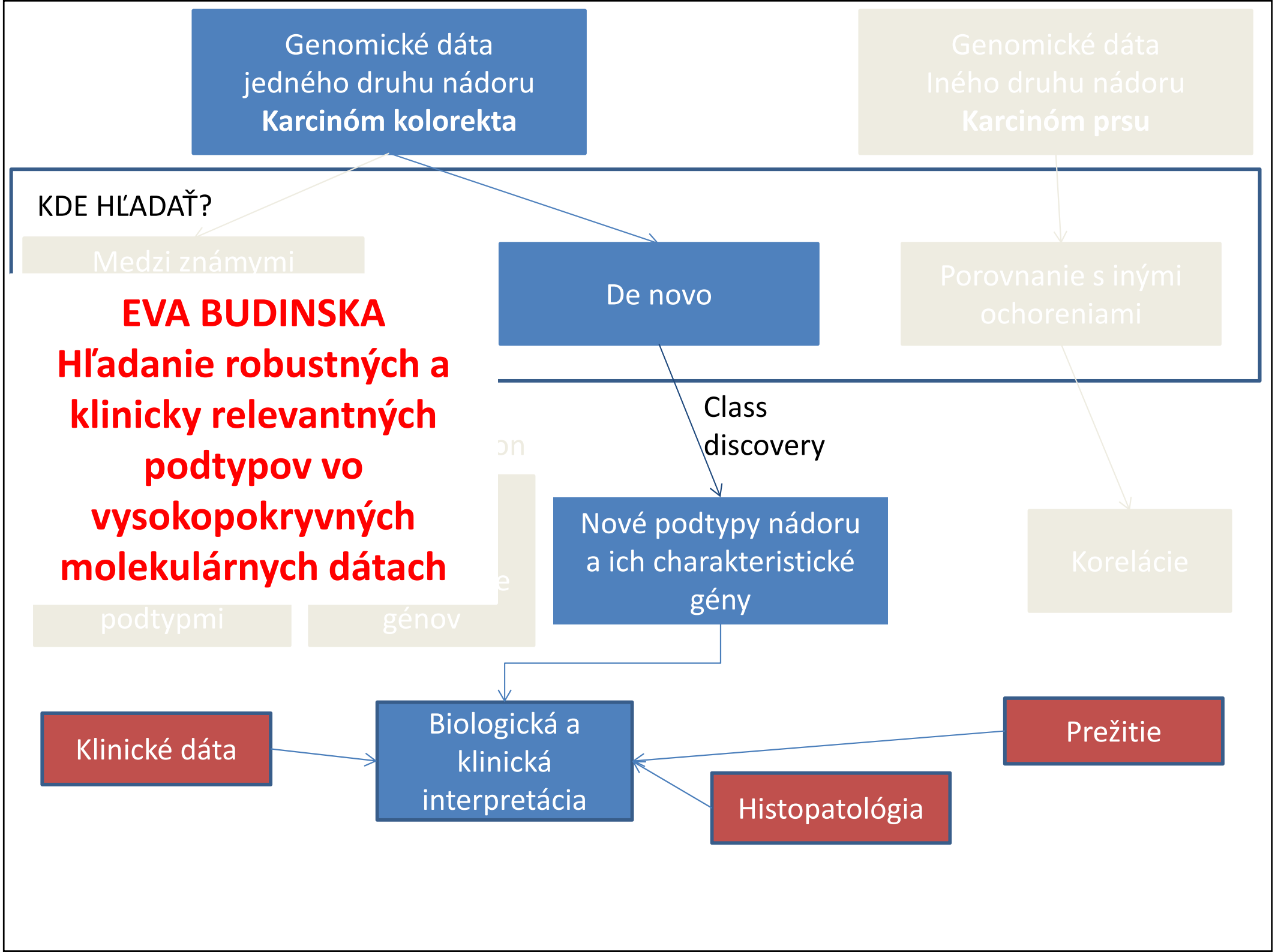
Hľadanie robustných a klinicky relevantných podtypov vo vysokopokryvných molekulárnych dátach

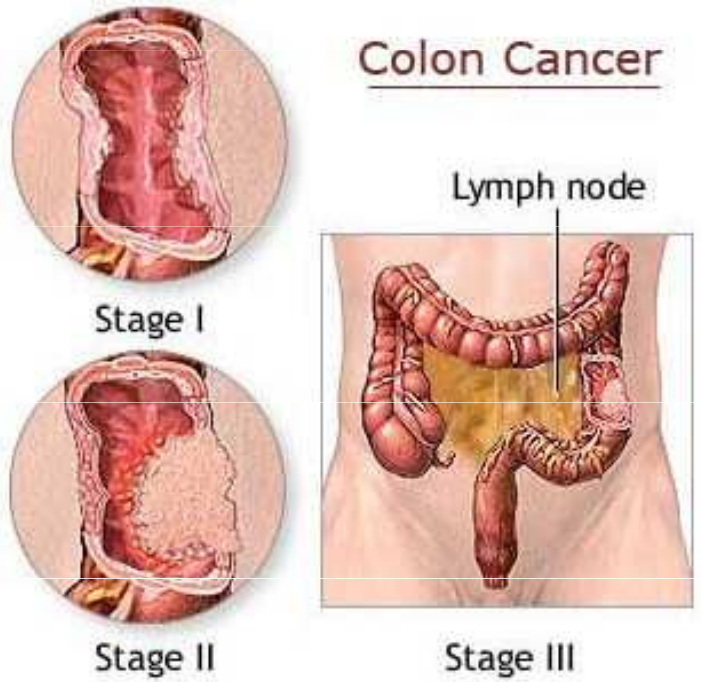
príkladová štúdia: kolorektálny karcinóm

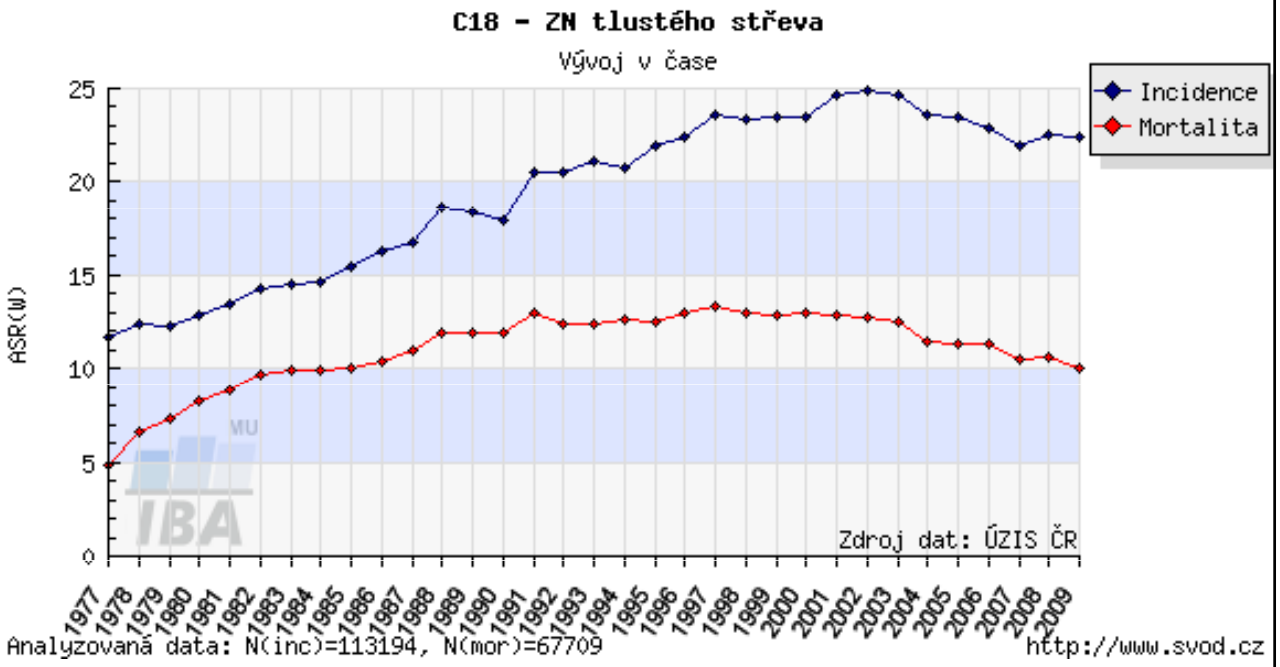
Eva Budinská

Letná škola Matematickej Biológie
Mikulov, 14.9. 2012


Genomické dáta
jedného druhu nádoru
Karcinóm kolorekta
Genomické dáta
Iného druhu nádoru
Karcinóm prsu
KDE HĽADAŤ?
Medzi známymi
De novo
Porovnanie s inými
ochoreniami
EVA BUDINSKA
Hľadanie robustných a
klinicky relevantných
podtypov vo
vysokopokryvných
molekulárnych dátach
Class
discovery
Nové podtypy nádoru
a ich charakteristické
gény
Korelácie
podtypmi
génov
Klinické dáta
Biologická a
klinická
interpretácia
Prežitie
Histopatológia

KOLOREKTÁLNY KARCINÓM





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*

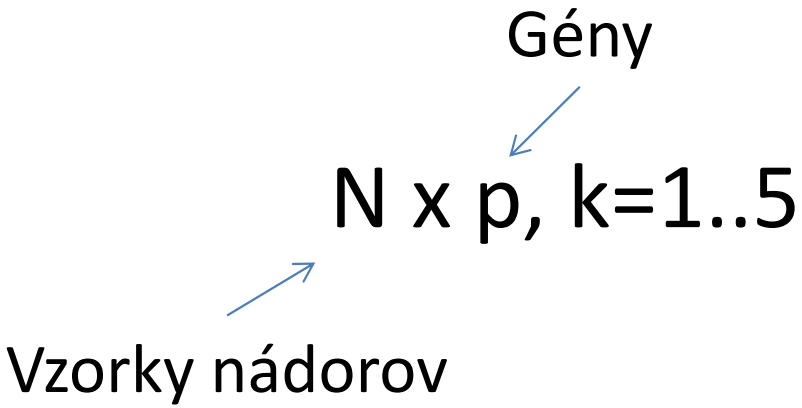
Cieľ:

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



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

Podtypy:

Neznáma pravda => Snaha o čo najobjektívnejšiu a najvšeobecnejšiu identifikáciu skutočnosti

Prístup:

Zhlukovanie – class discovery

Obmedzenia:

- (Ne) reprezentatívnosť populácie v dátach
- Technické možnosti
- Typ dát určuje uhol pohľadu – génová expresia, epidemiológia, histológia, DNA mutácie...)
- Odchýlky v dátach môžu podstatne ovplyvniť výsledok

Nevýhody zhlukovania:

- Rozdelenie dát do zhlukov i v prípade neexistencie skutočnej vnútornej štruktúry dát (štruktúra je náhodná)
- Potrebné metódy ad hoc stanovenia správneho počtu zhlukov
- Šum v dátach negatívne ovplyvňuje výsledok
- **Jedna veľká množina génov z biologického motívu ovplyvní zásadne zhluky**

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

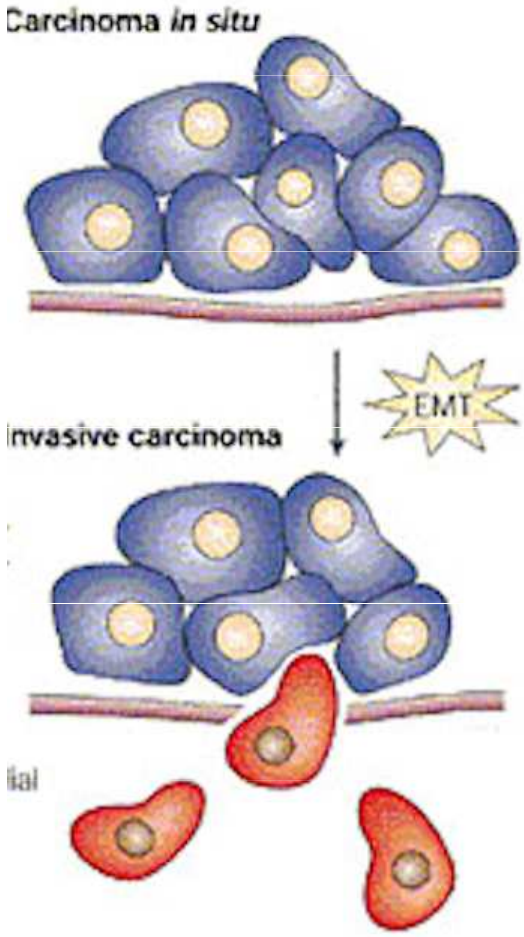
Jedná sa o istú formu váženia efektu biologických motívov

Predpoklad:

sada korelovaných génov ~ biologický motív

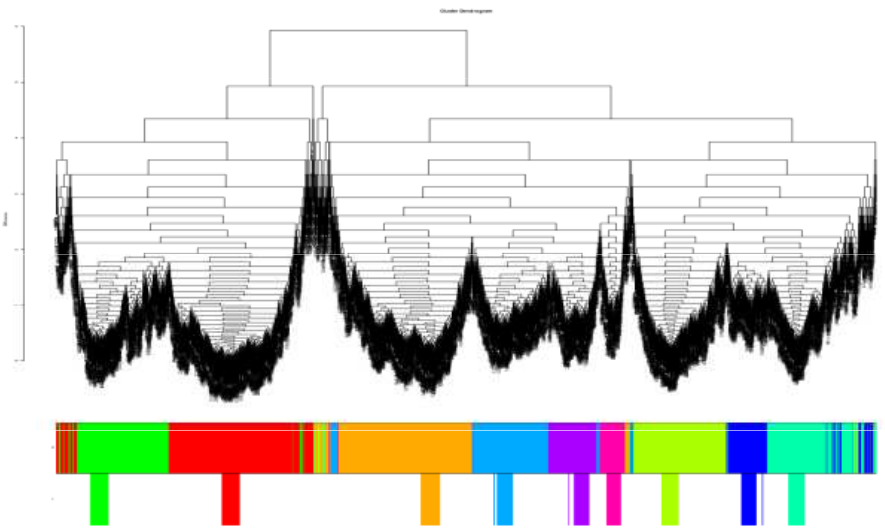
Príklad EMT

- 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

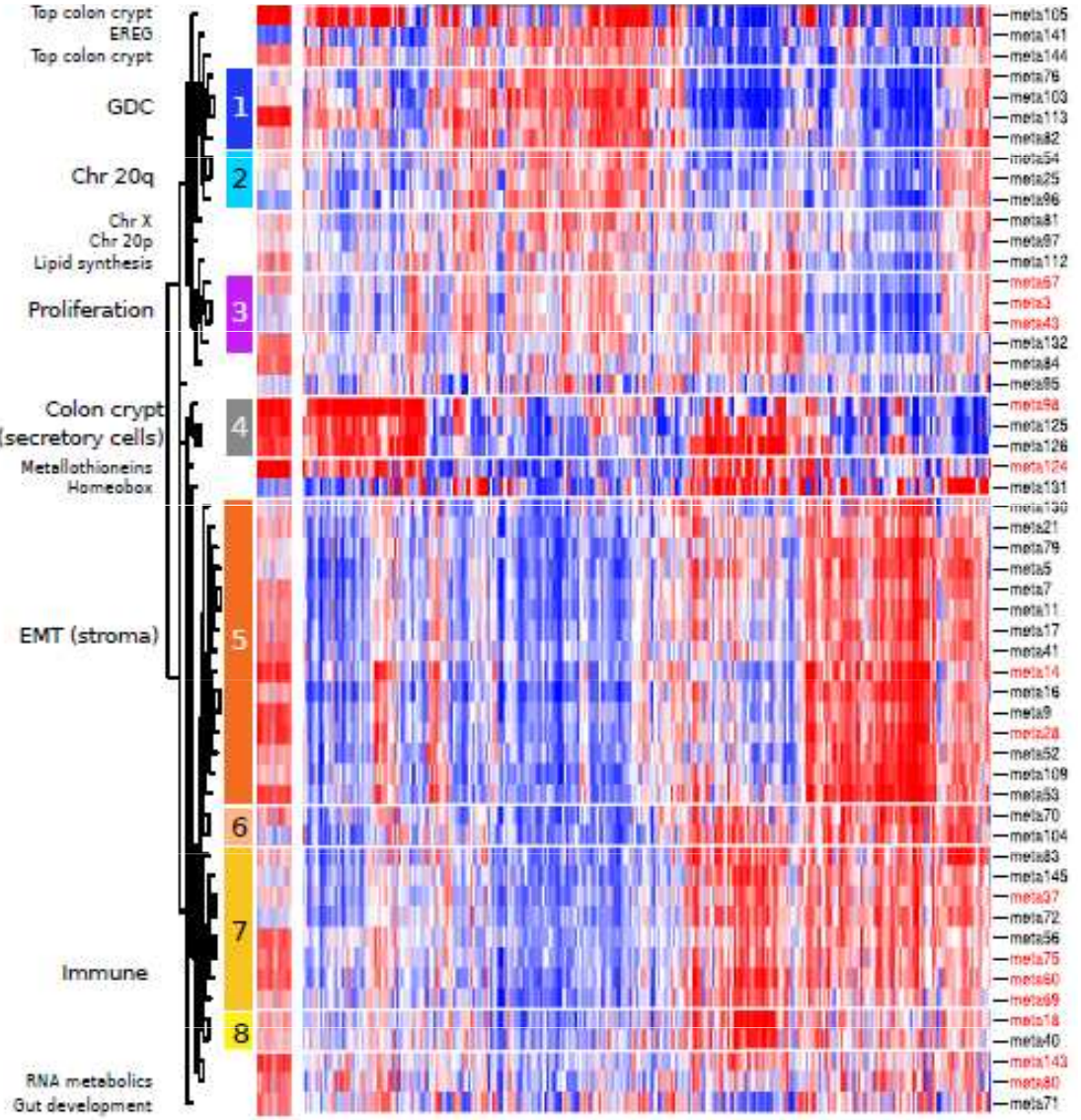
- Analýza génových sád

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

Biol motív:

EMT, proliferácia,

Chromozóm 20q...

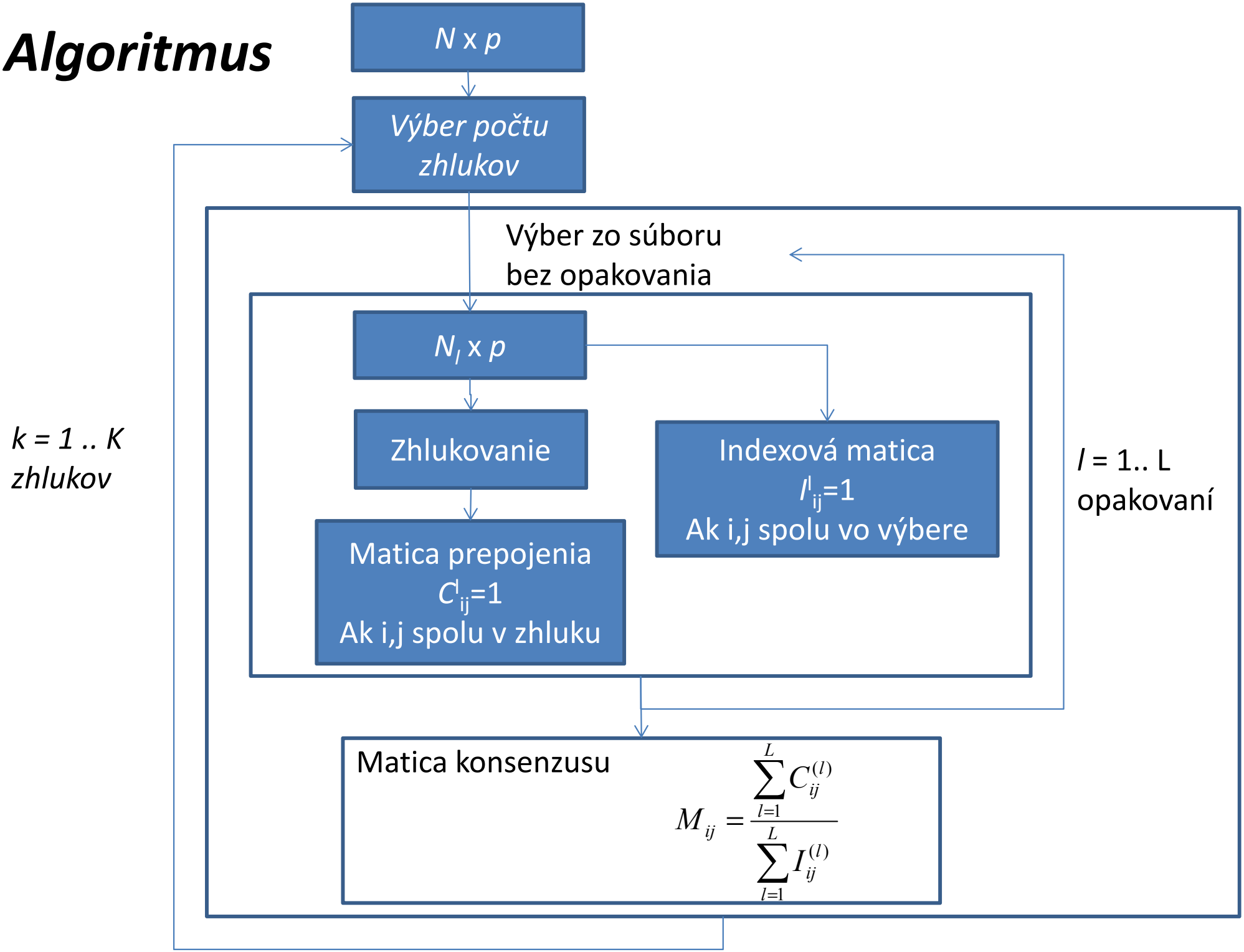


ROBUSTNOSŤ V ZHLUKOVANÍ

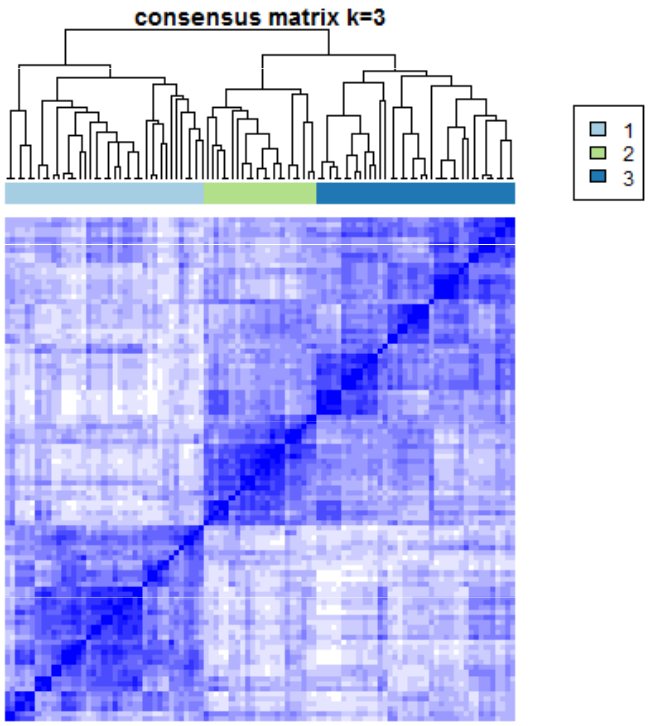
Consensus clustering

Metóda založená na opakovanom prevzorkovaní (<90% pôvodnej populácie a/alebo premenných) a zhlukovaní, za účelom vytvorenia konsenzusu medzi viacerými opakovaniami zhlukovania.

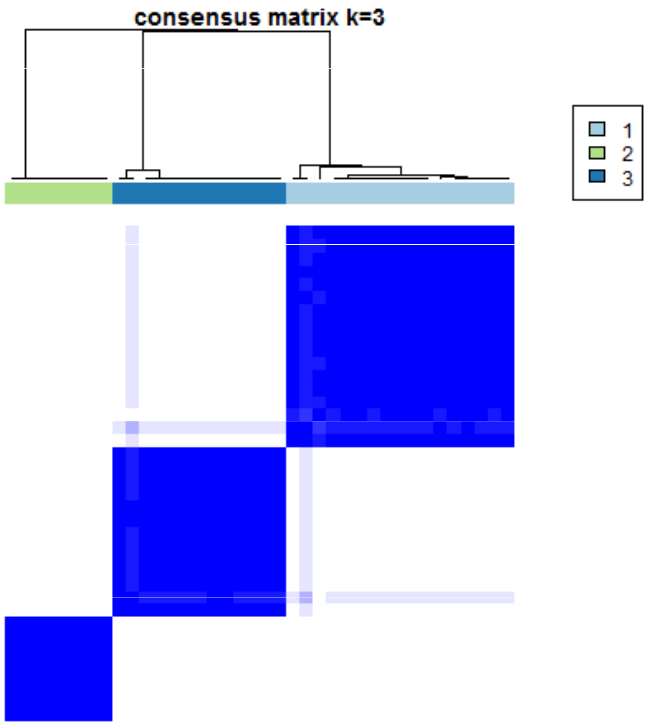
Algoritmus


N x p
Výber počtu zhlukov
Výber zo súboru bez opakovania
N_l x p
Zhlukovanie
Matica prepojenia
C^l_ij=1
Ak i,j spolu v zhluku
Indexová matica
I^l_ij=1
Ak i,j spolu vo výbere
Matica konsenzusu
M_ij = Σ_{l=1}^{L} C_ij^(l) / Σ_{l=1}^{L} I_ij^(l)
k = 1 .. K zhlukov
l = 1.. L opakovaní

- Ukazuje zhlukovanie stabilné aj pri drobnom narušení štruktúry dát – zhluky sú robustné
- Konsenzus – je nová metrika podobnosti (0,1)
- Pomáha určiť : počet a stabilitu zhlukov



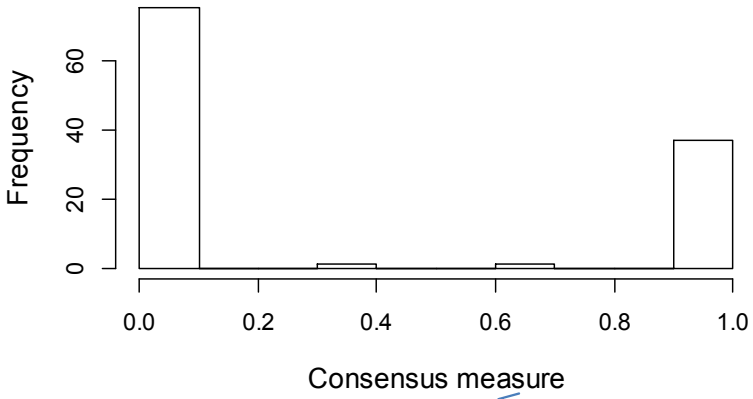
Náhodný výber z normálneho rozloženia

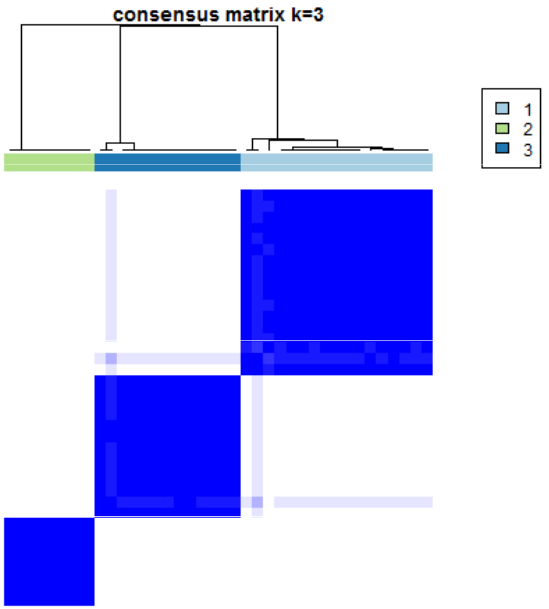


Leukémia:
ALL-T, ALL-B, AML

Odhad počtu zhlukov I

Histogram miery konzenzusu



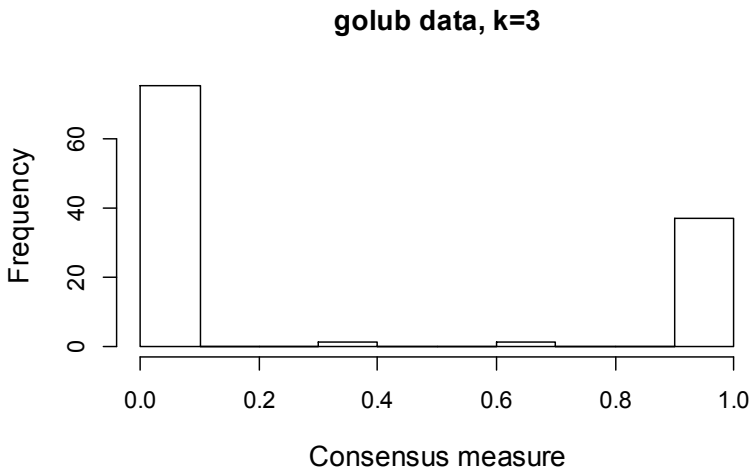


Konzenzus medzi dvomi vzorkami

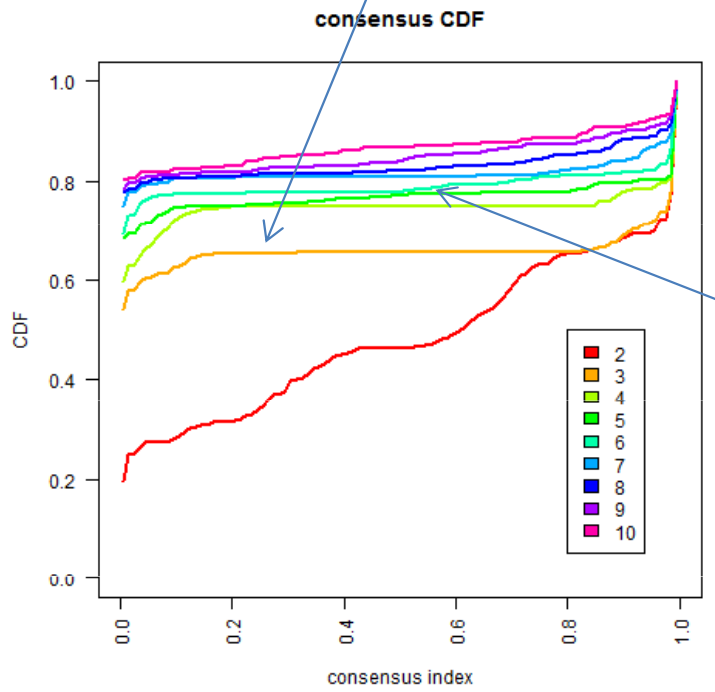
$$CDF^{x} = \frac{\sum_{i<j} 1\{M_{ij} \leq x\}}{N(N-1)/2}$$

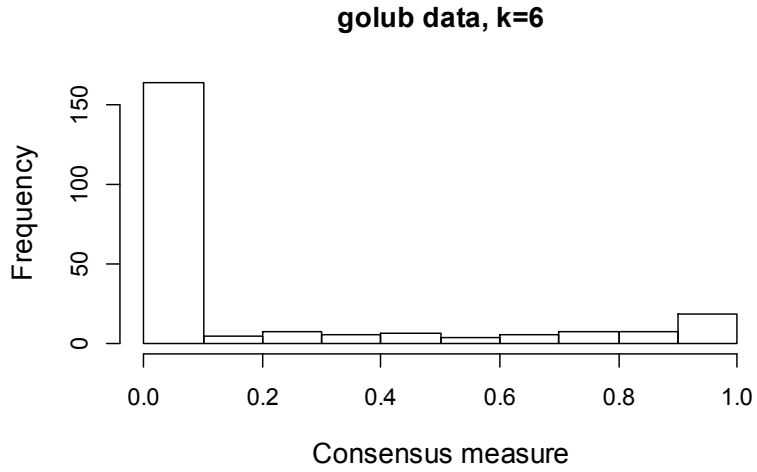
Kumulatívna distribučná funkcia

Odhad počtu zhlukov II



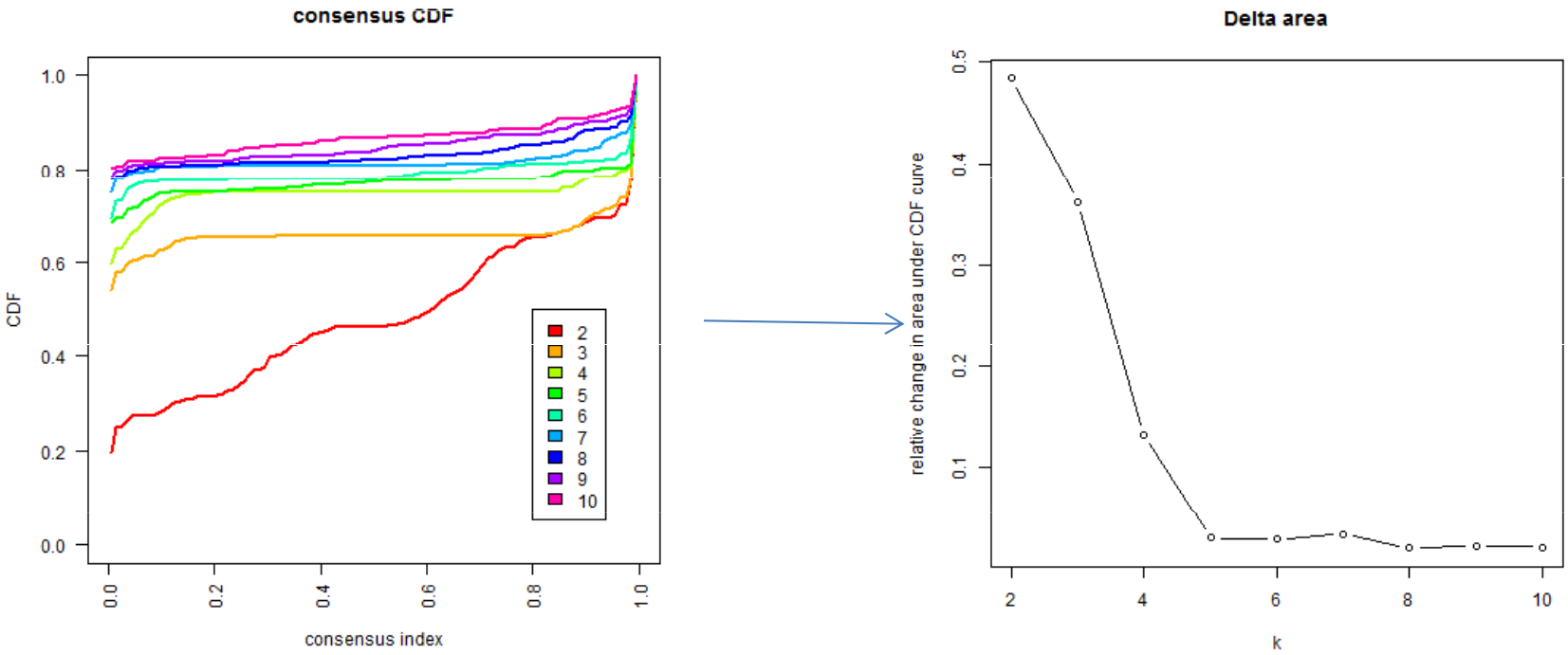
$$CDF^x = \frac{\sum_{i<j} 1\{M_{ij} \le x\}}{N(N-1)/2}$$



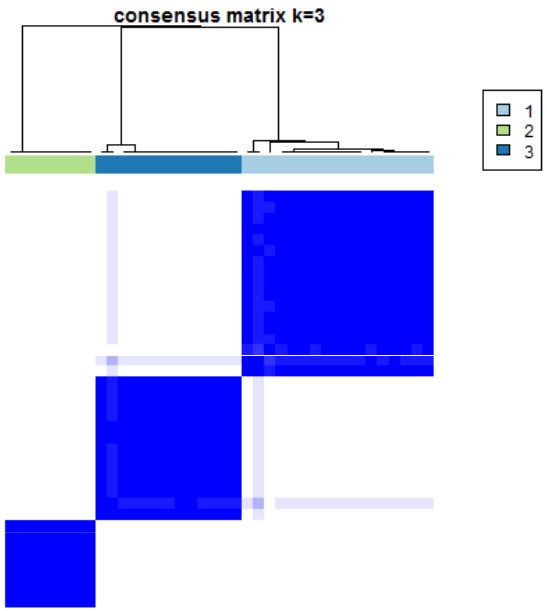


Odhad počtu zhlukov III

Delta = relatívna zmena plochy pod CDF krivkou medzi dvoma k



Odhad stability zhlukov



Kumulatívna distribučná funkcia

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

Consensus clustering v hierarchickom zhlukovaní - nevýhoda

Statické rezanie dendrogramu

Všetky vzorky sú zaradené do zhluku

Dynamic hybrid cut

Poskytuje dynamické rezanie dendrogramu na základe minimálnej veľkosti zhlukov, maximálnej výšky rezu a ďalších parametrov.

Ak vzorka nespĺňa kritériá, nie je zaradená do zhluku.

Dynamic hybrid cut - základy

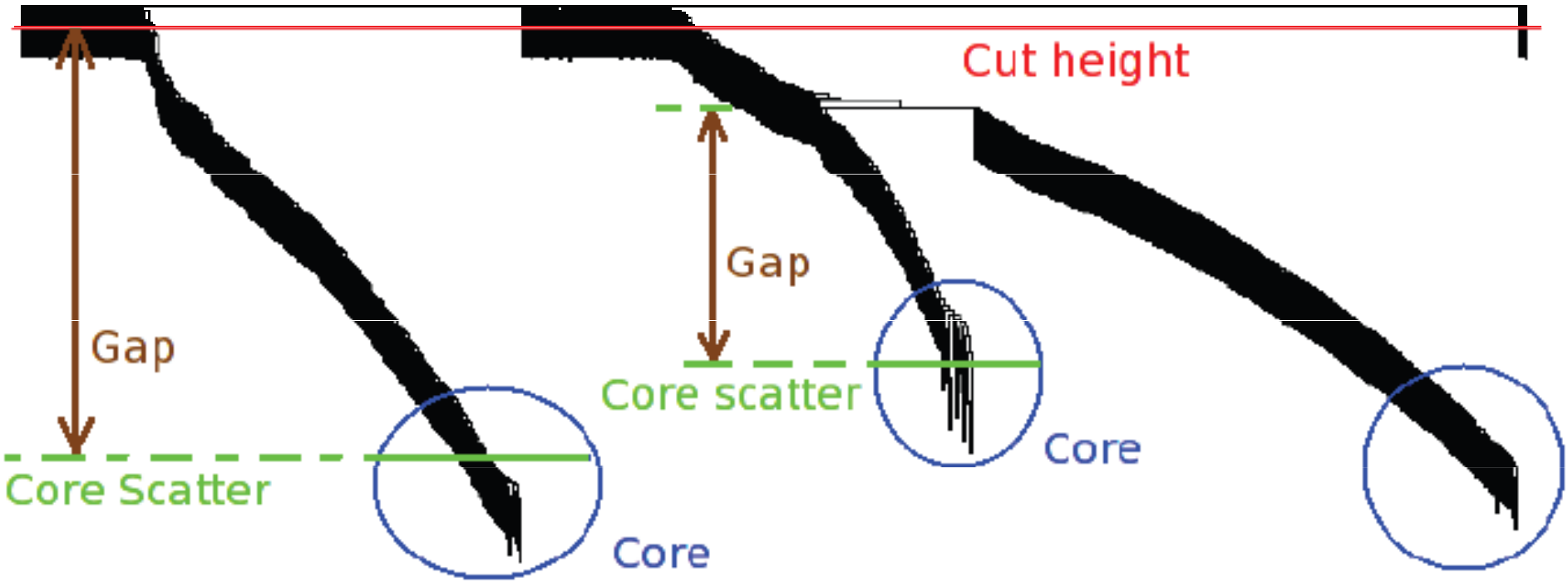
(1) Zhluk musí mať minimálny počet členov

(2) Príliš vzdialené objekty sú zo zhluku vylúčené aj v prípade, že patria do rovnakého ramena dendrogramu

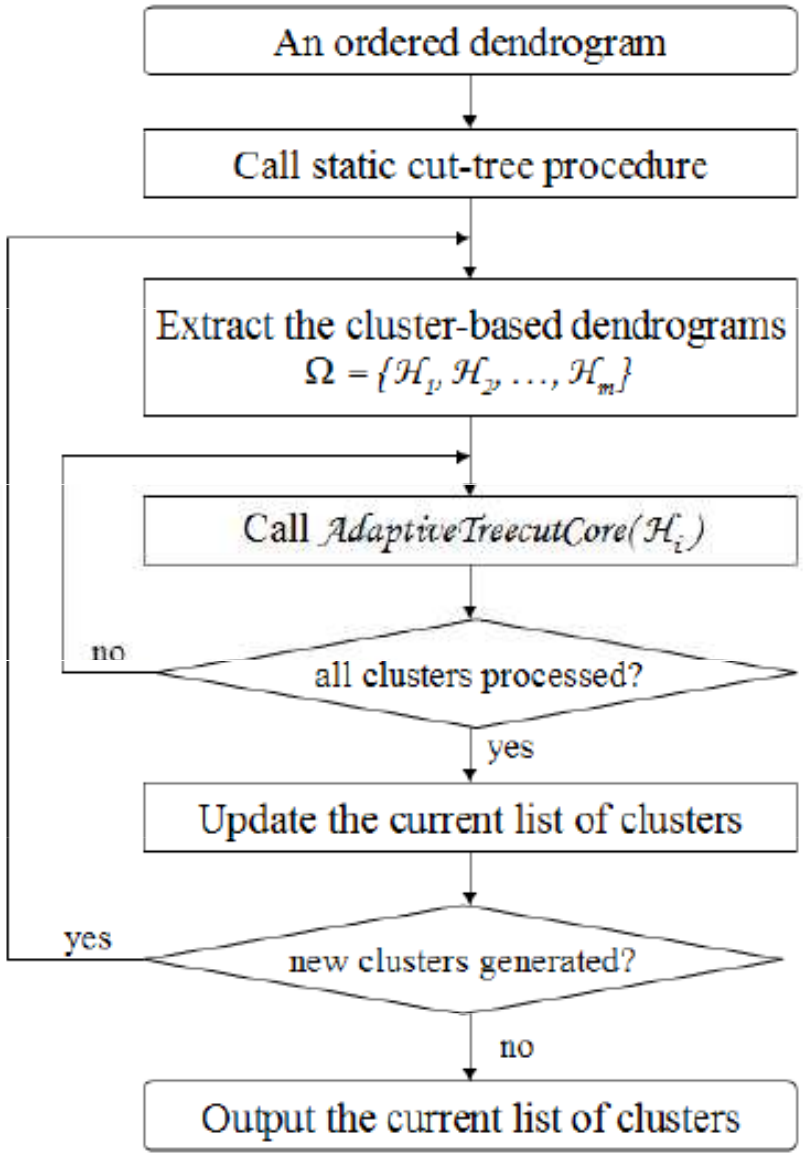
(3) Každý zhluk by mal byť separovaný od okolia medzerou

(4) Jadro (najnižšie prepojené objekty) každého zhluku musí byť silne prepojené

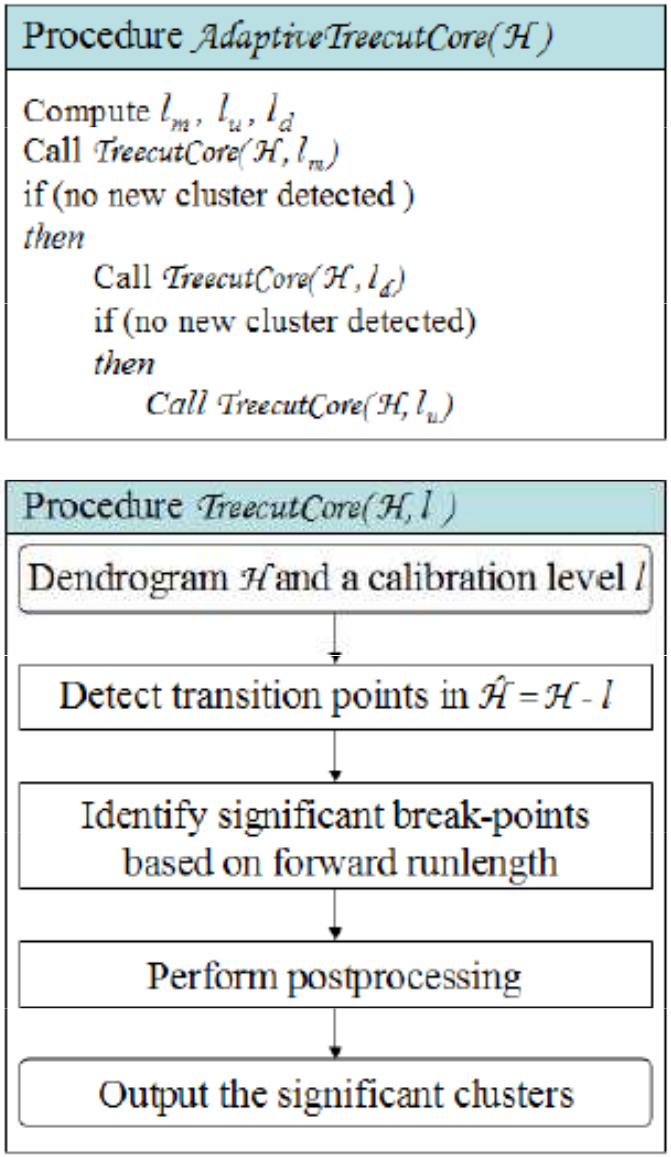
Parametre dynamic hybrid cut



Algoritmus dynamic hybrid cut

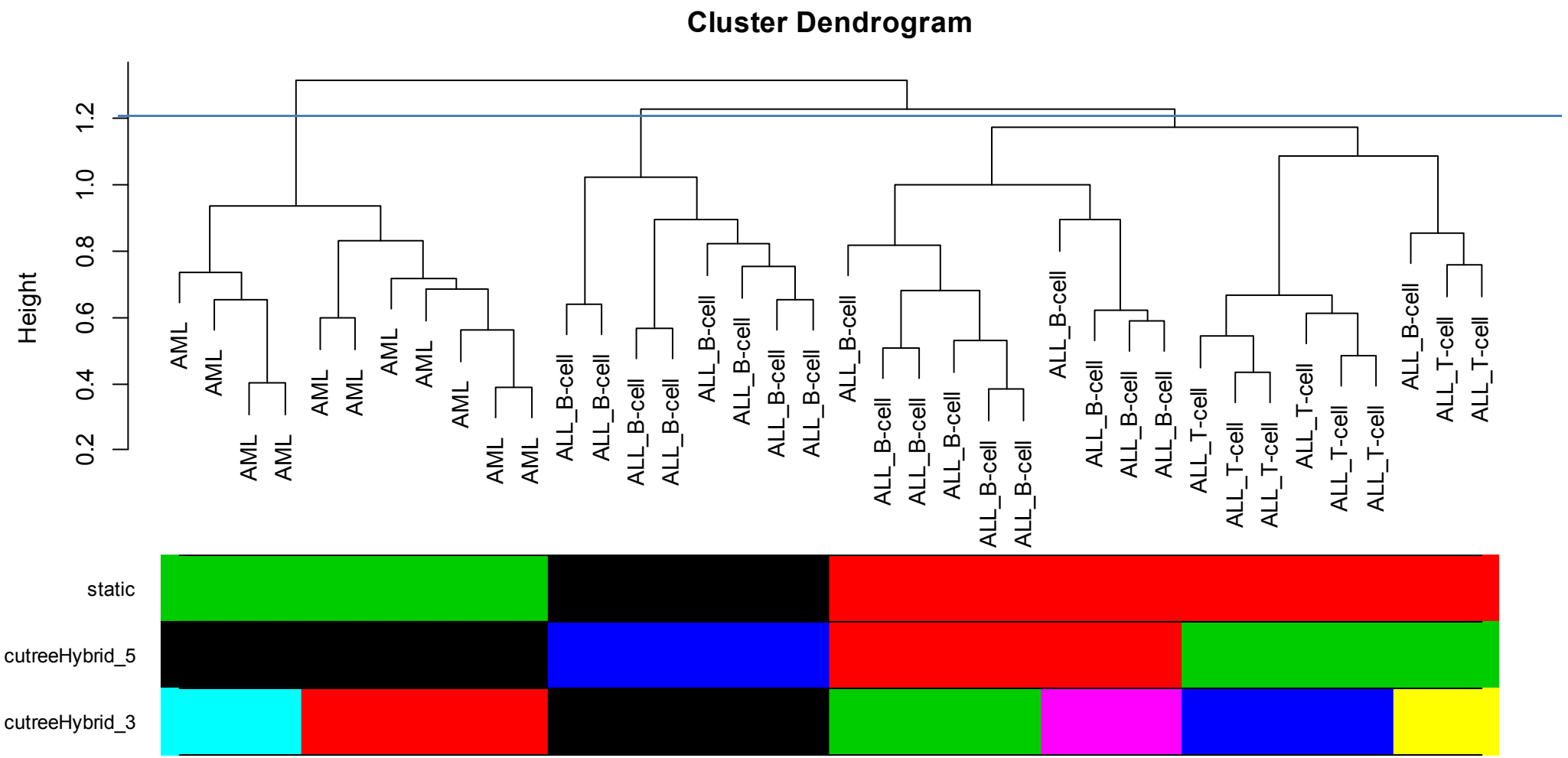


(a)



(b)

Porovnanie statického rezu s dynamickým rezom



Cieľ:

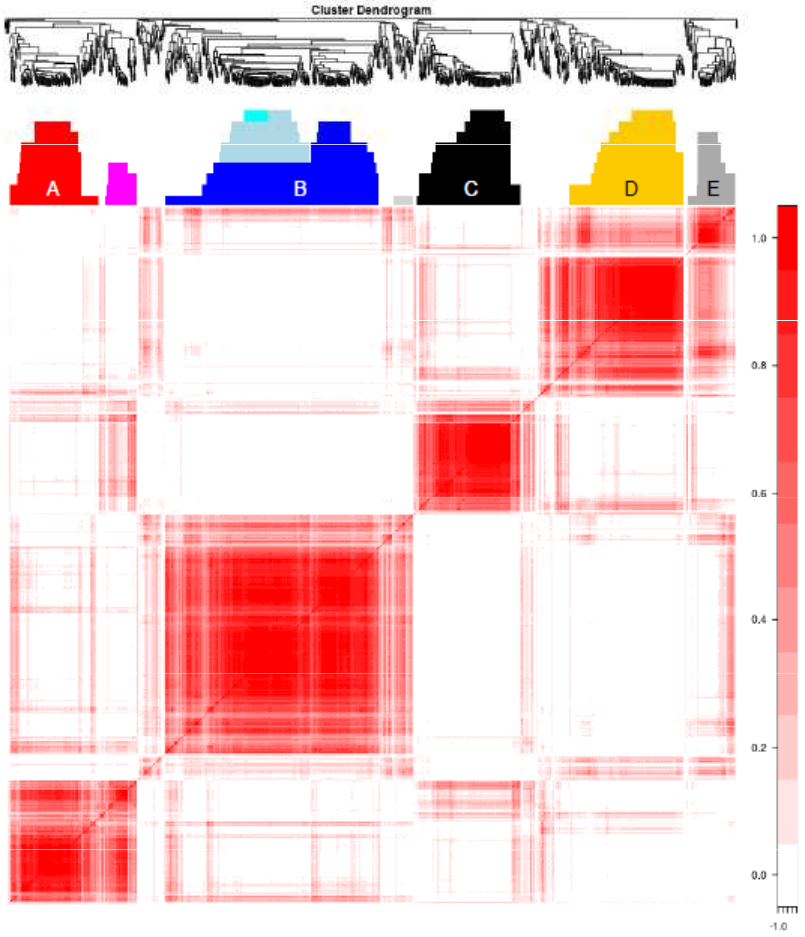
Nájsť skupiny nádorov kolorekta s podobnou expresiou génov (podobným génovým profilom) ~ podtypy

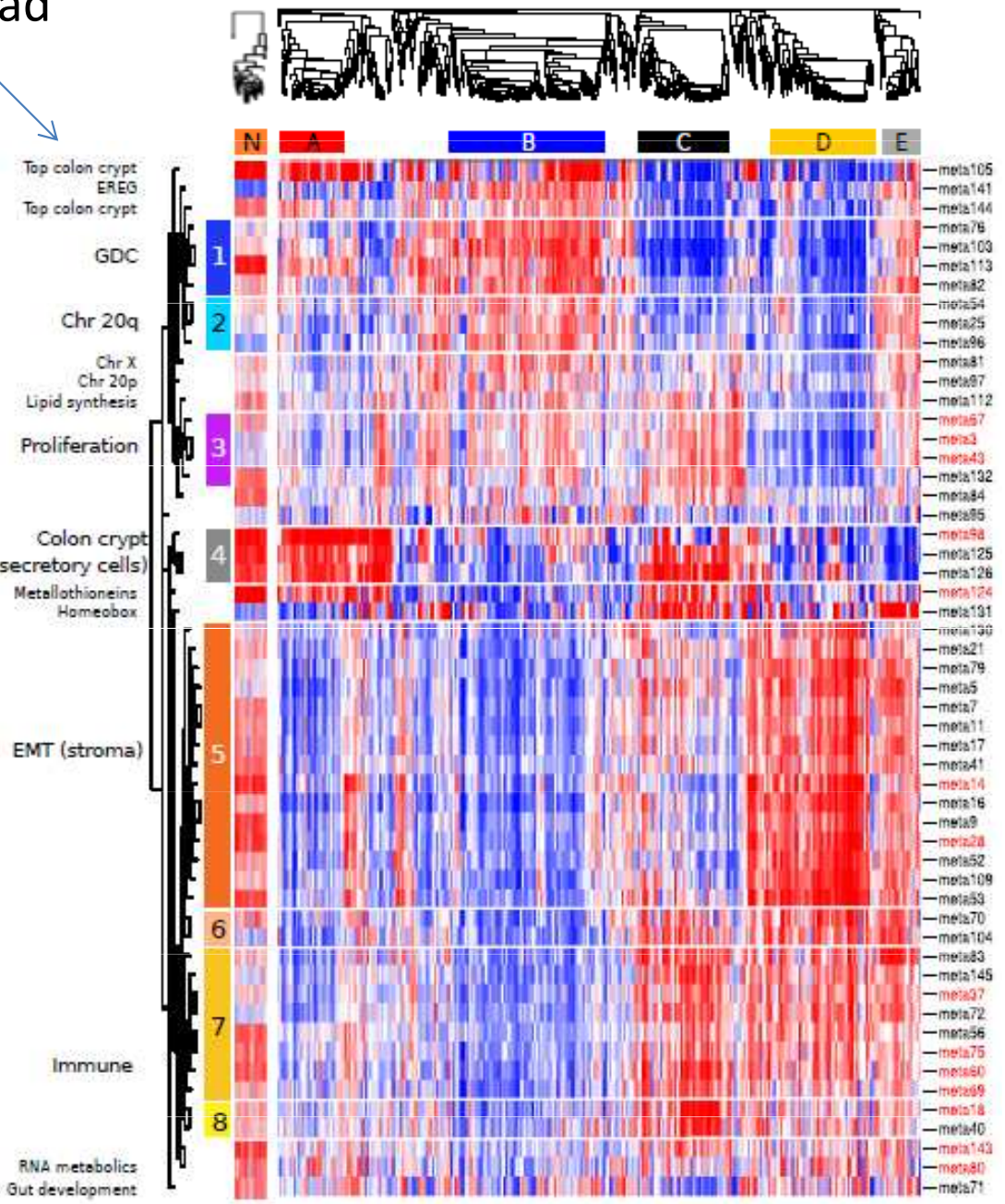
Charakterizovať tieto podtypy pomocou klinických a známych molekulárnych parametrov.

Motívy génovej expresie v podtypoch

Analýza génových sád

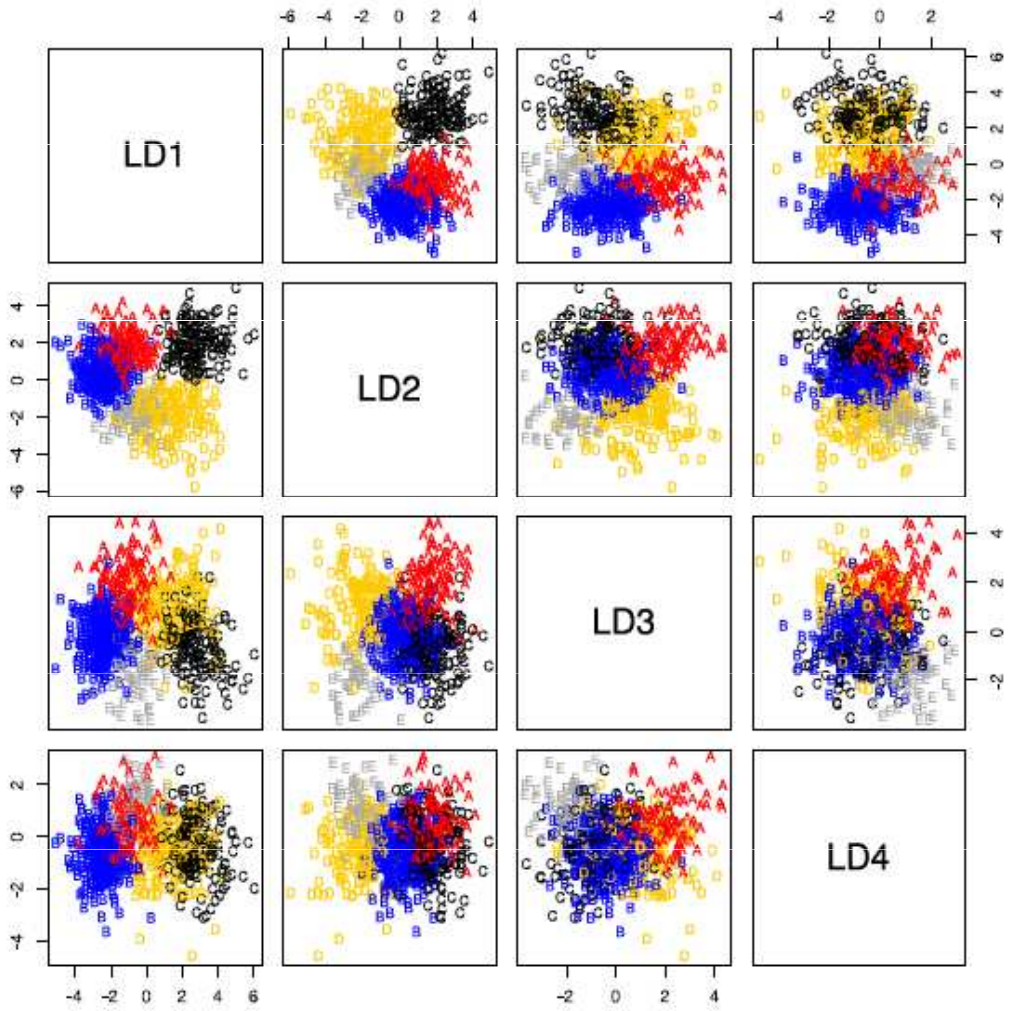
Hierarchické zhlukovanie na matici konsenzusu





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)

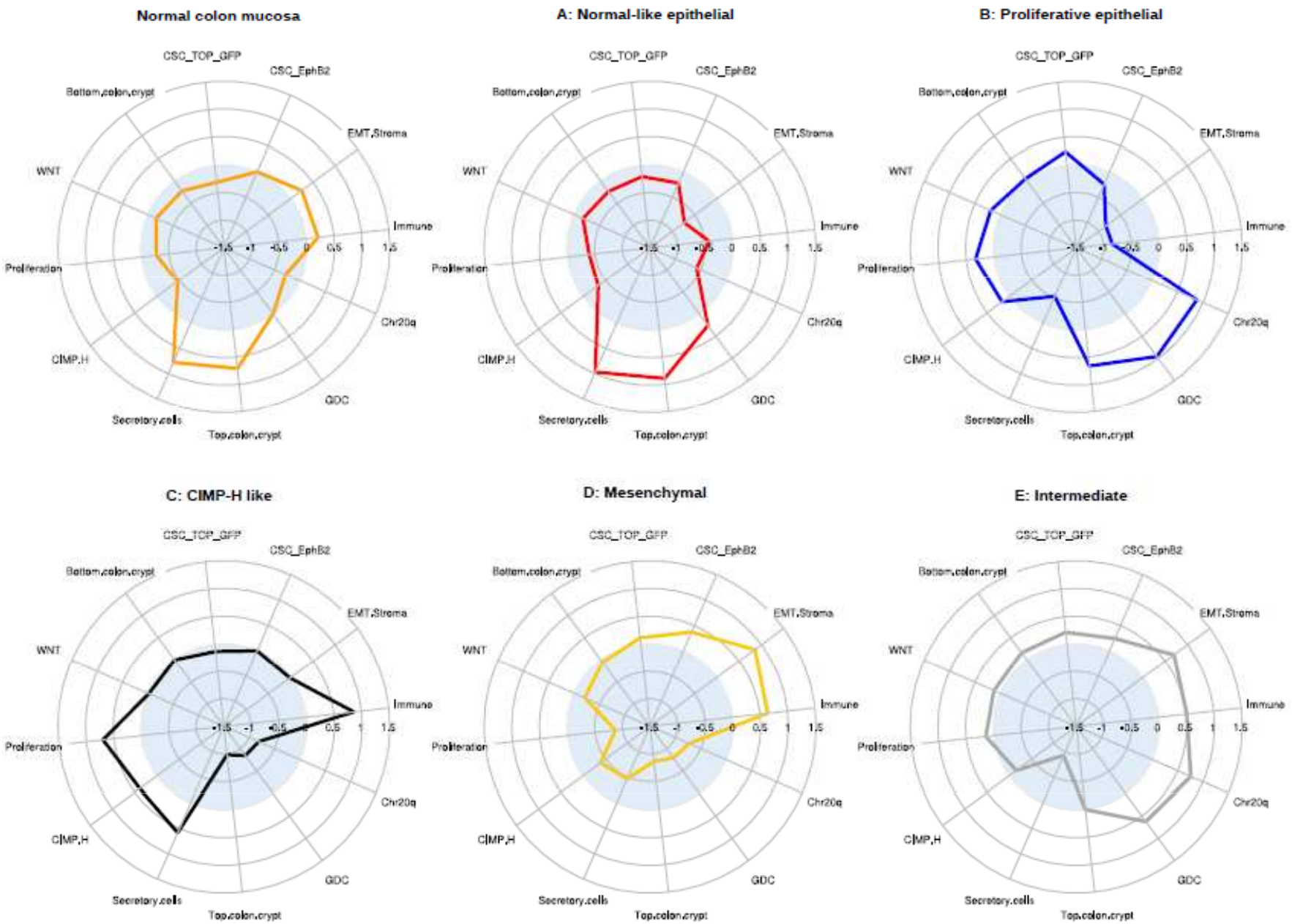
Subtype	Minimálna génová sada
A	CLCA1, PADI2, ADTRP, RETNLB, TIMP3, MUC2, FNDC1, NR3C2, SULF1, B3GNT7, STYK1, CHI3L1
B	FARP1, ALOX5, FSCN1, HNF4A, RARRES3, MYRIP, GPSM2, TSPAN6, CCDC113, CDHR1, KCTD12, SGK1, BASP1, MT1E, GPX8, RPS6KA3, SOCS3, SLC5A6, PRR15, PLAGL2, IHH, CREB3L1, TP53RK, YAE1D1, EPB41L3,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

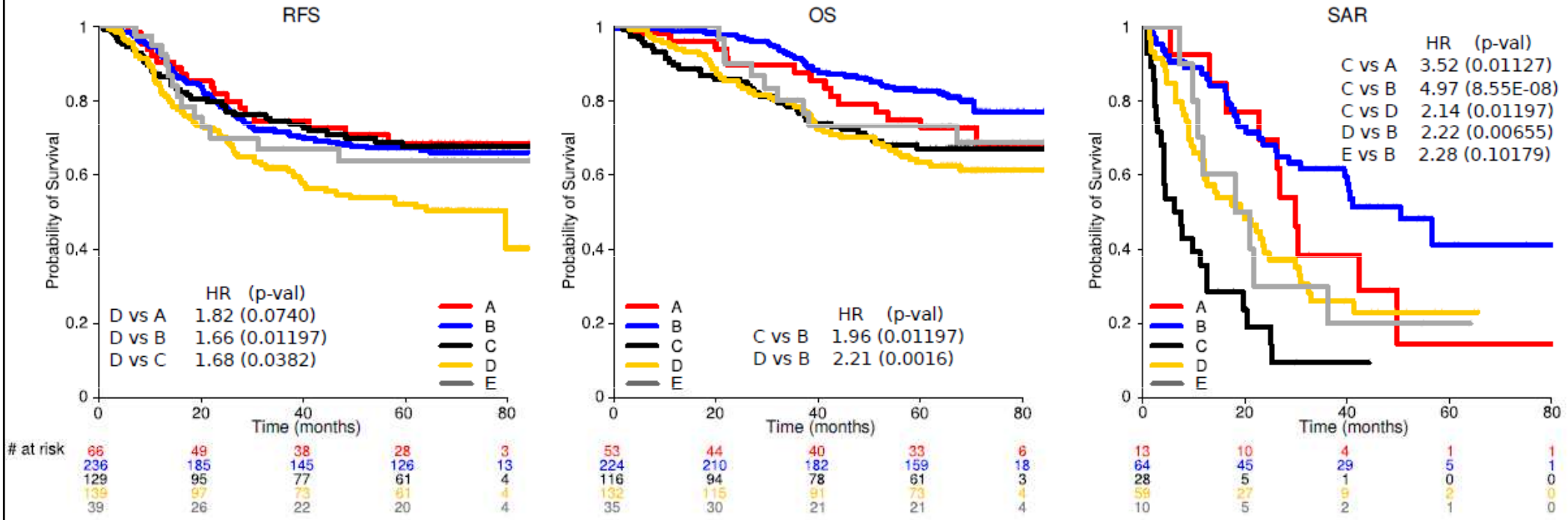
Charakterizovať tieto podtypy pomocou klinických a známych molekulárnych parametrov.

Vzor expresie biologických motívov



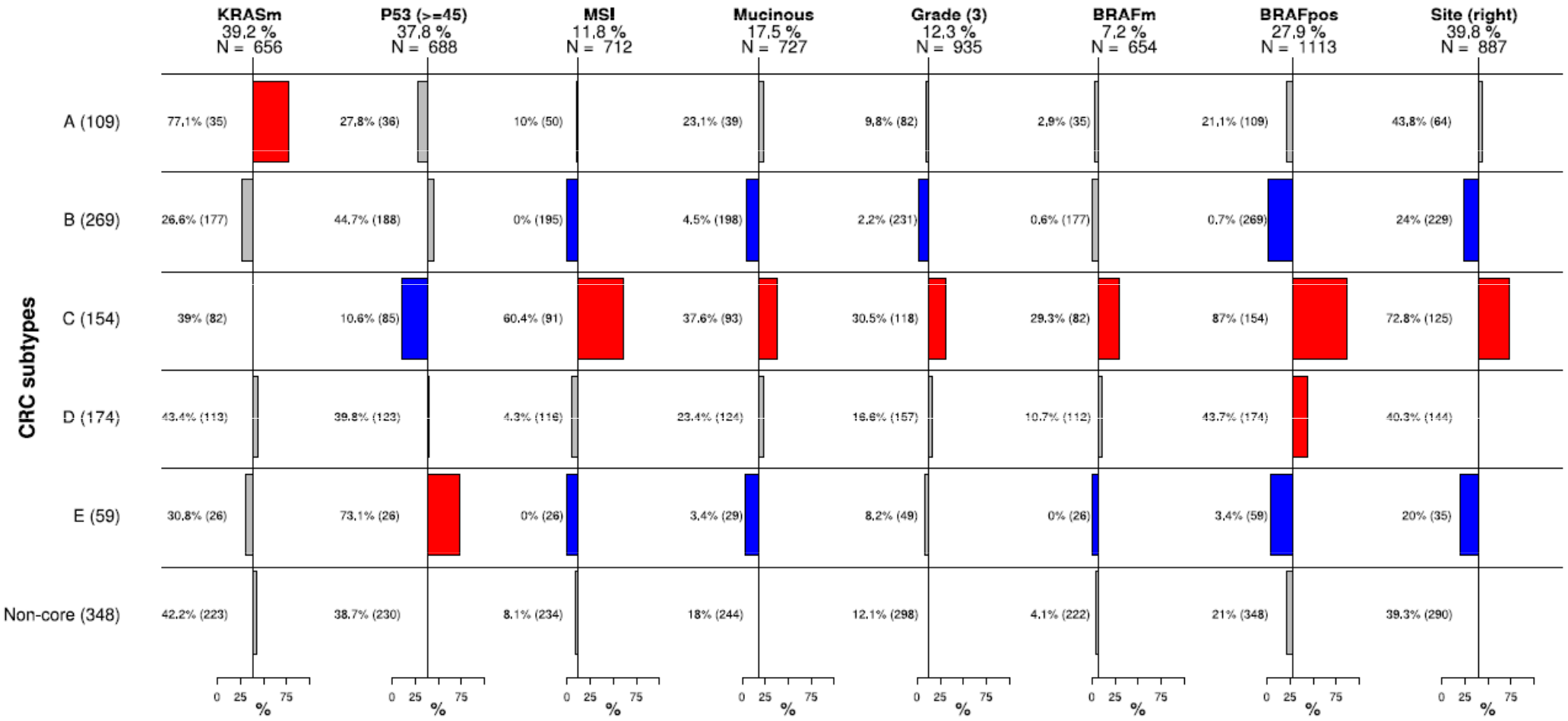
- Analýza génových sád pomocou KS testu

Rozdiely v prežití



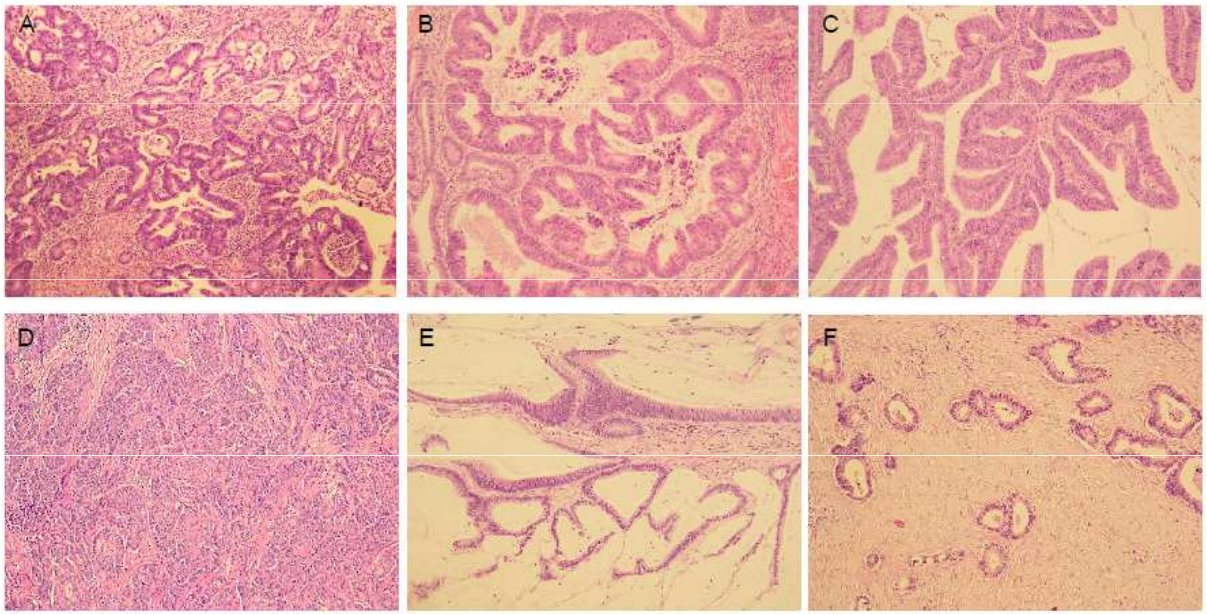
- Kaplan-Meierove krivky prežitia
- Coxov model proporcionálnych rizík (efekt štádia vs podtypov)

Charakterizácia podtypov klinickými a molekulárnymi premennými

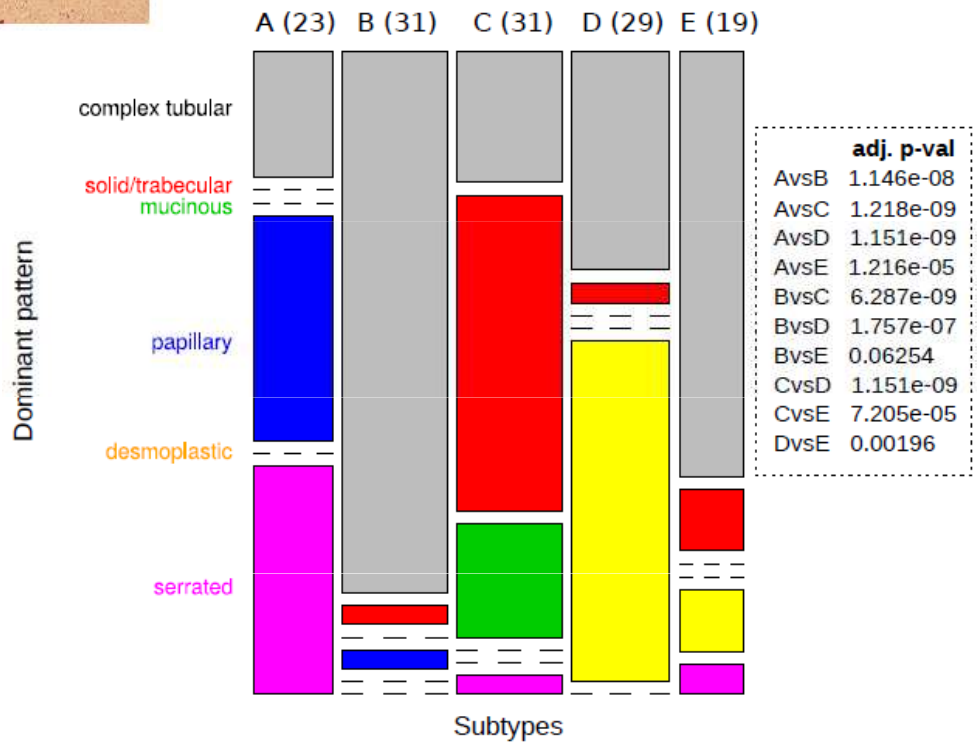


- Rozdiel od populačnej baseline u každého podtypu pomocou Fisherovho exaktného testu, FDR úprava p-hodnôt

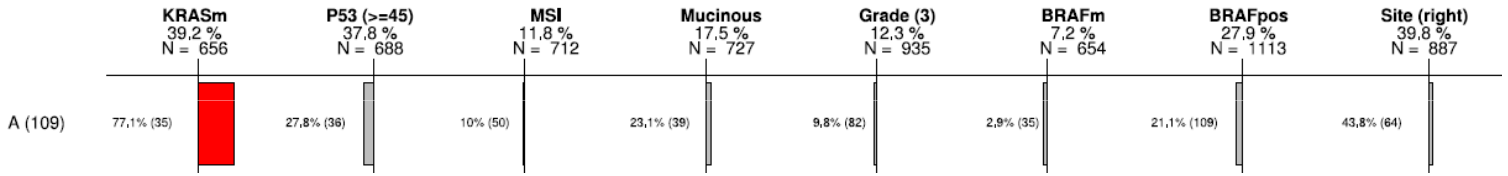
Histologické rozdiely

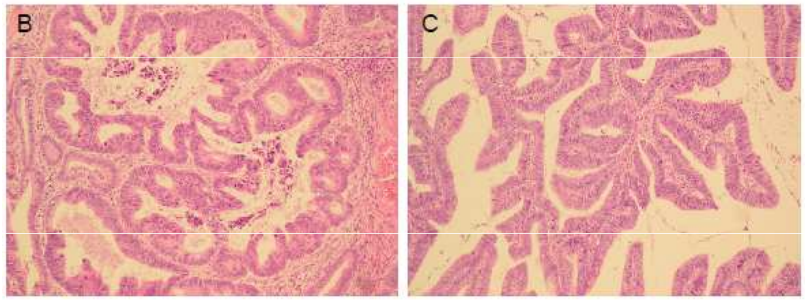


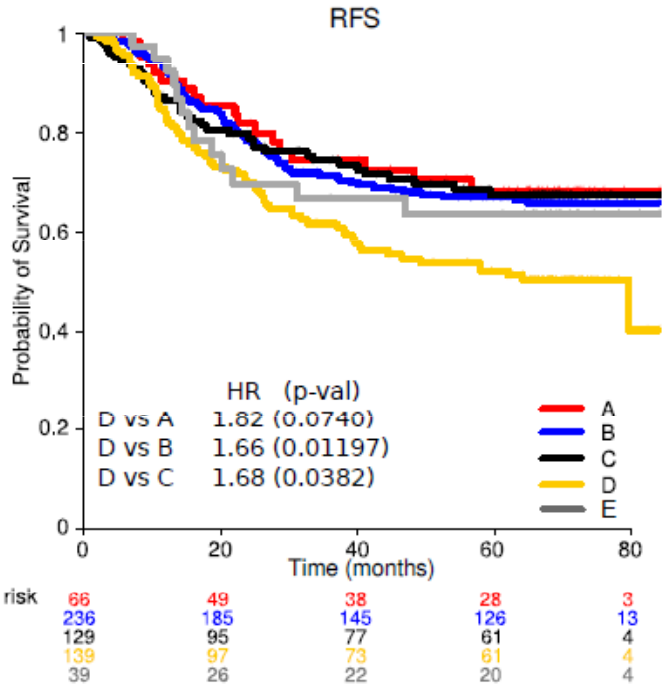
- Fisherov test, adjustácia na FDR

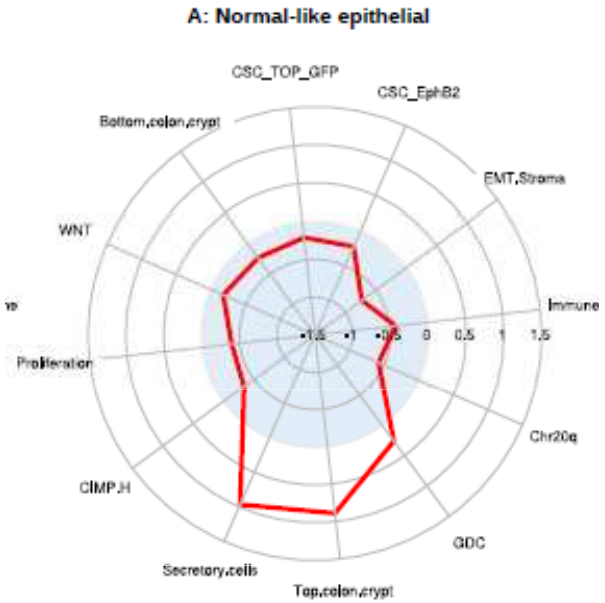


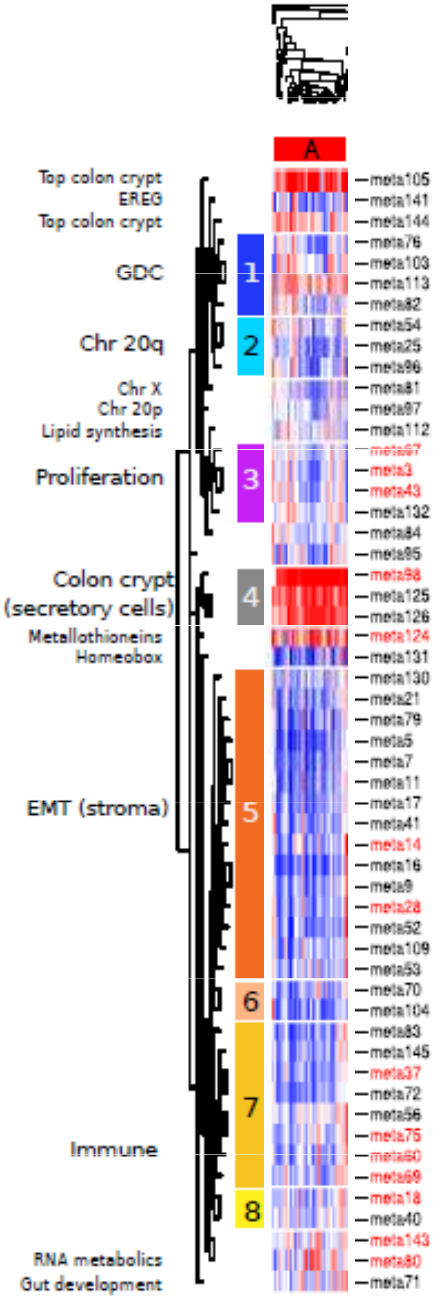
Podtyp A – Surface crypt like - *KRAS* mutanti, papillary a serrated morfotyp, najviac diferencovaný, bez aktívnej Wnt signálnej dráhy. Dobrý OS a RFS.



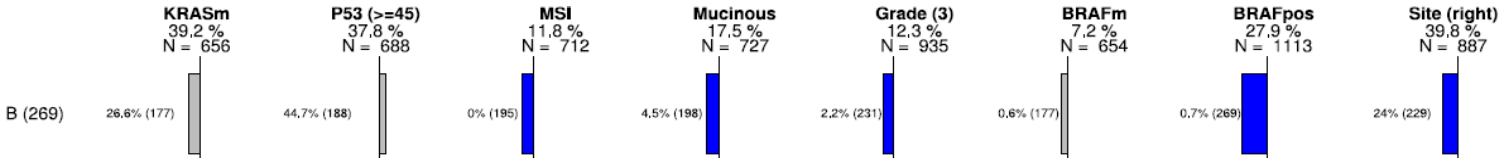


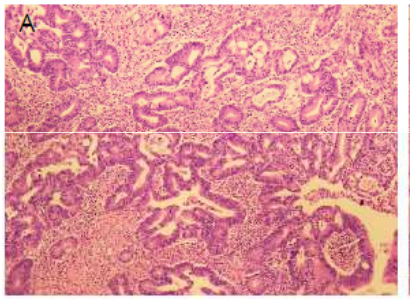


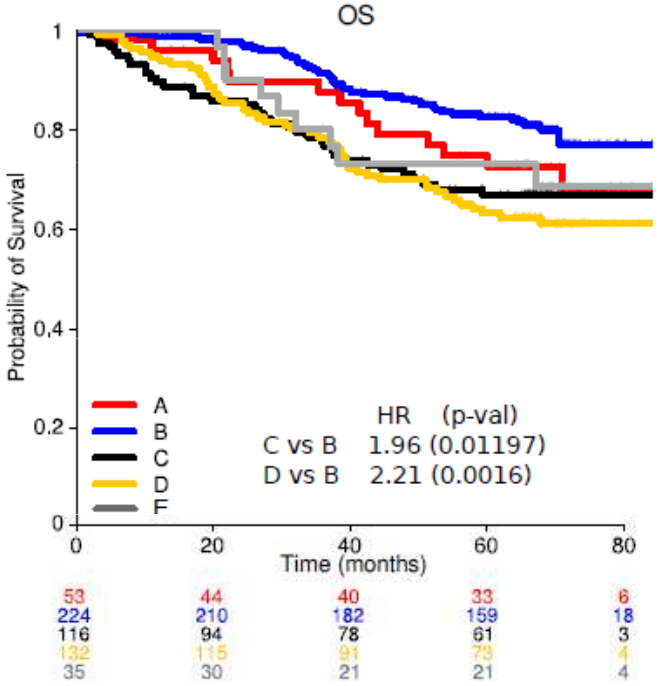


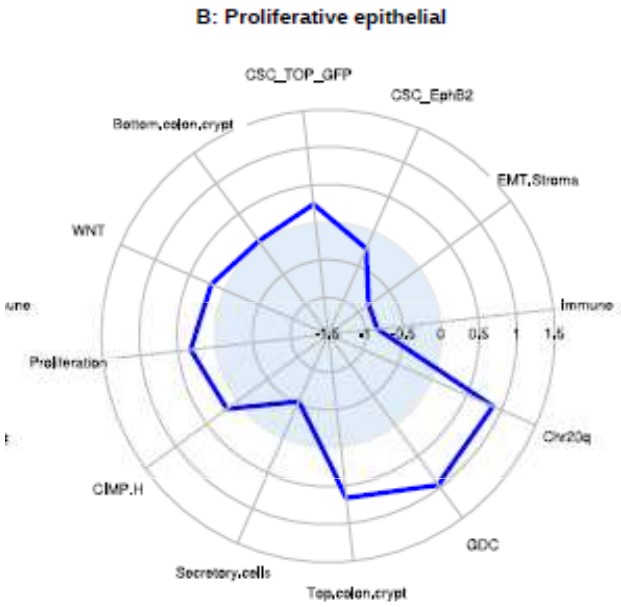


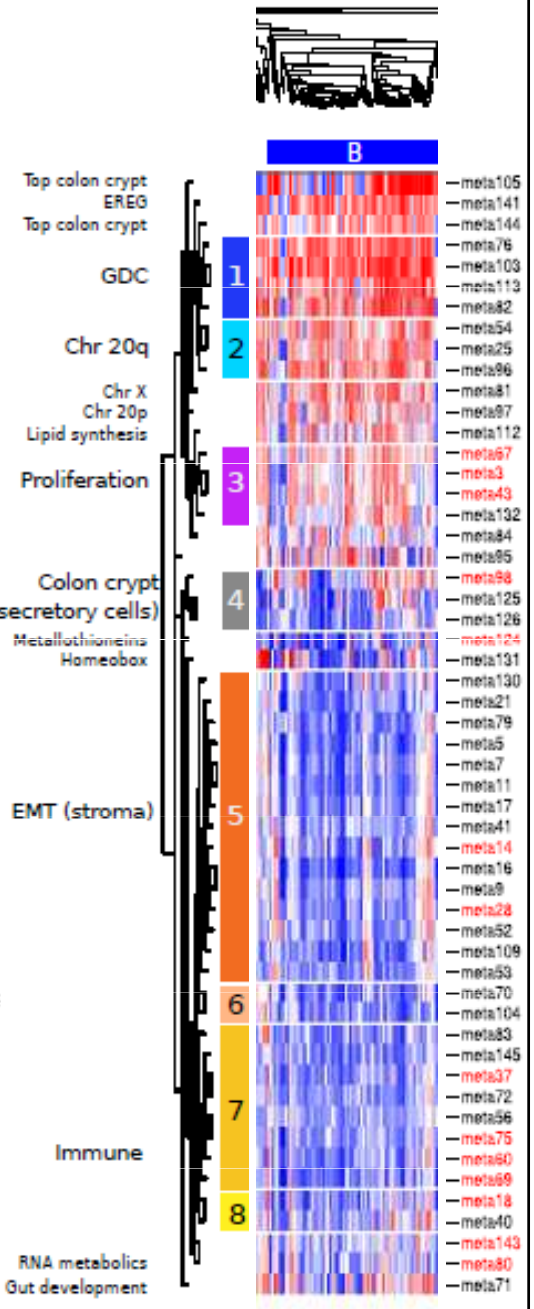
Podtyp B – **Lower crypt like** – diferencované ale bez sekrečných buniek, proliferujúce, a aktívnou Wnt signálnou dráhou. Komplexný tubulárny morfotyp. Časo MSS, BRAFwt, nižšieho grádu, dobré prežitie v OS, RFS i SAR.



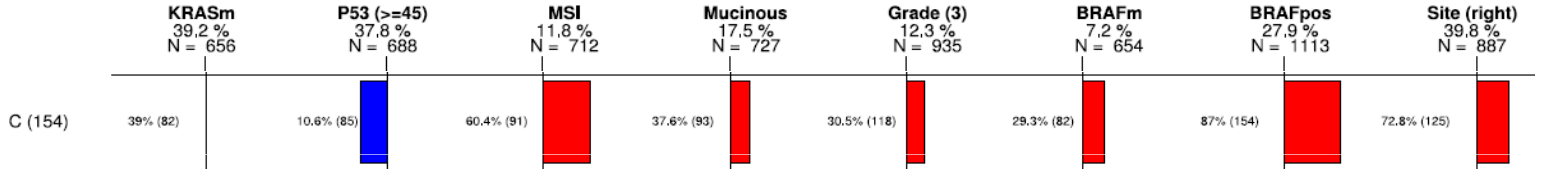


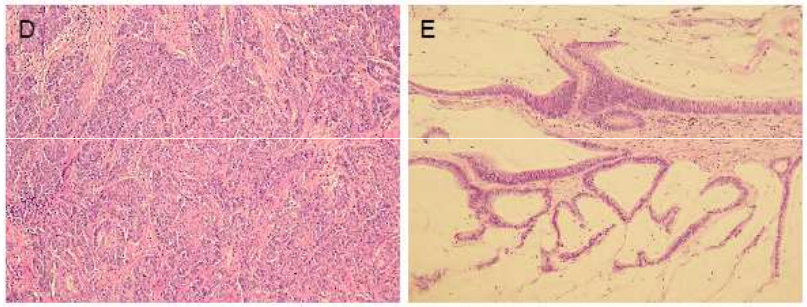


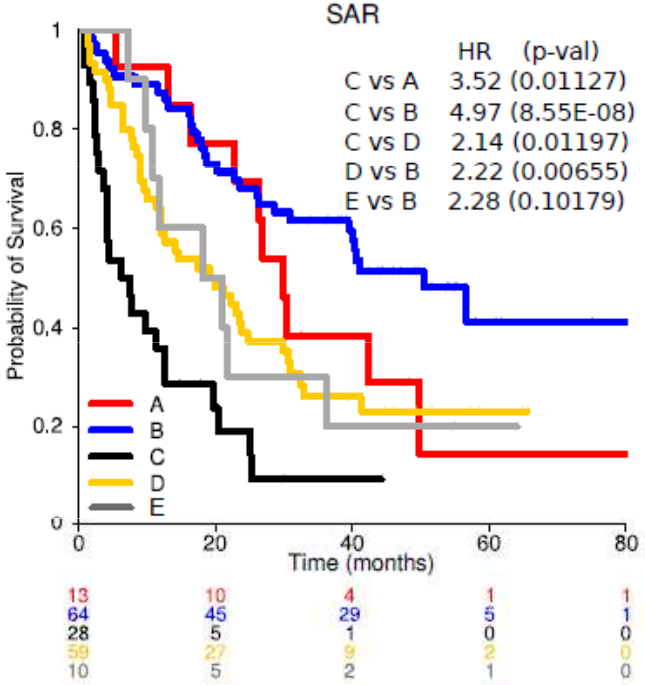


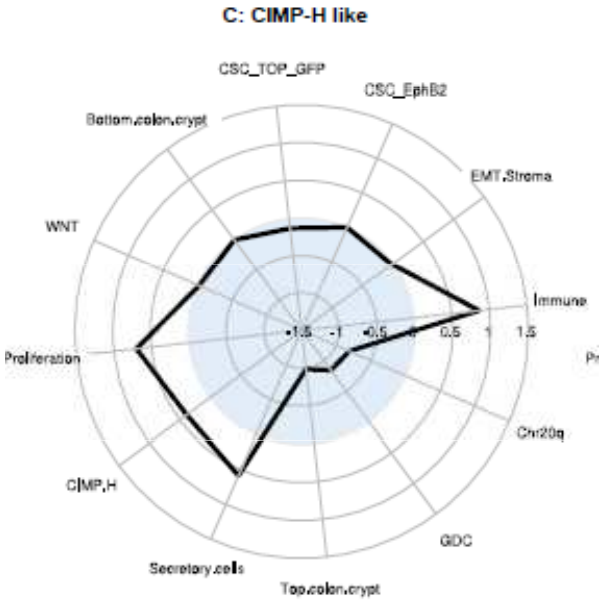


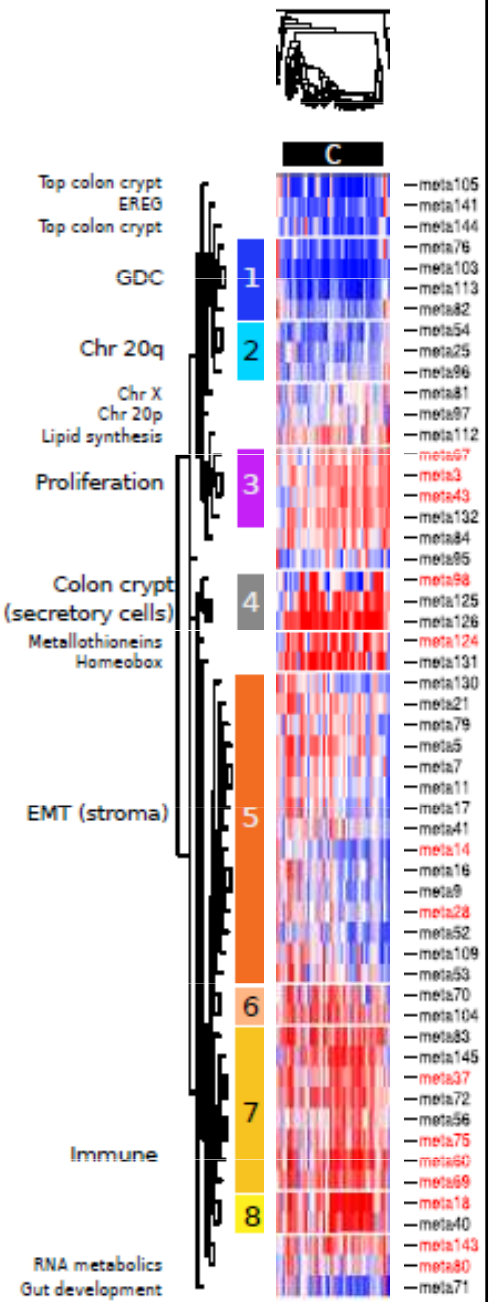
Podtyp C – CIMP-H like - časo MSI, *BRAF*-mutantné, hypermutované, z pravej časti hrubého čreva. Histologicky - horšie diferencované, solídno-trabekulárne s mucínovéým morfotypom. Aktívne proliferujú a majú silnú imunitnú reakciu. Dobrý RFS, ale zlý OS and SAR.



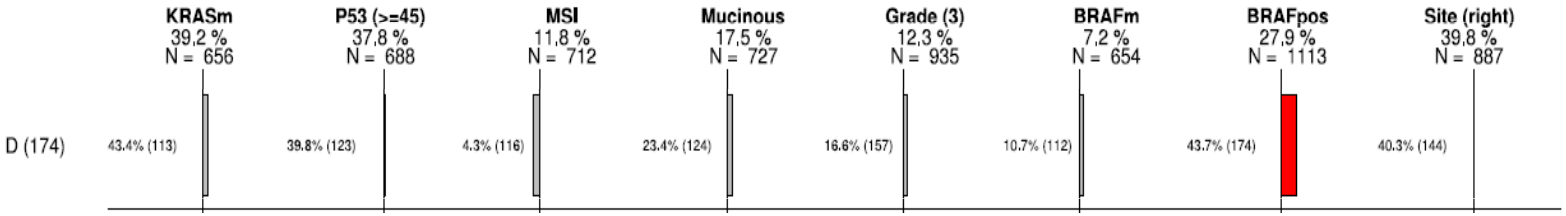


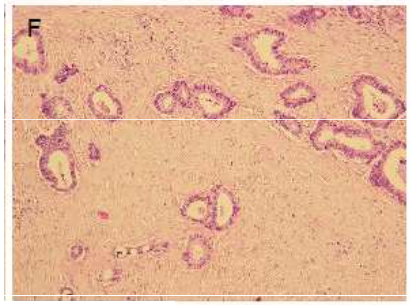


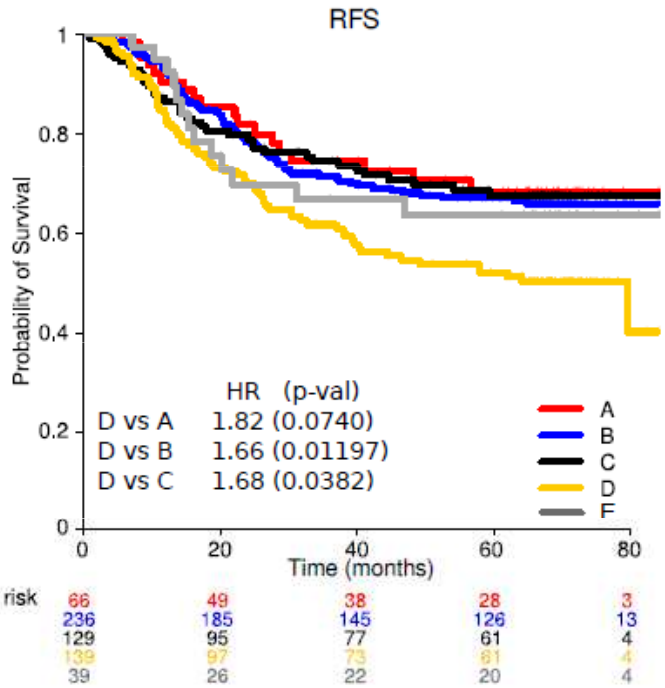


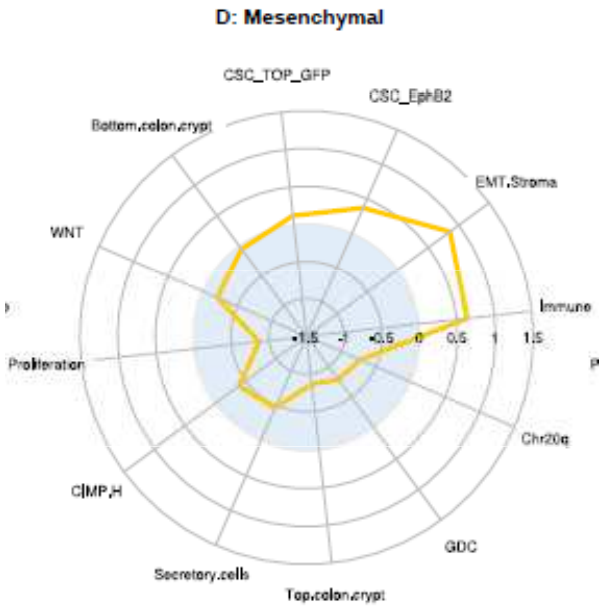


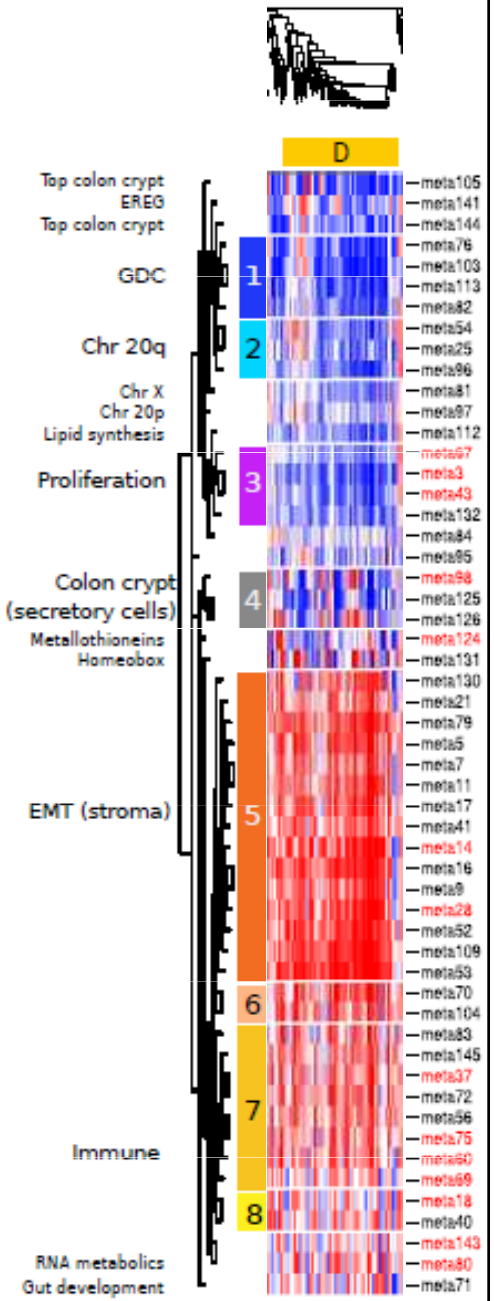
Podtyp D – Mesenchymal – markery kmeňových buniek, veľa mezenchymálnych buniek, ktoré sa prejavujú expresiou EMT génov. Wnt signálna dráha je neaktívna a proliferácia nízka. Klinické a mutačné charakteristiky sa nelíšia od populačnej baseline. Majú najkratšie prežitie do relapsu, zlý OS a SAR.



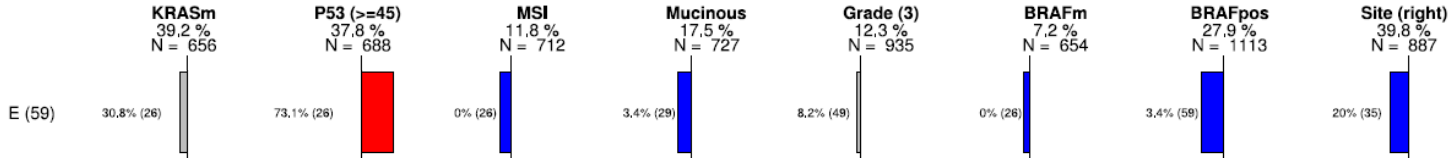


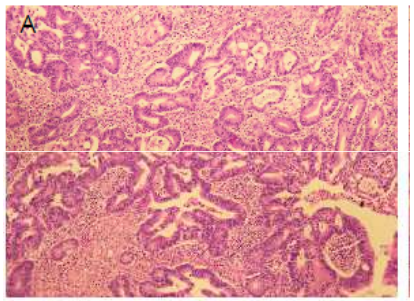


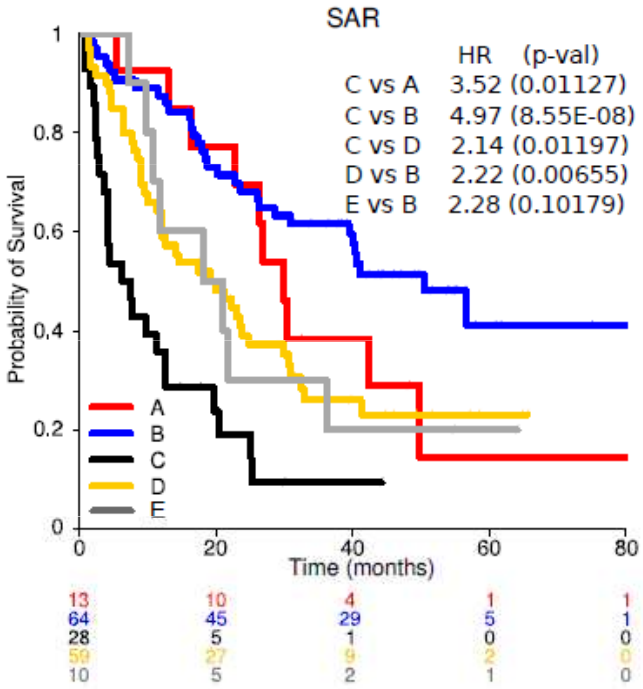


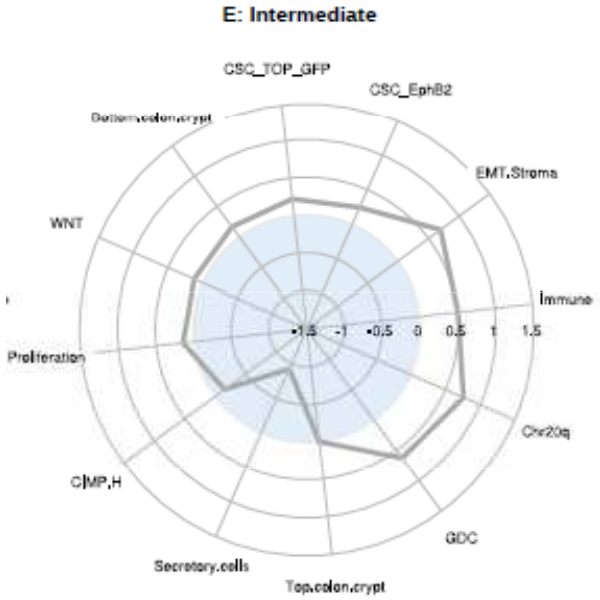


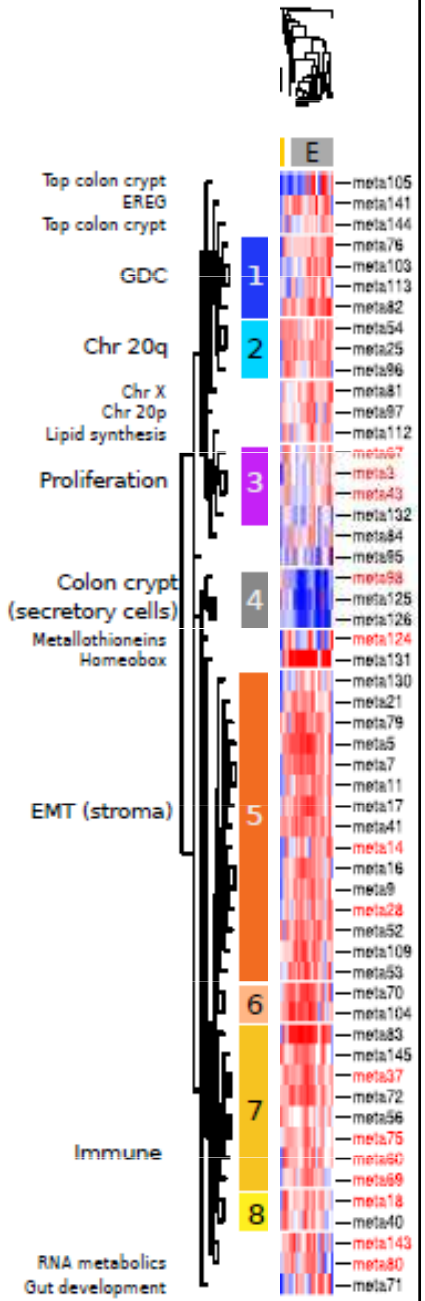
Podtyp E – **Mixed** – často MSS, *BRAF*wt, z ľavej strany hrubého čreva. Podobne ako podtyp D exprimuje gény kmeňových buniek a EMT procesu, avšak podobne s B má vysokú aktiviu kanonickej Wnt dráhy a vyzerá viac diferenciovaný. Je podobný B – komplexný tubulárny, len častejšie obsahuje mutáciu *p53*.







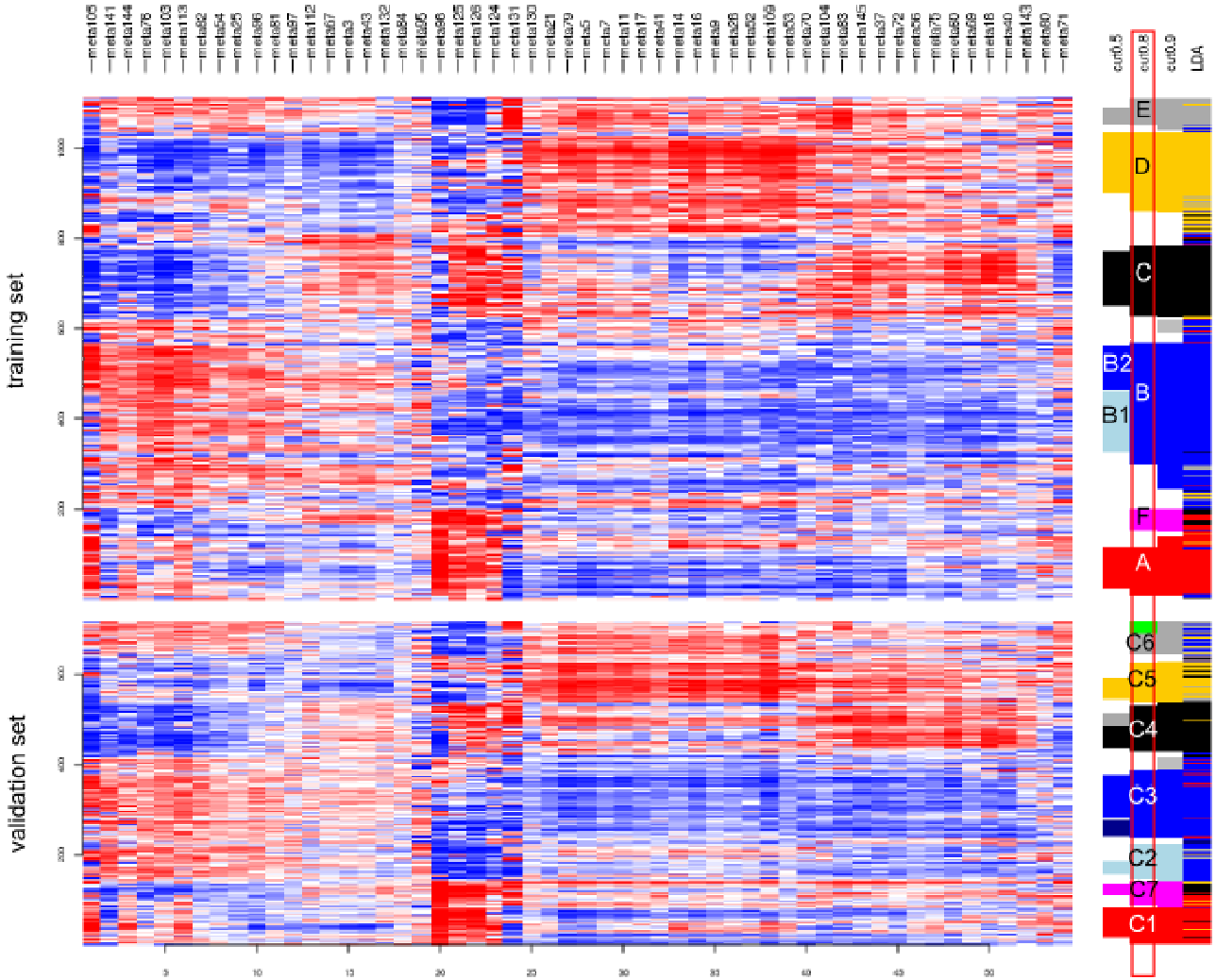




VALIDÁCIA V ZHLUKOVANÍ MOLEKULÁRNYCH DÁT

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

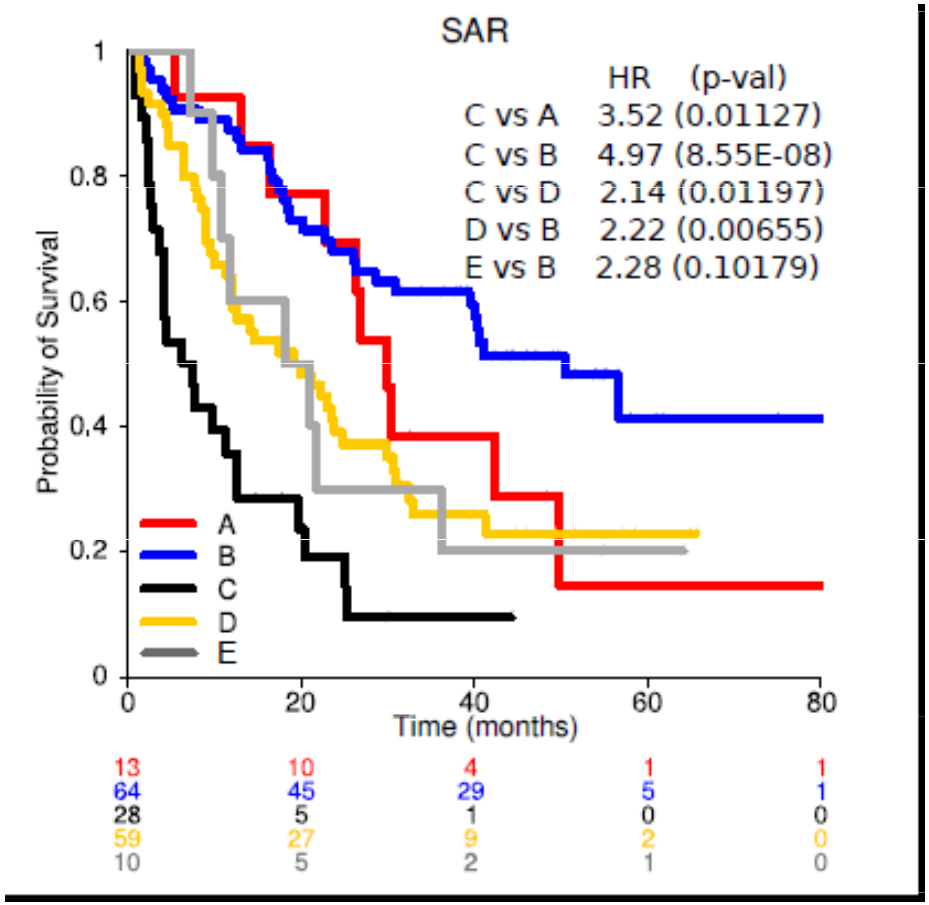
Validácia konceptu pomocou klinických, molekulárnych a histologických charakteristík objavených skupín

Majú objavené skupiny biologickú podstatu / odrážajú známe vedecké poznatky?

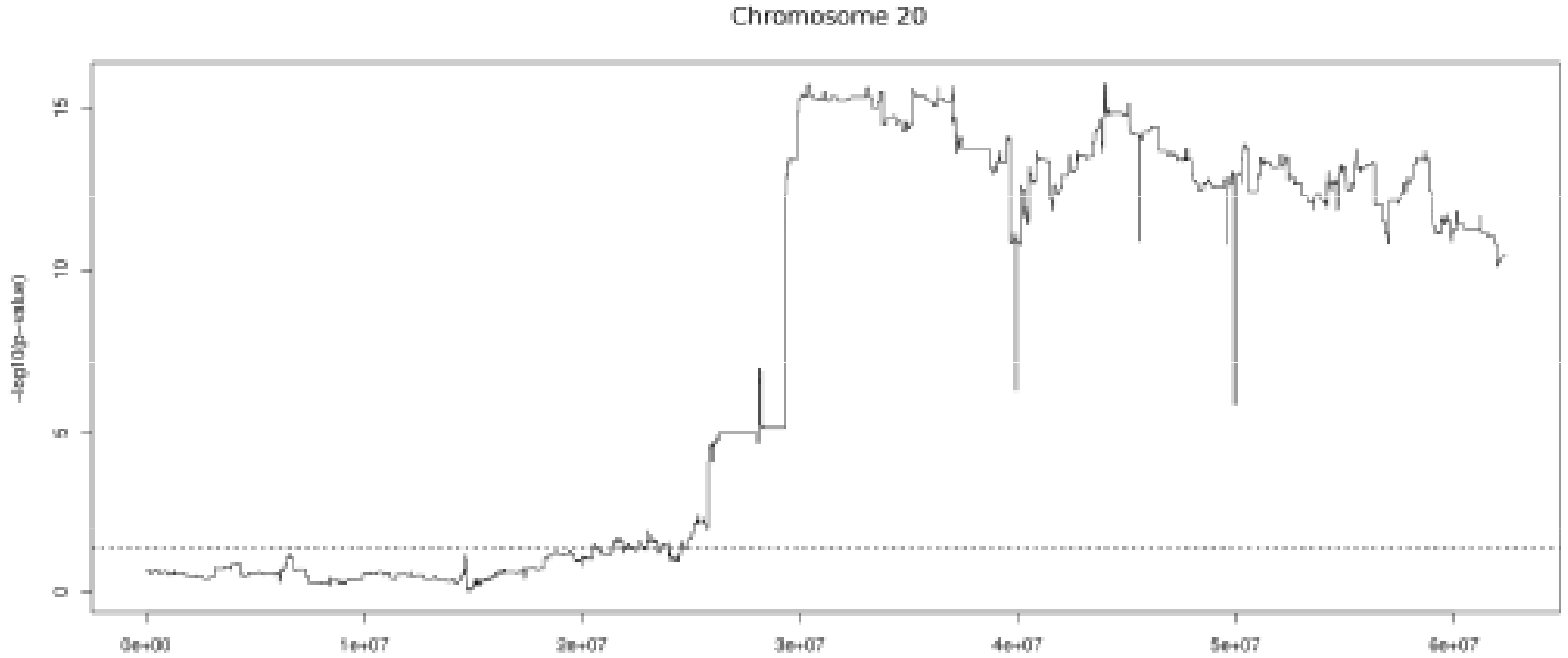
Je rozloženie týchto charakteristík medzi podtypmi porovnateľné vo validačnom súbore?

Majú objavené skupiny biologickú podstatu / odrážajú známe vedecké poznatky?

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



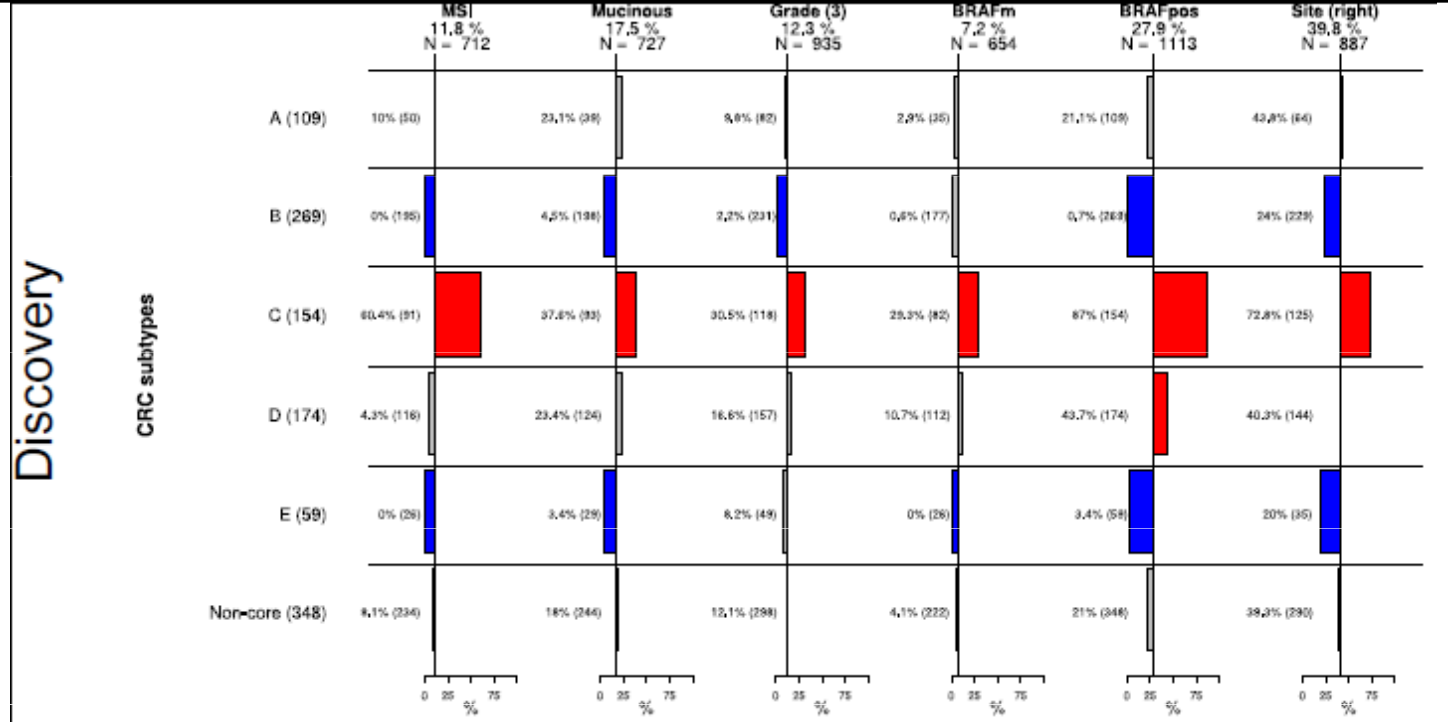
Zvýšená expresia génov chr. 20q v podtype B by mohla znamenať amplifikáciu chr20q regiónu.

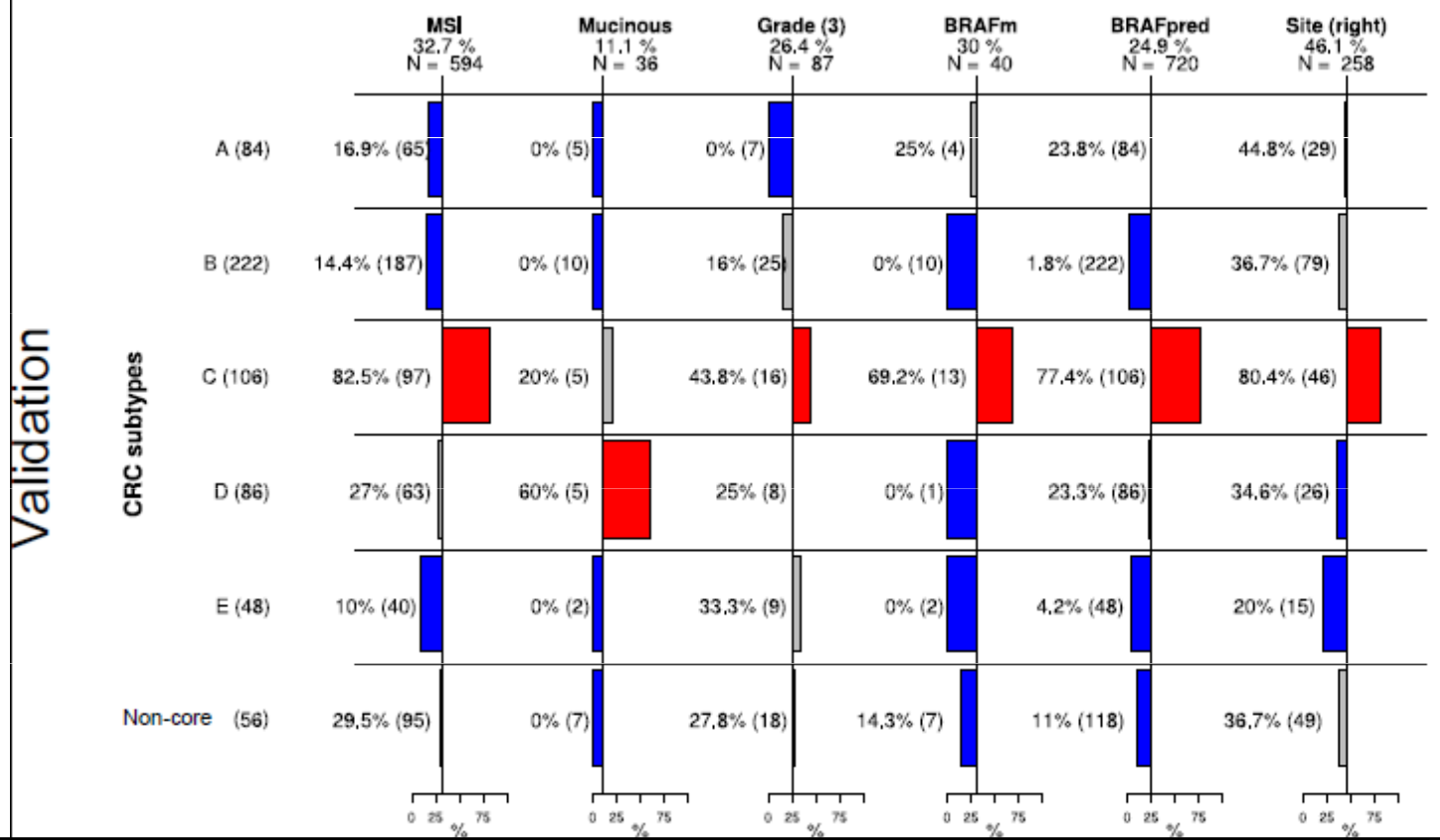


Podtyp D – mezenchymálny – histologické vyhodnotenie: v nádore prítomná silná desmoplastická reakcia (mezenchymálne tkanivo)

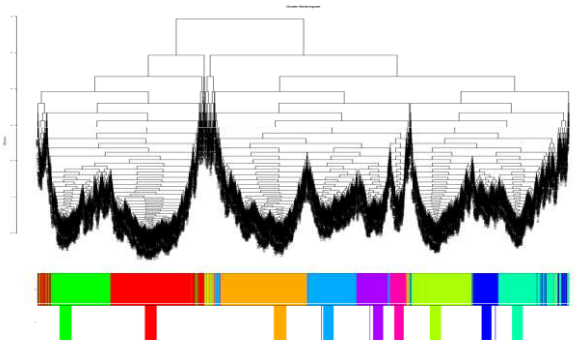
Nádor

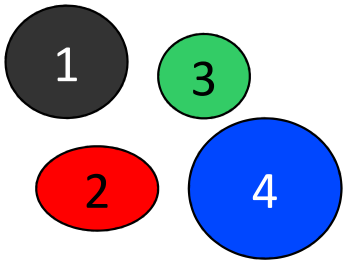
Je rozloženie klinických charakteristík medzi podtypmi porovnateľné vo validačnom súbore?


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)
10% (50)
23.1% (38)
9.6% (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% (28)
3.4% (29)
8.2% (49)
0% (26)
3.4% (59)
20% (35)
6.1% (234)
16% (244)
12.1% (288)
4.1% (222)
21% (348)
38.3% (290)
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)
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
%

Dataset 1 | Dataset 2 ..... | Dataset N



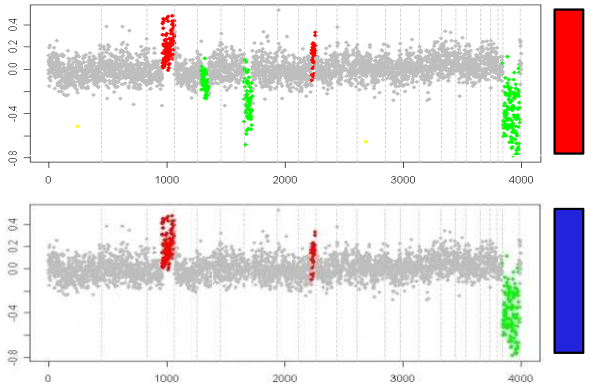


Klinické dáta

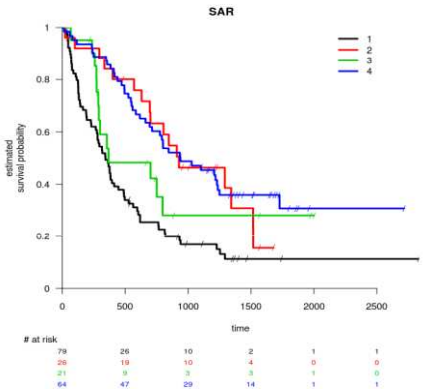
Right	MSI	KRAS	BRAF
35%	25%	10%	10%
15%	...		
0%		...	
50%			...

Definícia podtypov karcinómu

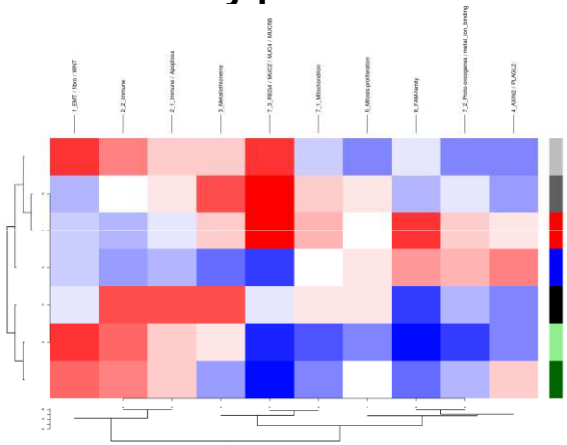
Iné zdroje dát (CNV, metylácia...)



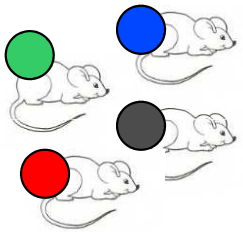
Prežitie

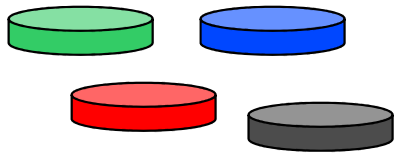


Molekulárny profil



Rezistencia na liečbu





Zhrnutie

- Konsenzusové zhlukovanie
- Dynamické rezanie stromu
- Validácia neznámej pravdy – medzi dátovými súbormi
- Množstvo metód v rámci jednej štúdie
- Je to úžasné