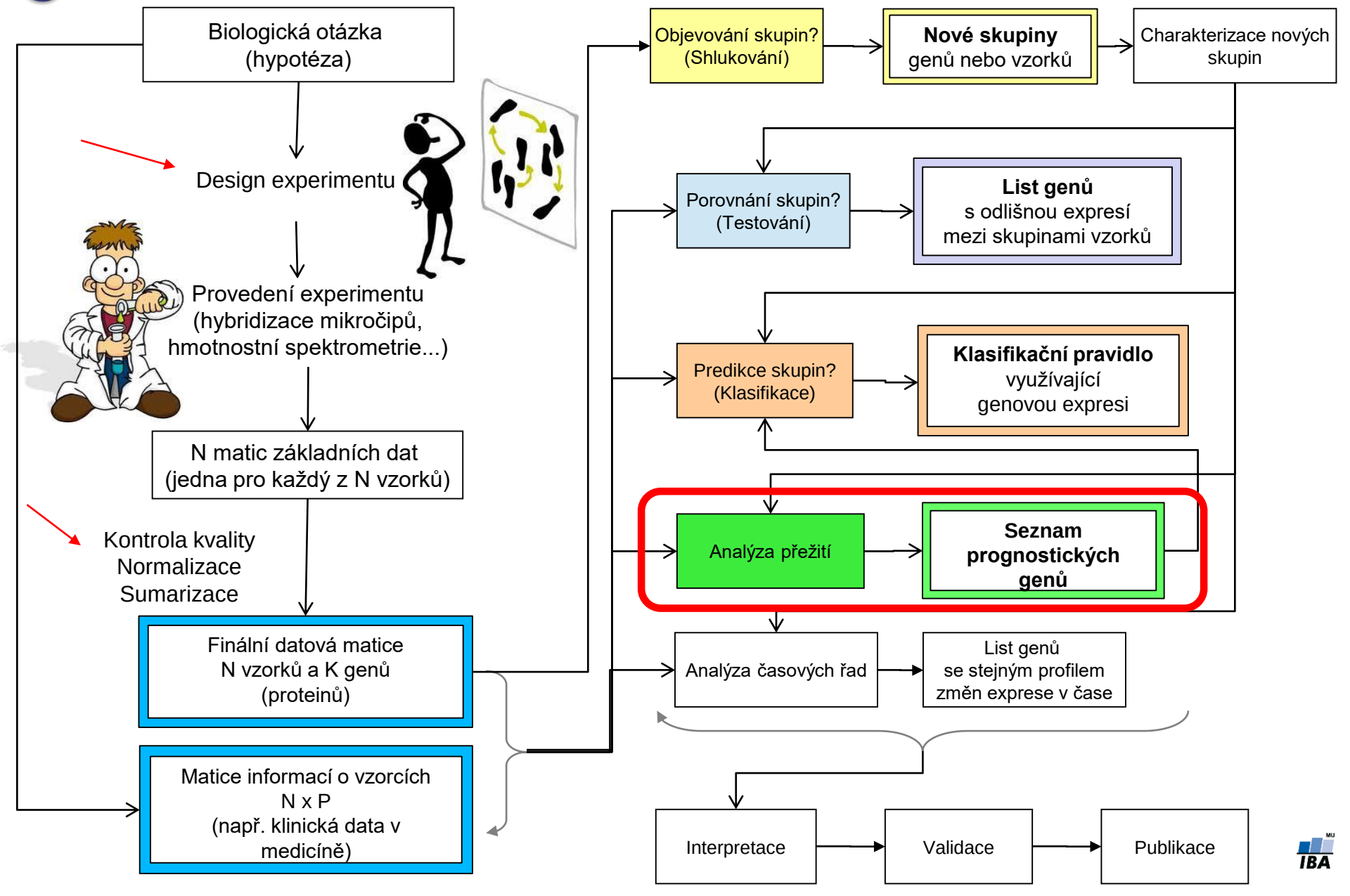
Analýza přežití

Společná schéma analýzy dat



Analýza přežití a genomická data

- Dva základní cíle s přesně definovanými otázkami:
  1. **Predikce rizika molekulárních skupin** – Mají skupiny definované pomocí genové exprese rozdílné přežití?
  2. **Predikce času přežití na základě genové exprese** – Má exprese genu vliv na přežití?

- **Přežití** je čas do nějaké námi sledované událost<u>i</u>
  - Úmrtí (overall survival)
  - Relaps (relaps-free survival)
  - Návrat onemocnění (disease-free survival)
  - …

Data analýzy přežití

- Sbírání dat od zadaného času (začátek studie, diagnóza)
- Dvě proměnné:
  - Výsledek – Nastala událost?

    1 = událost nastala

    0 = událost nenastala
  - Čas přežití

Pacient	Výsledek	Čas prežití (měsíce)
1	1	4
2	1	11
3	0	56*
4	1	8
5	0	44*
6	0	48*
7	0	57*
8	1	3

- Časy přežití s výsledkem 0 představují tzv. **cenzorované hodnoty** – je to čas do konce pozorování, a nebo posledního záznamu

Metody analýzy prežití

- V závislosti od otázky sa používají dvě základní funkce:
  - **Kaplan – Meierův odhad prežití** – Mají skupiny definované pomocí genové exprese rozdílné přežití?
  - **Coxův model proporcionálních rizik** – Má exprese genu vliv na přežití?

Kaplan-Meierův odhad prežití

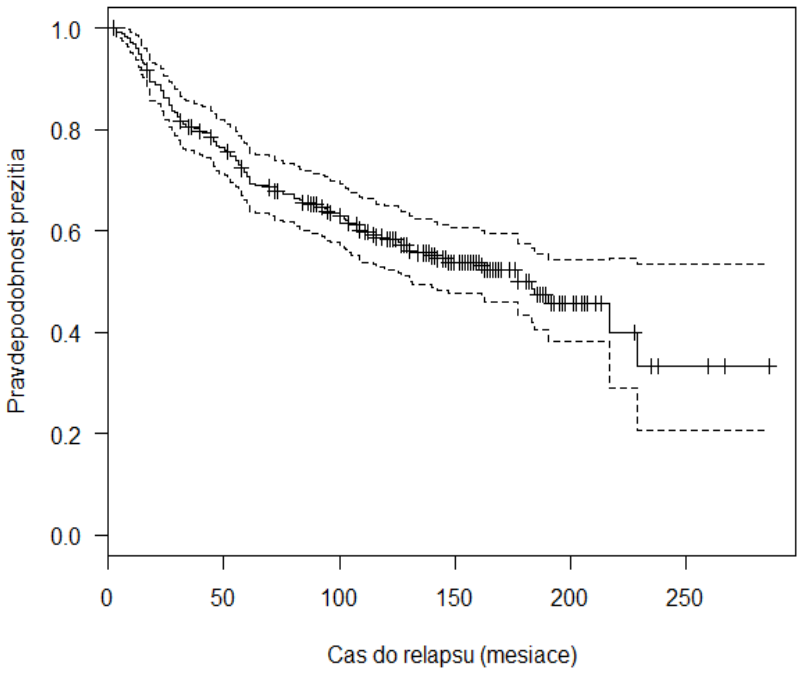
- Definovaný jako pravděpodobnost prežití do času *t*
- Pro každý časový interval *t* odhadne podíl přežívajících, za použití pravděpodobnosti
- Jedná se o neparametrický odhad

$$S(t) = p_1 * p_2 * p_3 * \ldots * p_t$$

$$p_i = \frac{r_i - d_i}{r_i}$$

r_i –počet živých pacientů na začátku časového intervalu t_i

d_i –počet úmrtí za časový interval t_i



Kaplan-Meierův odhad prežití - příklad

t(i)	N	# úmrtí	# censor.	Risk
0	21	0	0	1
6	21	3	1	1*(18/21) = 0.8571
7	17	1	1	0.8571*(16/17) = 0.8067
10	15	1	2	0.8067*(14/15) = 0.7529
13	12	1	0	0.7529*(11/12) = 0.6902
16	11	1	3	0.6902*(10/11) = 0.6275
22	7	1	0	0.6275*(6/7) = 0.5378
23	6	1	5	0.5378*(5/6) = 0.4482

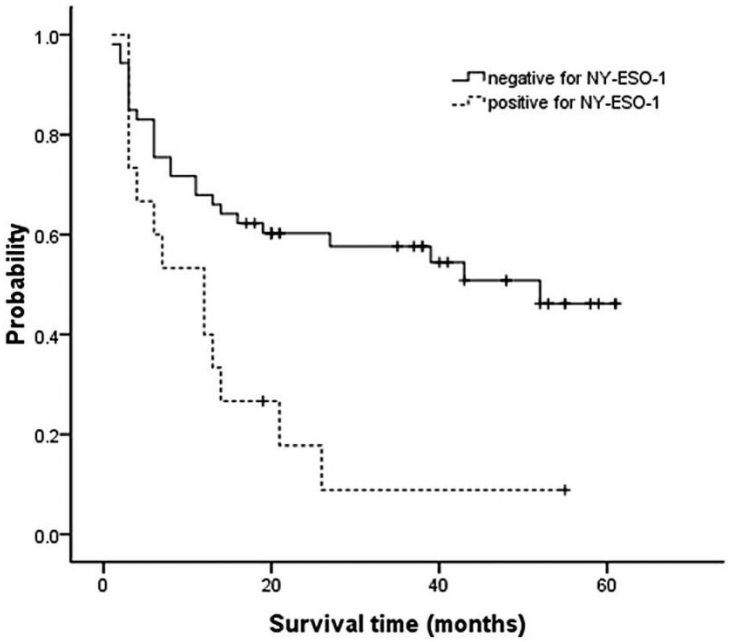
Porovnání křivek přežití

- Dva testy pro zjištění párových rozdílů v analýze přežití:
  - **Gehanův-Breslowův-Wilcoxonův test**
    - Přiřazuje větší váhy úmrtím v dřívějších časových bodech
    - Může být zavádějící, pokud je velké procento pacientů cenzorováno v dřívějších časových bodech
  - **Mantelův-Haenszův log-rank test**
    - Standardně používaný
    - Předpokládá nezávislost cenzorování a výskytu jednotlivých událostí
    - Přiřazuje stejné váhy úmrtím ve všech časových bodech
    - Silný, pokud je předpoklad proporcionality rizik splněný

Log-rank test

- Umožňuje statistické zhodnocení rozdílu v přežití, ale neposkytuje kvantifikaci tohoto rozdílu a nebere v úvahu vliv dalších proměnných
- j – čas, Z – testová statistika, hypotéza je zamítnutá $Z>z_a$

$$Z=\frac{\sum_{j=1}^{J}(O_{1j}-E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}$$



$$V_j = \frac{O_j(\frac{N_{1j}}{N_j})(1-\frac{N_{1j}}{N_j})(N_j-O_j)}{N_j-1}$$

$$O_j = O_{1j} + O_{2j}$$

$$N_j = N_{1j} + N_{2j}$$

$$E_{1j} = \frac{O_j}{N_j} N_{1j}$$

O_j – počet pozorovaných událostí v čase j
N_j – počet subjektů v riziku
E_{1j} – očekávané hodnoty

Coxův model proporcionálních rizik

- Měříme okamžité riziko události
- Můžeme testovat víc proměnných (nejen genovou expresi, ale i další jako věk, …)

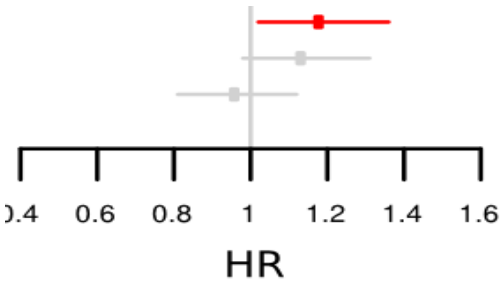
$$h(t,x) = h_0(t)e^{(b_1x_1+\cdots+b_px_p)}$$

- h(t) je riziko v čase t
- $h_0(t)$ je základní riziková funkce společná pro všechny subjekty
- $x_1, x_2, \ldots x_p$ jsou vysvětlující proměnné
- $b_1, b_2, \ldots b_p$ jsou odhadnuté regresní koeficienty

Coxův model proporcionálních rizik II.

- Předpoklad: proporcionalita rizika
- Interpretace parametrů: změna o jednotku znamená změnu rizika h(t) o hodnotu odhadnutého koeficientu
- Změna genové exprese o jednotku se nedá přímo porovnat mezi experimenty (čísla jsou relativní, mají jinou škálu) => škálování expresních hodnot genu g($\mathbf{x_g}$) před analýzou:

$$y_g = \frac{x_g\text{-median}(x_g)}{IQR(x_g)}$$

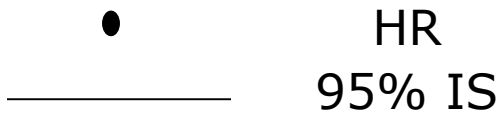


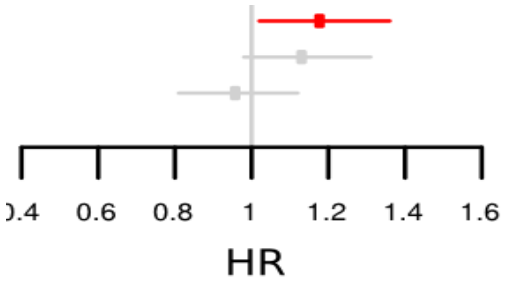
Coxův model proporcionálních rizik III

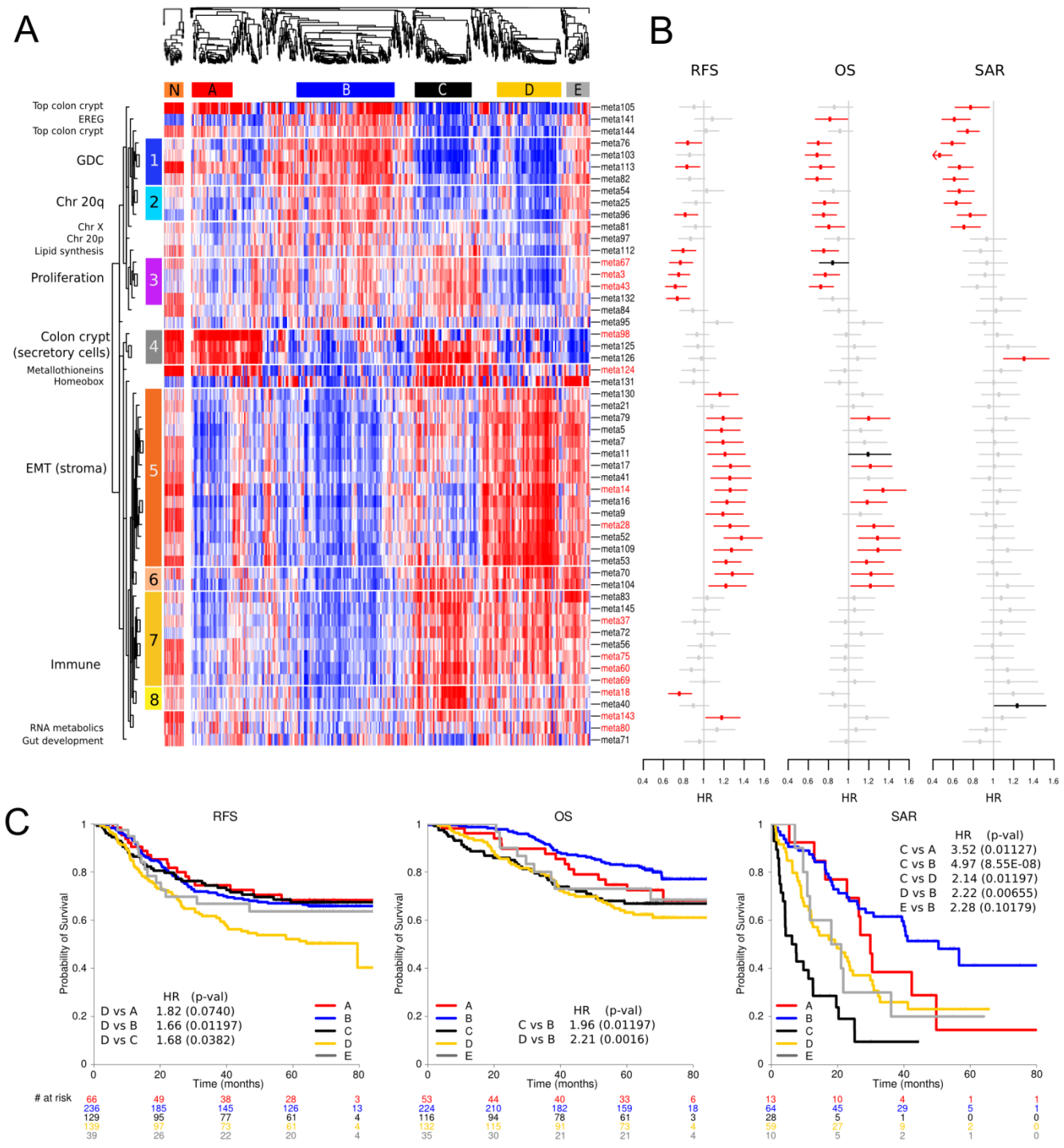
- Predpoklad: proporcionalita rizik:
  - Podíl rizikových funkcí libovolných dvou jedinců je proporcionální, nezávislý na čase (třeba otestovat)

- Jinak použít parametrické metody

- Zobrazeni (tzv. forestplot)






A
N
A
B
C
D
E
Top colon crypt
EREG
Top colon crypt
GDC
Chr 20q
Chr X
Chr 20p
Lipid synthesis
Proliferation
Colon crypt
(secretory cells)
Metallothioneins
Homeobox
EMT (stroma)
Immune
RNA metabolics
Gut development
1
2
3
4
5
6
7
8
B
RFS
OS
SAR
HR
C
RFS
Probability of Survival
Time (months)
HR (p-val)
D vs A 1.82 (0.0740)
D vs B 1.66 (0.01197)
D vs C 1.68 (0.0382)
OS
HR (p-val)
C vs B 1.96 (0.01197)
D vs B 2.21 (0.0016)
SAR
HR (p-val)
C vs A 3.52 (0.01127)
C vs B 4.97 (8.55E-08)
C vs D 2.14 (0.01197)
D vs B 2.22 (0.00655)
E vs B 2.28 (0.10179)
at risk

Zavolat statistika po vykonání experimentu je asi jako požádat doktora o posmrtné vyšetření:
pravděpodobně bude schopný říct, na co experiment zemřel.

Ronald Fisher