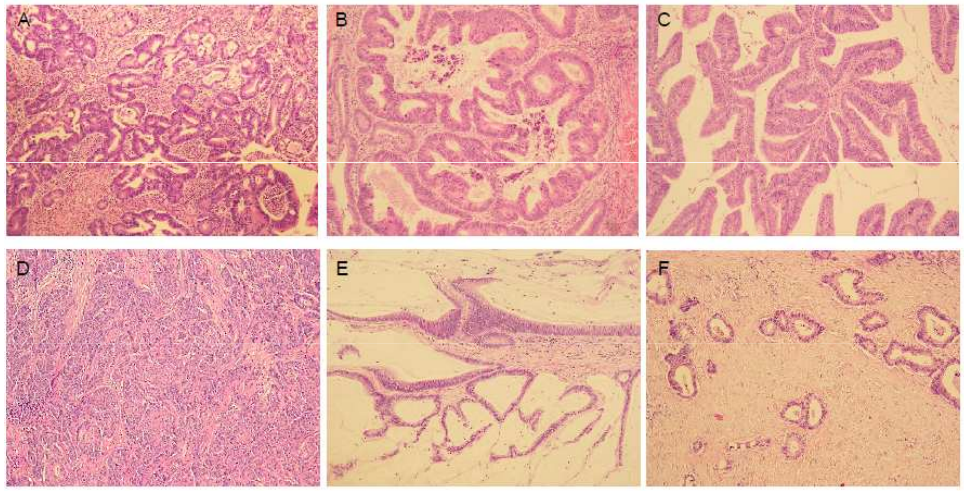
- Kaplan-Meierove krivky prežitia
- Coxov model proporcionálnych rizík (efekt štádia vs podtypov)

Charakterizácia podtypov klinickými a molekulárnymi premennými

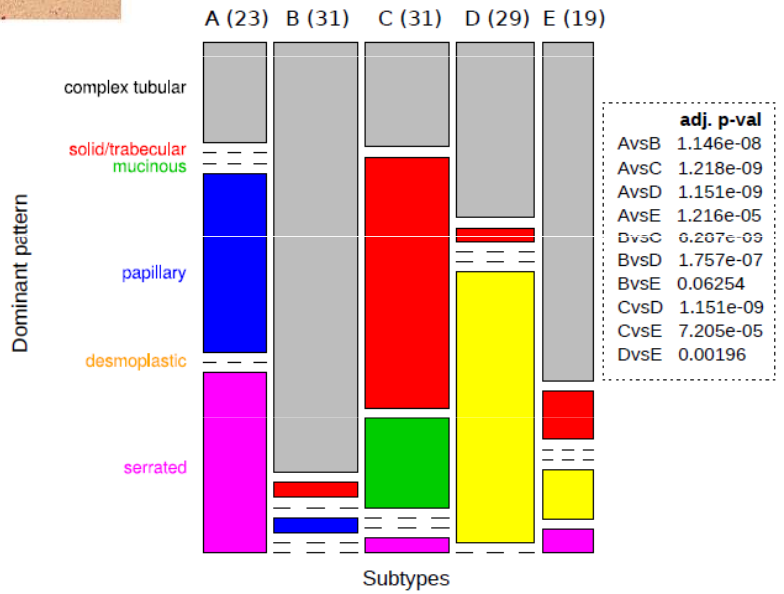


- Rozdiel od populačnej baseline u každého podtypu pomocou Fisherovho exaktného testu, FDR úprava p-hodnôt

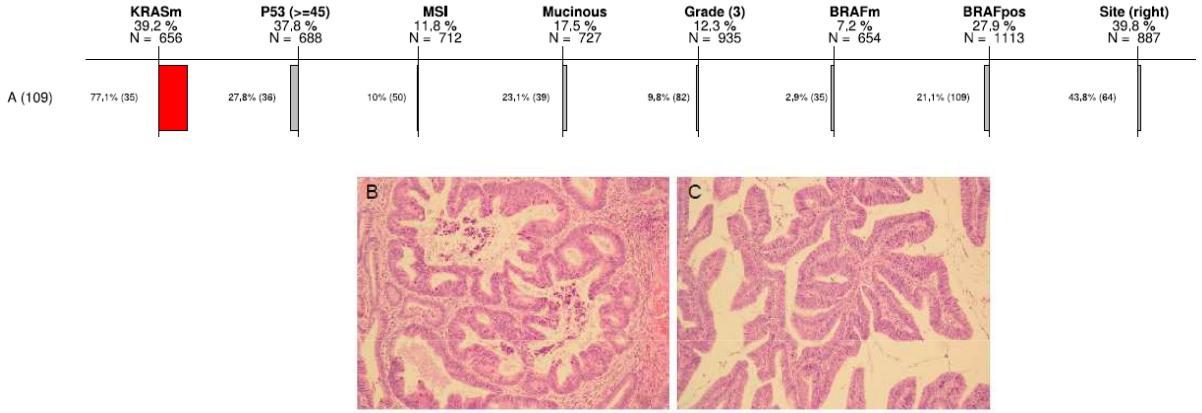
Histologické rozdiely

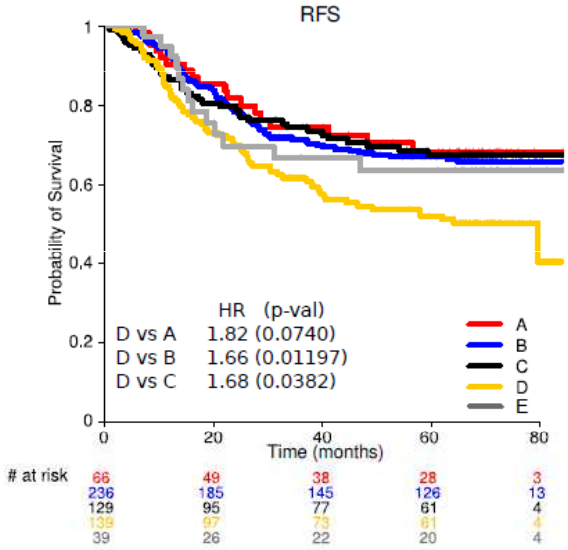


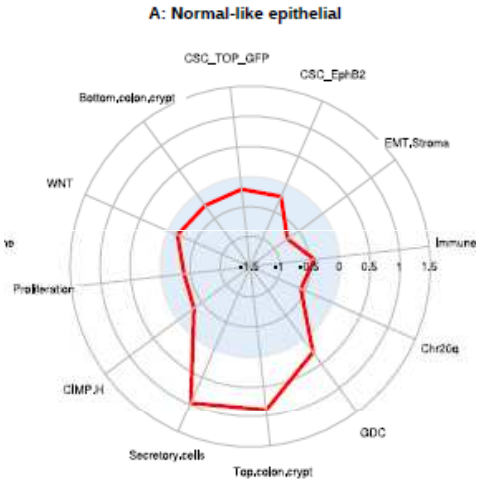
- Fisherov test, adjustácia na FDR

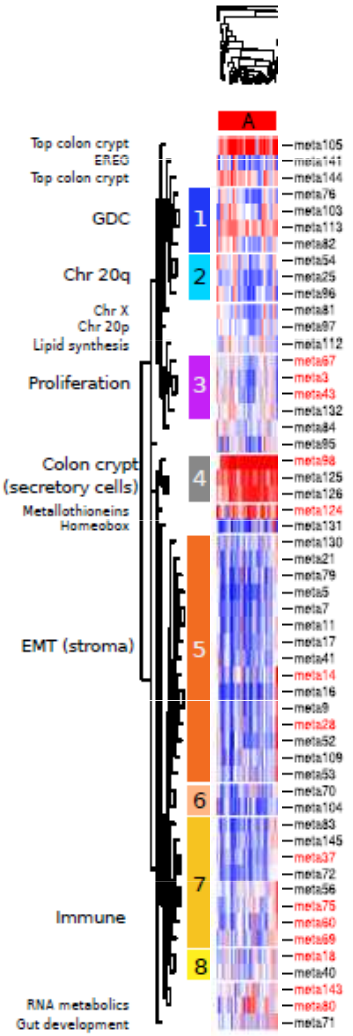


Podtyp A – Surface crypt like - *KRAS* mutanti, papillary a serrated morfotyp, najviac diferencovaný, bez aktívnej Wnt signálnej dráhy. Dobrý OS a RFS.

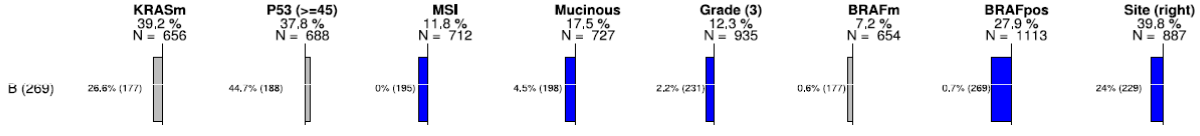


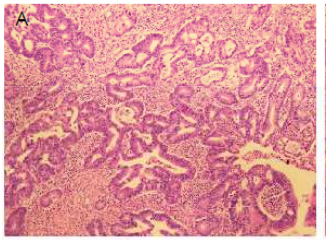


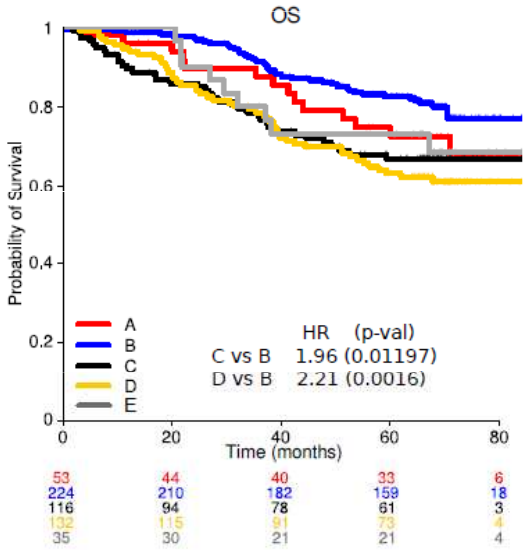


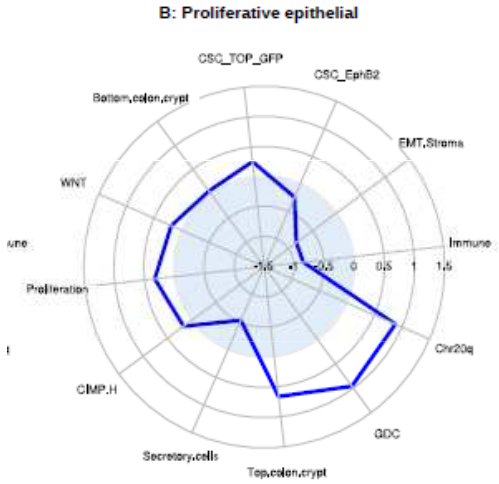


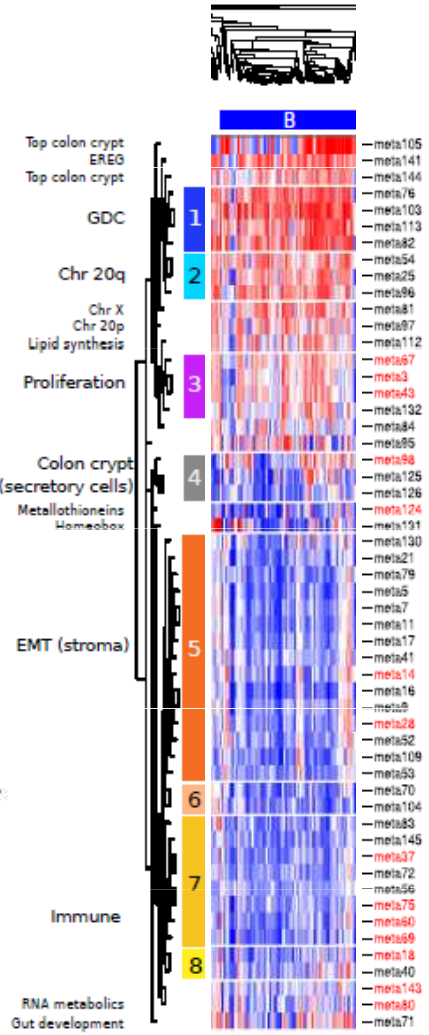
Podtyp B – Lower crypt like – diferencované ale bez sekrečných buniek, proliferujúce, a aktívnou Wnt signálnou dráhou. Komplexný tubulárny morfotyp. Časo MSS, BRAFwt, nižšieho grádu, dobré prežitie v OS, RFS i SAR.



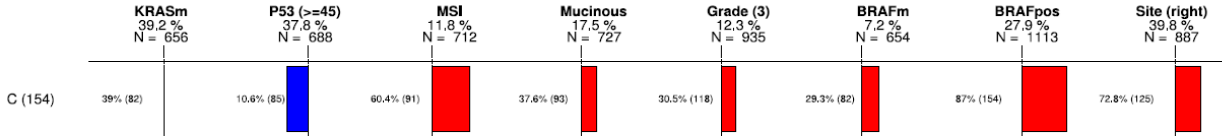


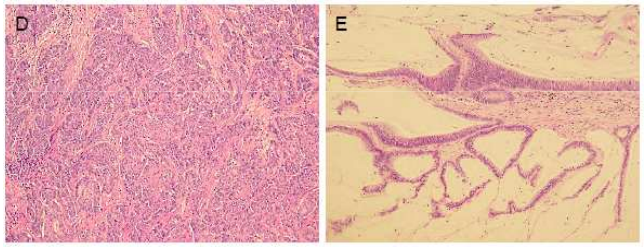


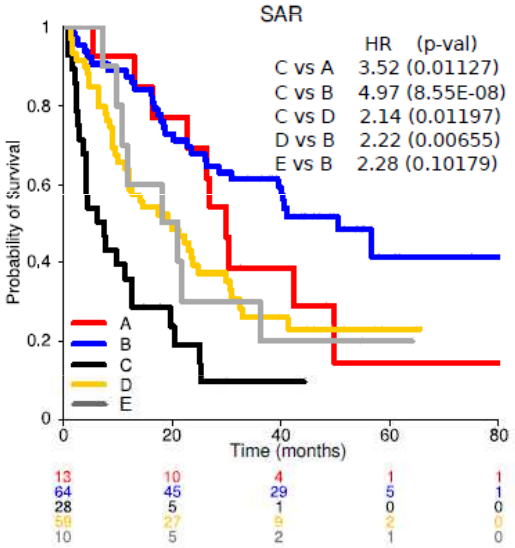


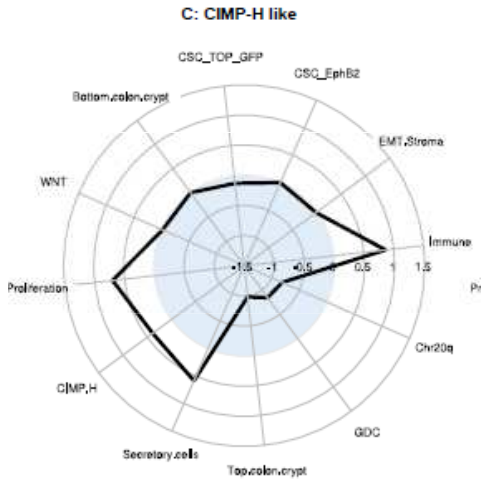


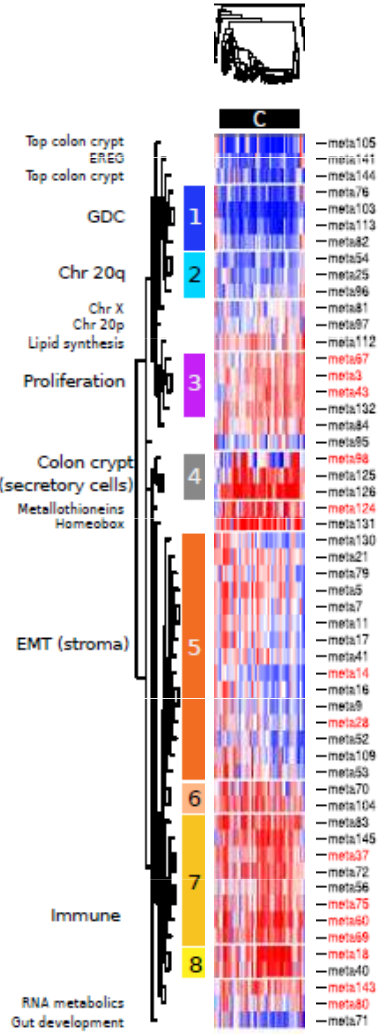
Podtyp C – CIMP-H like - časo MSI, *BRAF*-mutantné, hypermutované, z pravej časti hrubého čreva. Histologicky - horšie diferencované, solídno-trabekulárne s mucínovéým morfotypom. Aktívne proliferujú a majú silnú imunitnú reakciu. Dobrý RFS, ale zlý OS and SAR.



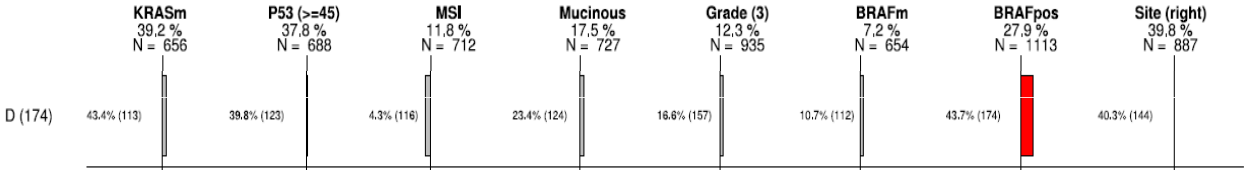


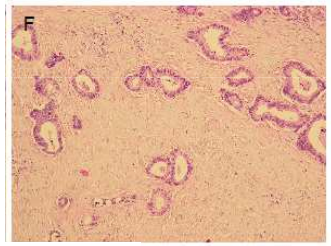


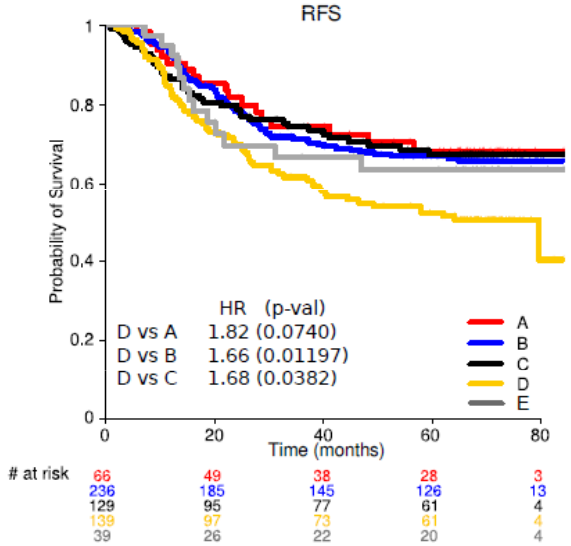


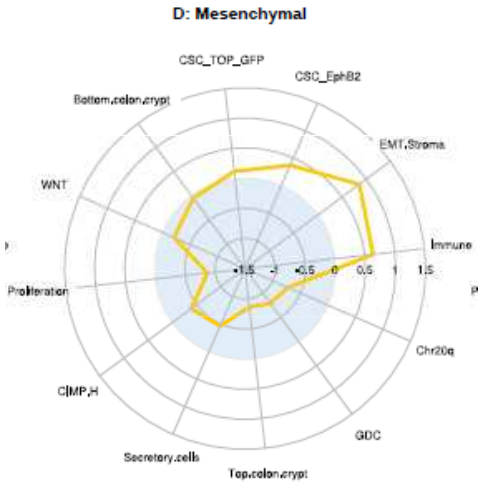


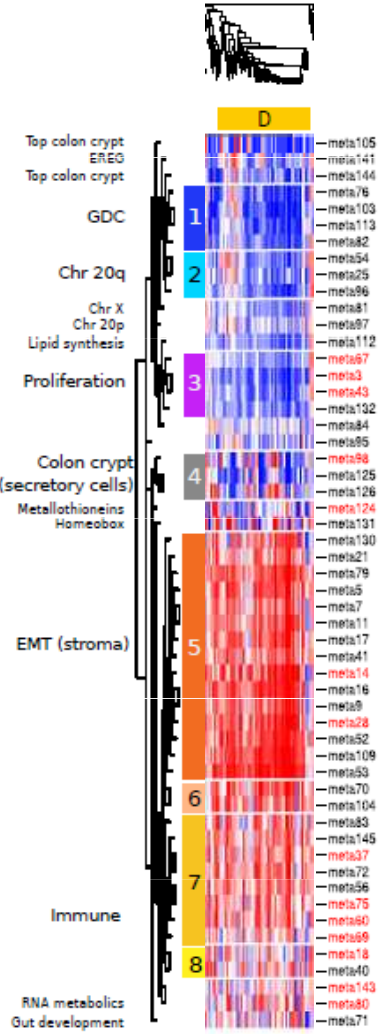
Podtyp D – Mesenchymal – markery kmeňových buniek, veľa mezenchymálnych buniek, ktoré sa prejavujú expresiou EMT génov. Wnt signálna dráha je neaktívna a proliferácia nízka. Klinické a mutačné charakteristiky sa nelíšia od populačnej baseline. Majú najkratšie prežitie do relapsu, zlý OS a SAR.



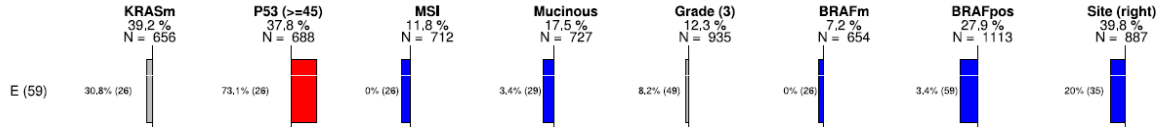


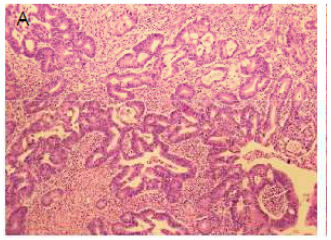


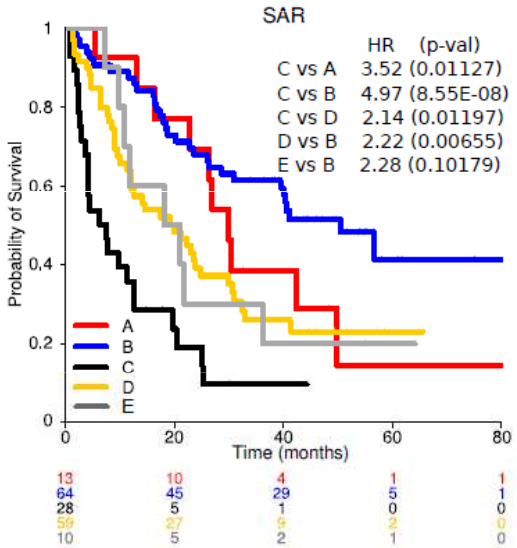


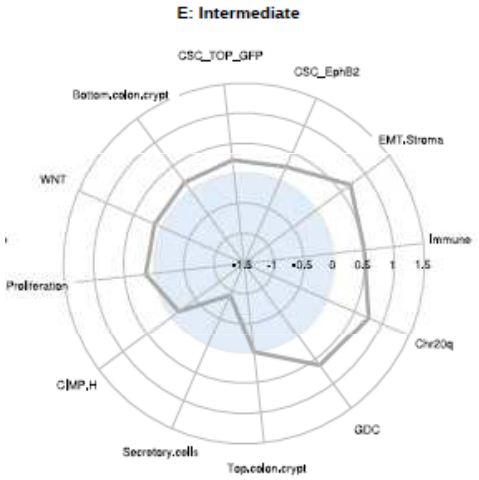


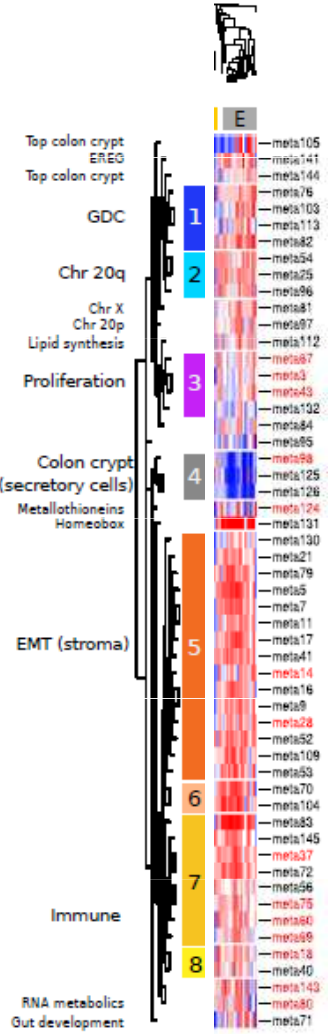
Podtyp E – Mixed – často MSS, *BRAF*wt, z ľavej strany hrubého čreva. Podobne ako podtyp D exprimuje gény kmeňových buniek a EMT procesu, avšak podobne s B má vysokú aktiviu kanonickej Wnt dráhy a vyzerá viac diferenciovaný. Je podobný B – komplexný tubulárny, len častejšie obsahuje mutáciu *p53*.







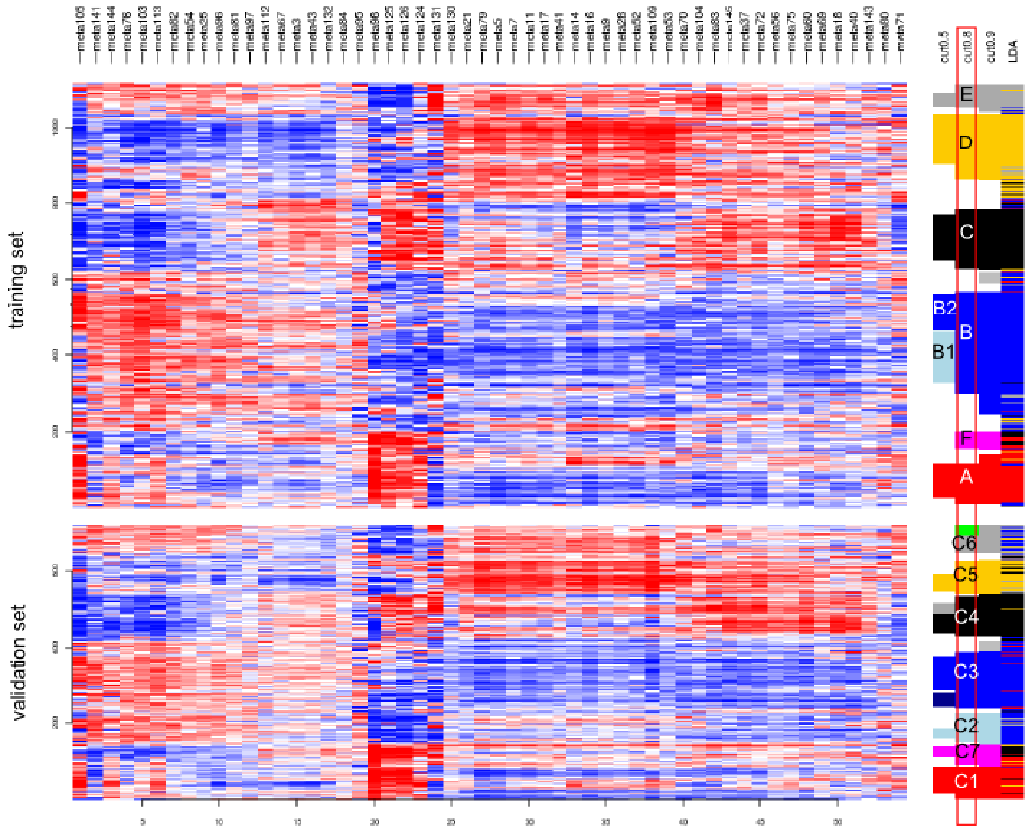




Validace podtypů kolorektálního karcinomu

Validácia algoritmu a parametrov modelu na testovacom súbore

Keď zopakujem celú procedúru na inom súbore, dostanem 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

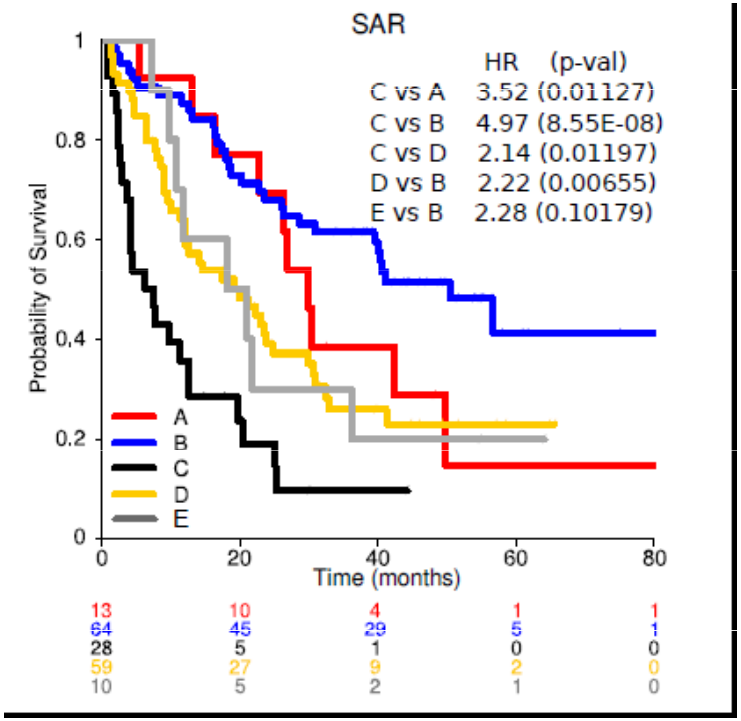
Validácia konceptu pomocou klinických, molekulárnych a histologických charakteristík objavených skupín

Majú objavené skupiny biologickú podstatu / odrážajú známe vedecké poznatky?

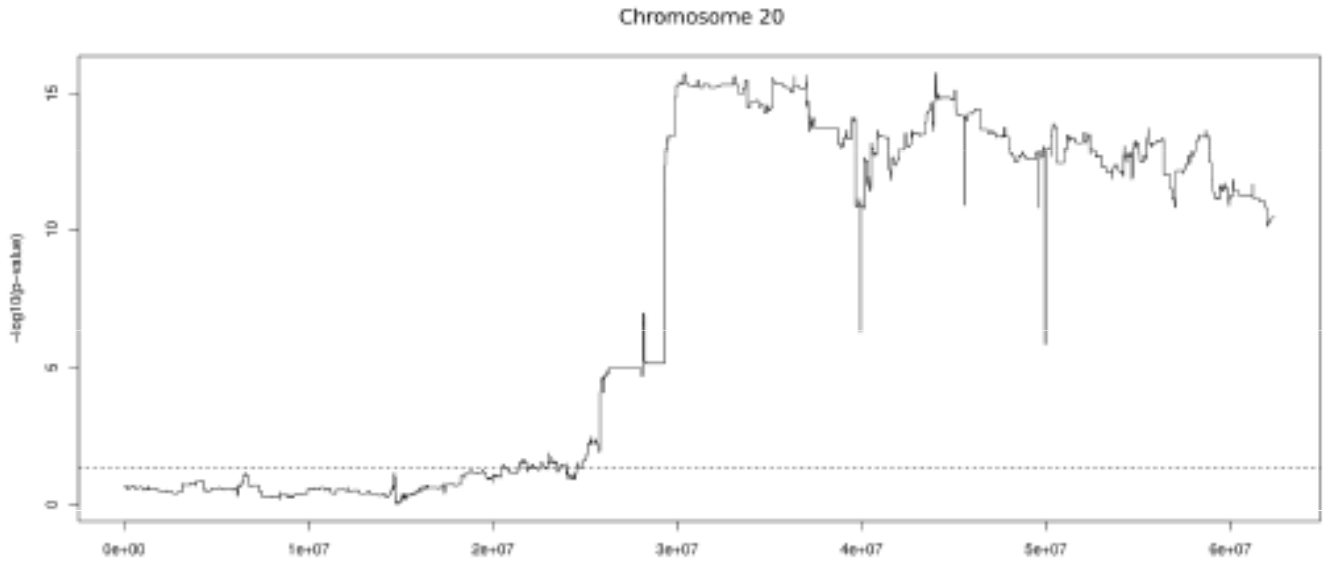
Je rozloženie týchto charakteristík medzi podtypmi porovnateľné vo validačnom súbore?

Majú objavené skupiny biologickú podstatu / odrážajú známe vedecké poznatky?

Podtyp C – pravostranné, BRAFm, MSI nádory, ktoré sú **známe zlým prežitím po relapse**

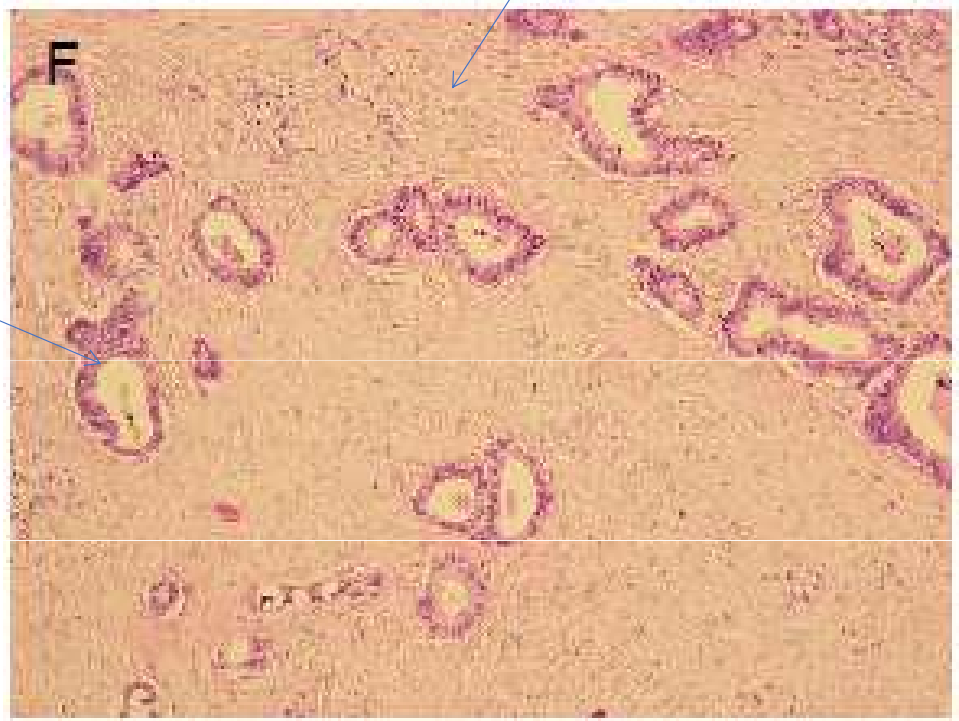


Zvýšená expresia génov chr. 20q v podtype B by mohla znamenať amplifikáciu chr20q regiónu.

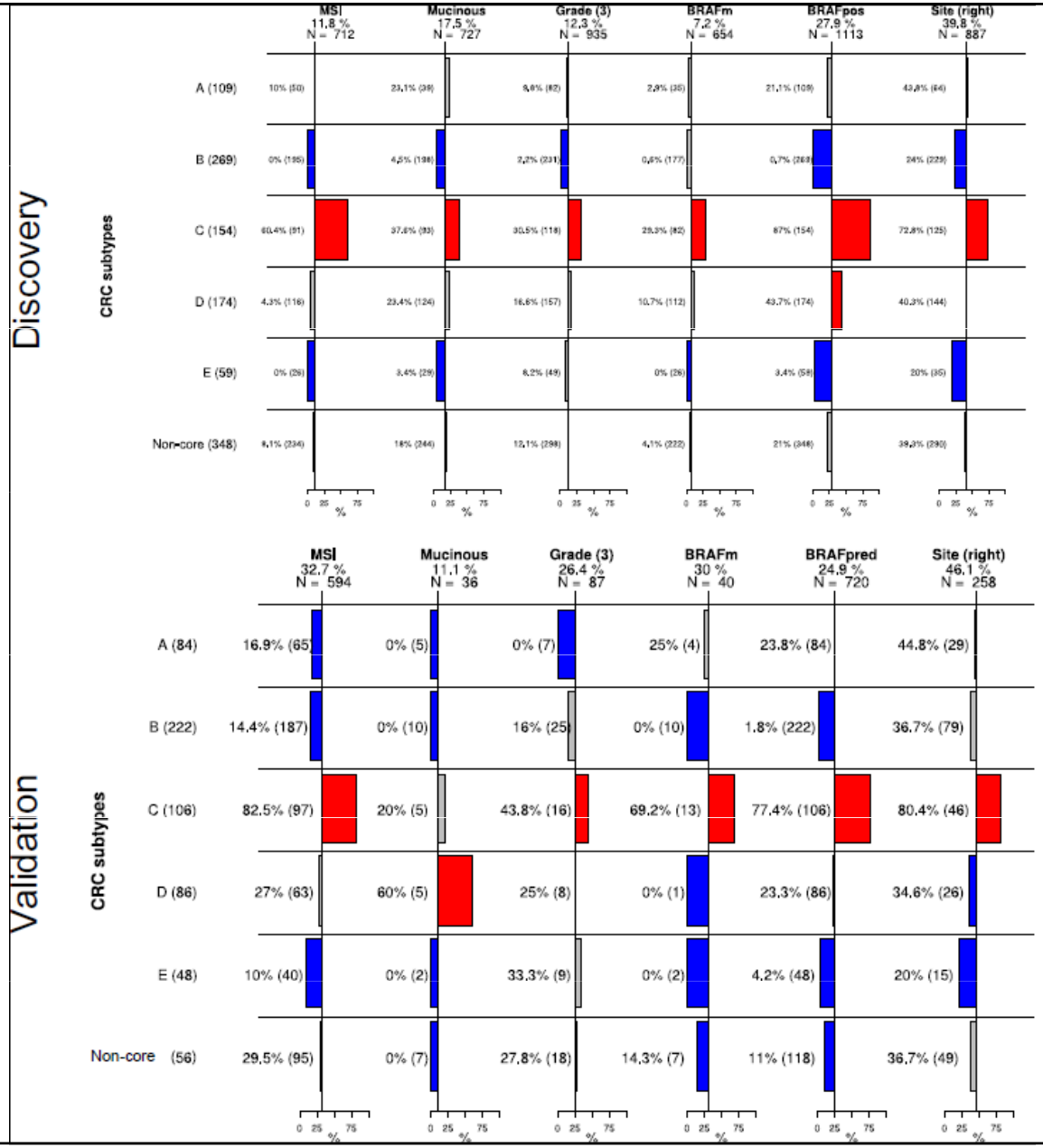


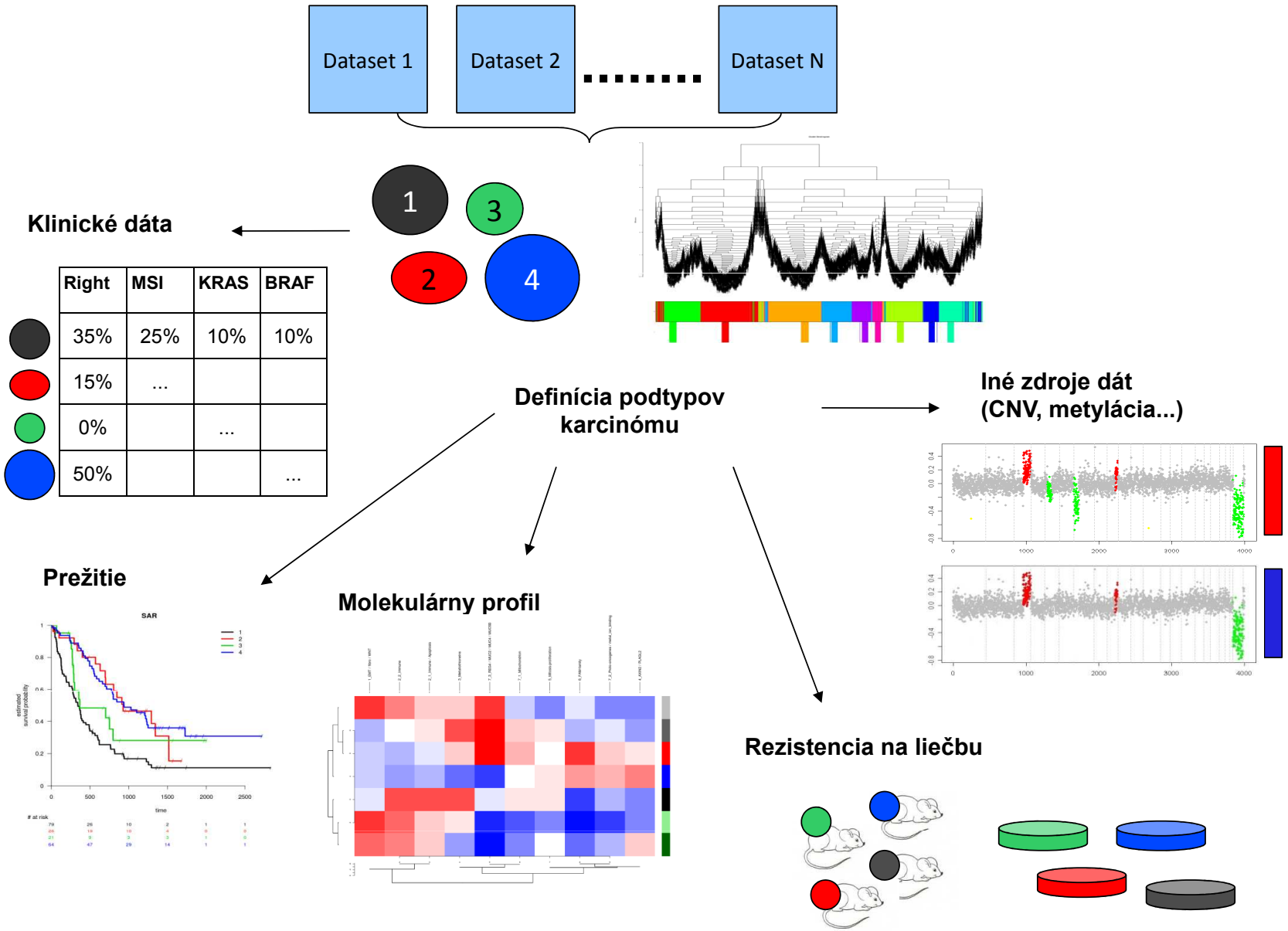
Podtyp D – mezenchymálny – histologické vyhodnotenie: v nádore prítomná silná desmoplastická reakcia (mezenchymálne tkanivo)

Nádor



Je rozloženie klinických charakteristík medzi podtypmi porovnateľné vo validačnom súbore?


Discovery
CRC subtypes
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
A (109)
B (269)
C (154)
D (174)
E (59)
Non-core (348)
0 25 75
%
Validation
CRC subtypes
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
A (84)
16.9% (65)
0% (5)
0% (7)
25% (4)
23.8% (84)
44.8% (29)
B (222)
14.4% (187)
0% (10)
16% (25)
0% (10)
1.8% (222)
36.7% (79)
C (106)
82.5% (97)
20% (5)
43.8% (16)
69.2% (13)
77.4% (106)
80.4% (46)
D (86)
27% (63)
60% (5)
25% (8)
0% (1)
23.3% (86)
34.6% (26)
E (48)
10% (40)
0% (2)
33.3% (9)
0% (2)
4.2% (48)
20% (15)
Non-core (56)
29.5% (95)
0% (7)
27.8% (18)
14.3% (7)
11% (118)
36.7% (49)
0 25 75
%


Dataset 1
Dataset 2
Dataset N
1
2
3
4
Klinické dáta
Right
MSI
KRAS
BRAF
35%
25%
10%
10%
15%
...
0%
...
50%
...
Definícia podtypov karcinómu
Iné zdroje dát (CNV, metylácia...)
Prežitie
SAR
Molekulárny profil
Rezistencia na liečbu