MUNI | RECETOX

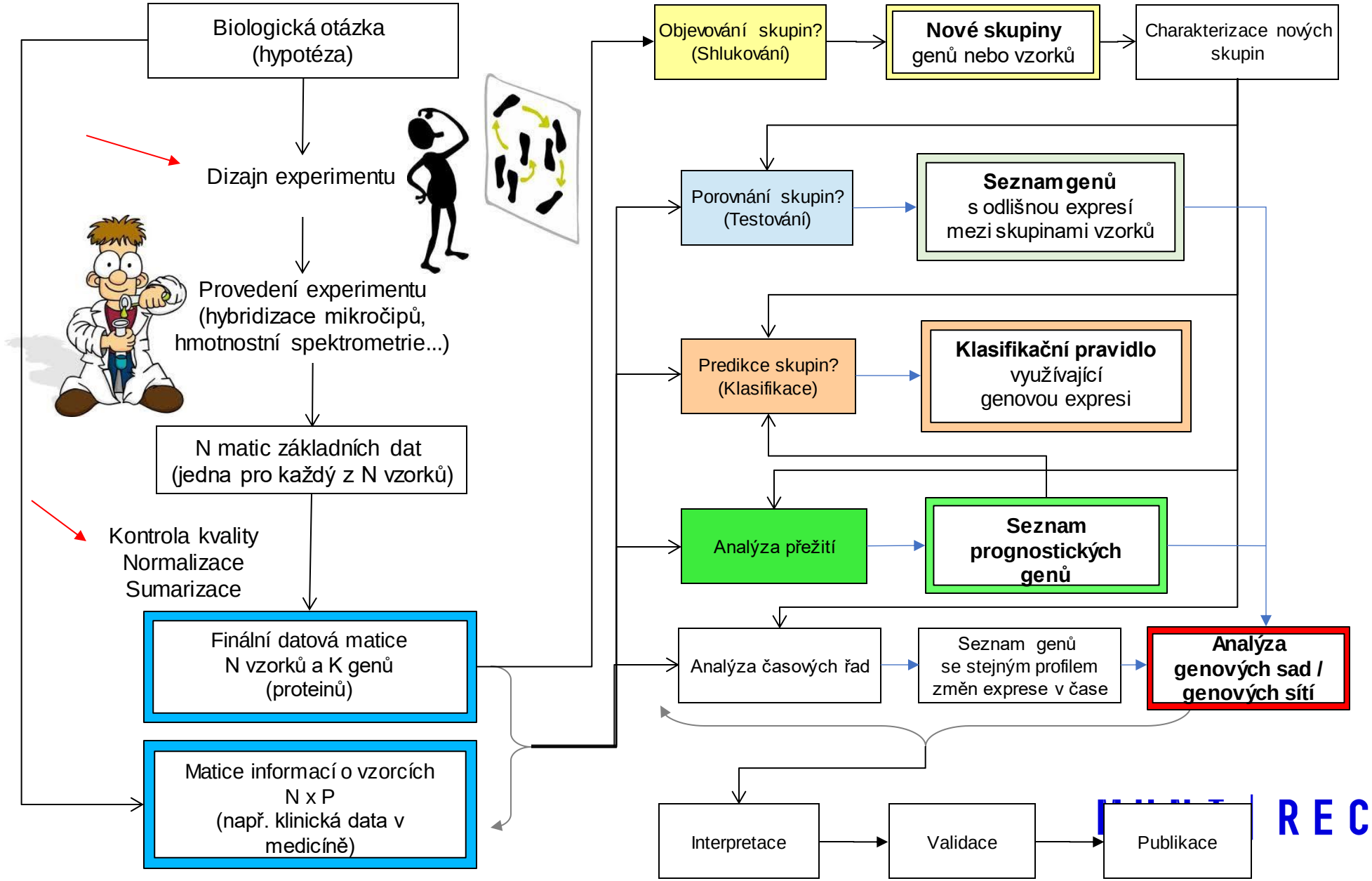
Analýza genomických a proteomických dat

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

- Jaro 2022

MUNI | RECETOX

Společné schéma analýzy dat



RECETOX

Analýza genových sad

(pathway analýza)

MUNI | RECETOX

Motivace

- Geny, proteiny a další molekuly jsou navzájem propojené ve velké spleti různých signálních, metabolických a různych jiných drah
- Jak odhalit tyto závislosti?
  - Geny, které najdeme odlišně exprimované mezi skupinami (porovnání skupin) můžeme ad-hoc vložit do databáze a podívat se kam patří (KEGG, MsigDB....)
    - nevýhoda – nemáme statistickou významnost, která z drah je zastoupená nejvíce
  - Můžeme přímo porovnávat všechny geny se skupinami genů v jednotlivých dráhách
- **Předpoklad těchto analýz:** operují s již definovanými skupinami genů

Genová sada vs dráha

Génová sada je jakákoliv množina genů, například

všechny geny patřící do jedné dráhy	všechny geny které mají podobnou funkci	…

Sada genů nemusí být dráha – je to všeobecnější a méně specifický pojem

Cíl

- Cíl je přiřadit každé genové sadě, případně dráze jedno číslo - skóre, a nebo p-hodnotu, abychom mohli odpovědět na otázku:

  Kolik genů je v sadě(pathway) odlišně exprimovaných a je to dostatečně statisticky významné, abychom mohli říct, že je tato dráha specifická jen pro naše porovnávané skupiny?

Databáze genových sad (pathways)

Gene Ontology (GO) databáze

- http://www.geneontology.org/
- Hierarchická databáze
- Rodičovské uzly: obecnější termíny
- Potomci uzlů: víc specifické
- Na konci hierarchie jsou molekuly (geny/proteiny)
- Na vrcholu jsou 3 rodičovské uzly:
  - Biologické procesy
  - Molekulární funkce
  - Buněčné složky

KEGG pathway databáze

- KEGG = Kyoto Encyclopedia of Genes and Genomes
- http://www.genome.jp/kegg/pathway.html
- Více informací než GO, máme tu již vztahy mezi geny a genovými produkty
- Detailní informáce jen pro některé organizmy a procesy
- Využívá hlavně ověřené poznatky, nemůže ji kdokoliv změnit
- Proto se tu nenachází všechny geny (obvykle tak třetina až polovina z hledaných)
- Aktualizovaná databáze není volně přístupná

MsigDB databáze

- https://www.gsea-msigdb.org/gsea/msigdb

Metody analýzy genových sad

MUNI | RECETOX

Rozdělení metod

Podle toho s jakou informací pracují na

- *metody dělící hranice* – berou v potaz jen informaci "významný" vs. "nevýznamný" gen
- *metody celého seznamu genů* – pracují přímo *p*-hodnotami (i nevýznamnými!) a tedy s pořadím

Podle skupiny molekul které analyzují na:

- *uzavřené* – analýza jen v rámci genů v sadě
- *kompetitivní* – porovnání se všemi geny experimentu

Nové metody pracují i s topologií dráhy

Dělení metod dle skupiny molekul které analyzují

Uzavřené vs. kompetitivní

Uzavřená metoda používá jen hodnoty genů z dané sady:

- H_0 : “Žádné geny z genové sady nejsou odlišně exprimované”

Kompetitivní test porovnává geny v genové sadě s ostatními geny v experimentu

- H_0 : “Podíl odlišně exprimovaných genů v genové sadě není odlišný od podílu odlišně exprimovaných genů mezi ostatními geny v experimentu”

Příklad

Datový soubor 12 639 genů. Z nich FDR< 5% má 1272 genů

96 genů v genové sadě, z toho 8 má FDR < 5%

Kolik odlišně exprimovaných genů v sadě očekáváme náhodně?

1. Dělící hranice byla 5% FDR

2. V případě, že platí nulová hypotéza, náhodně očekáváme 96 x 5% = 4.8 významných genů (falešná pozitivita)

3. Pomocí binomického testu vypočteme pravděpodobnost pozorování **8** a více významných genů: p = 0.1079, teda není významné

```
binom.test(x=8,n=96,p=0.05, alternative="greater")
```

Příklad, uzavřená metoda dělící hranice

Příklad, kompetitivní metoda dělící hranice

	V GS	Není v GS
Význ	8	1264
Nevýzn	88	11279

p = 0.73 (Fisherův test – jednostranný)

- 1272 z 12639 genů je odlišně exprimovaných v tomto datovém souboru (to je zhruba 10%)
- V množině náhodně vybraných 96 genů očekáváme tedy 96 x 10% = 9.6 významných genů
- p-hodnotu vypočítáme z kontingenční tabulky pomocí Fisherova nebo Chi-kvadrát testu

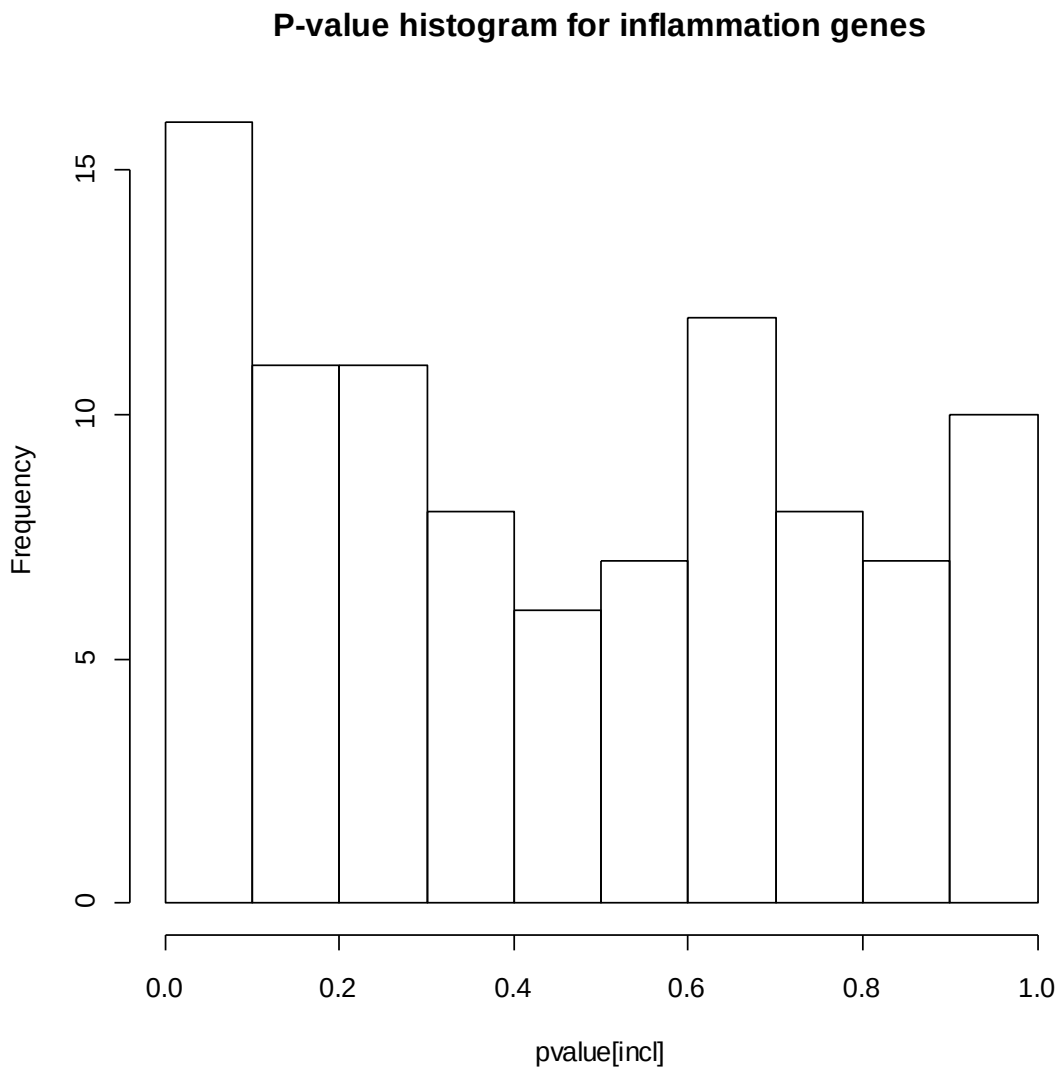
Dělení metod podle toho s jakou informací pracují

Metody dělící hranice vs. metody celého seznamu

- Dvě předchozí metody byly závislé na dělících hranicích – cut-offs a tedy závislé na N (p hodnota se mění v závislosti od zvyšujícího se N)
- V případě, že řekneme, že gen je pro nás významný již na 10% FDR, výsledek se změní!
- Dále ztrácíme informaci tím, že redukujeme p-hodnotu na binární proměnné (významné/nevýznamné)
- Je rozdíl jestli statisticky nevýznamné geny v naší množině jsou významné na hranici významnosti a nebo vůbec ne

Metoda celého seznamu genů: *uzavřená*

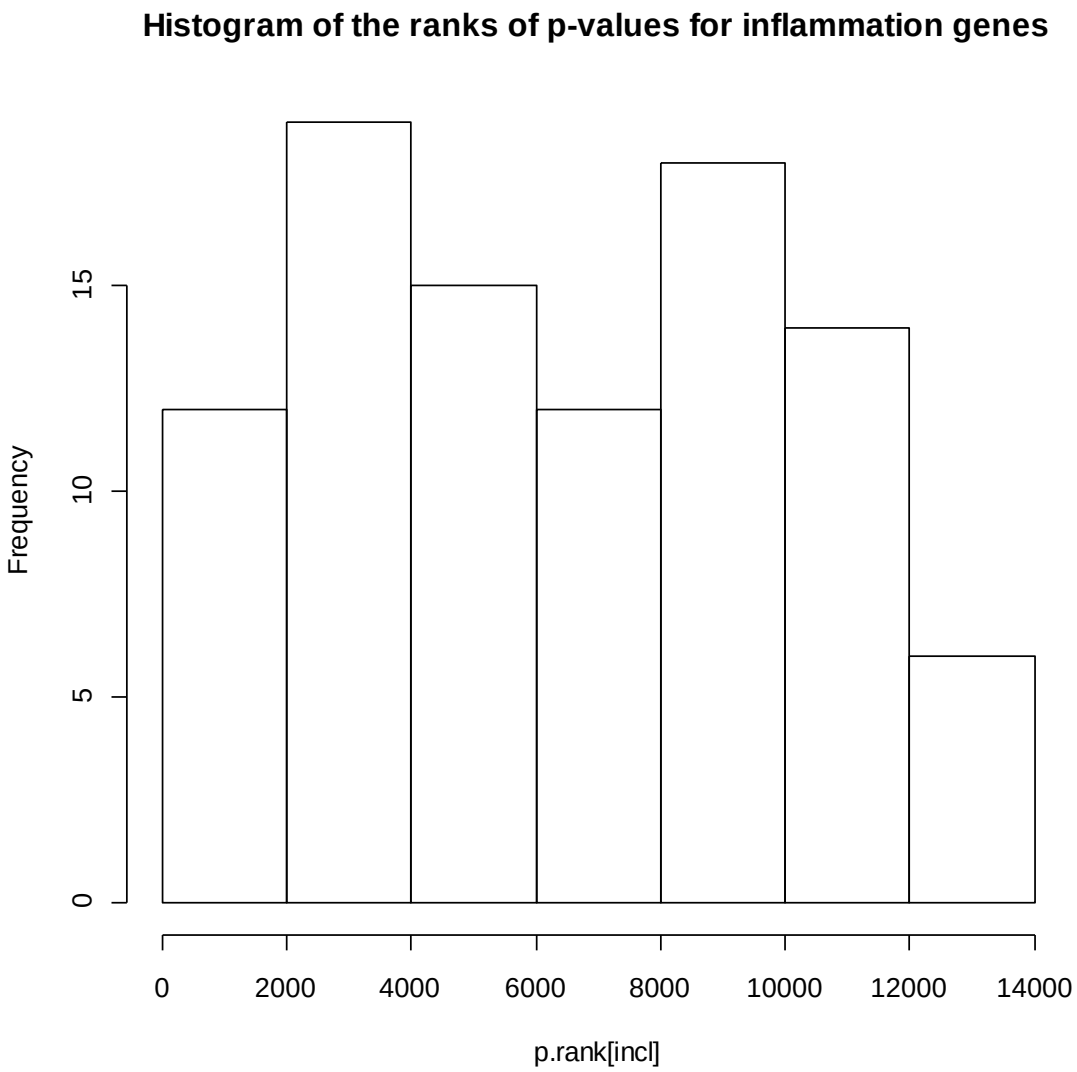
- Můžeme studovat **rozložení p-hodnot v genové sadě**
- V případě, že žádné geny nejsou odlišně exprimované (platí nulová hypotéza), měli by mít p-hodnoty **rovnoměrné** rozdělení
- Pík vlevo indikuje významnost některých genů
- Aplikujeme Kolmogorův-Smirnovův test pro porovnání rozložení
- p = 0.082, není velmi významné
- Je to **uzavřená** metoda, protože používáme jen geny z genové sady



MUNI | RECETOX

Metoda celého seznamu genů: *kompetitivní*

- Alternativně se můžeme dívat na rozložení **pořadí** p-hodnot
- Toto by byla kompetitivní metoda, protože porovnáváme naši genovou sadu s ostatními geny v experimentu
- Opět můžeme aplikovat KS test
- p=85.1%, velmi nevýznamné



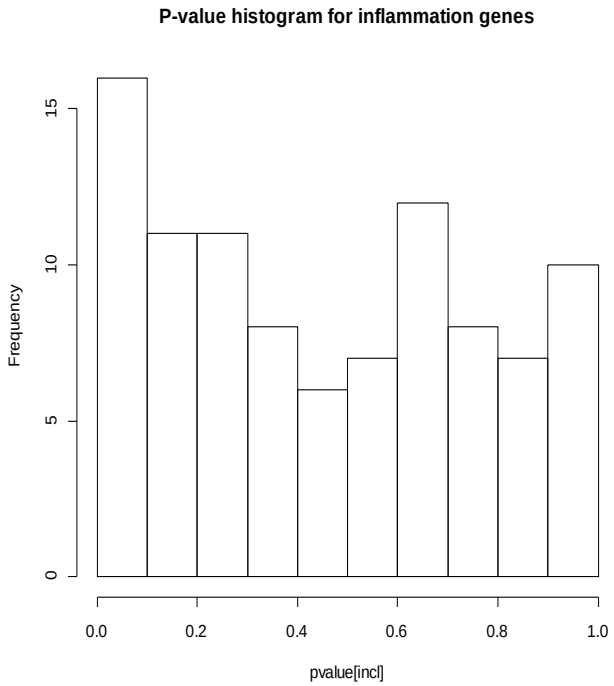
MUNI | RECETOX

1272 z 12639 genů je odlišně exprimovaných
z toho 8 v genové sadě o 96 genech

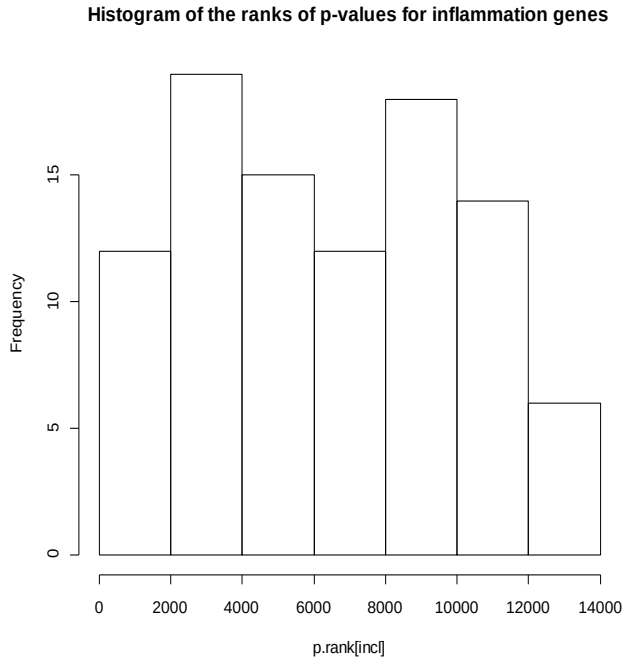
Metoda celého seznamu genů: *uzavřená*

Metoda celého seznamu genů: *kompetitivní*

	p-hodnota	pořadí
Gen A	0.001	1
Gen H	0.001	2
Gen Z	0.031	3
Gen G	0.024	4
.	.	.
Gen M	0.024	62
.	.	.
Gen O	0.049	1272
Gen J	0.351	5843
.		
Gen L	0.454	7390
.		
Gen B	0.752	10287
.		
.		
Gen C	0.989	12639



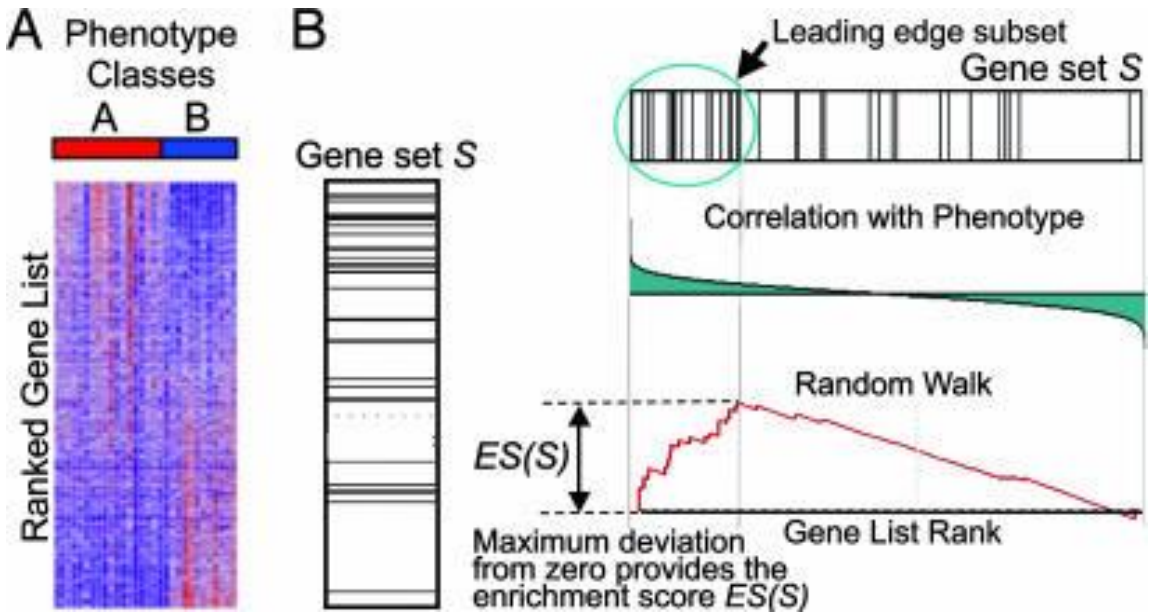
rozložení p-hodnot v genové sadě



MUNI | RECETOX

GSEA

- Najznámější je GSEA – gene set enrichment analysis (analýza obohacení genové sady)
- Jádrem je v podstatě pozměněný KS test
- Počítá se na seřazených p-hodnotách a sleduje se, zda jsou geny z genové sady náhodně rozložené v tomto seřazeném listě, a nebo se vyskytují v horních, významných pozicích
- Postup: 1. Výpočet skóre obohacení (ES)
    - 2. Odhad významnosti ES (p-hodnota) na základě permutačního testu
    - 3. Upravení p-hodnot na problém mnohonásobného porovnávání



MUNI | RECETOX

Uzavřené vs. kompetitivní II.

- Výsledky kompetitivních testů závisí na počtu testovaných genů (např. genů na mikročipu a předchozím filtrování)
  - Na malém mikročipovém sklíčku, kde jsou změněné všechny geny, kompetitivní metoda nenajde žádné odlišně exprimované množiny genů.
- Kompetitivní metody dávají méně významných výsledků než metody uzavřené

Další aspekty

Směr změny

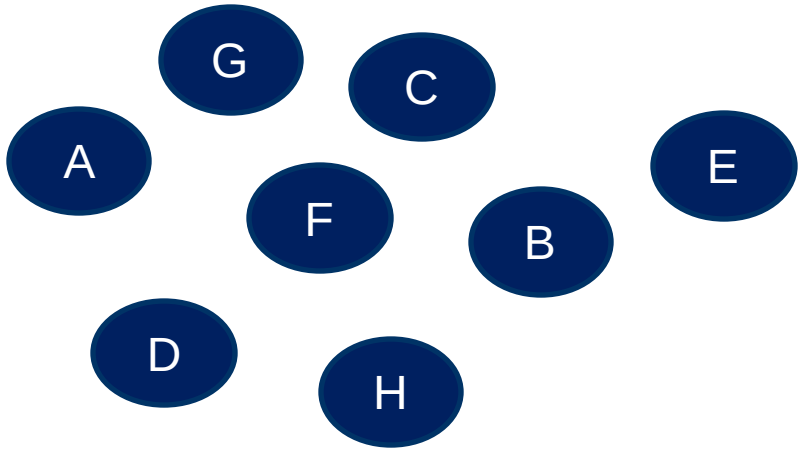
- Pokud chceme zjistit **směr** změny, musíme zopakovat analýzu pro jednostranný test
  - jen up-regulované
  - jen down-regulované

Mnohonásobné testování

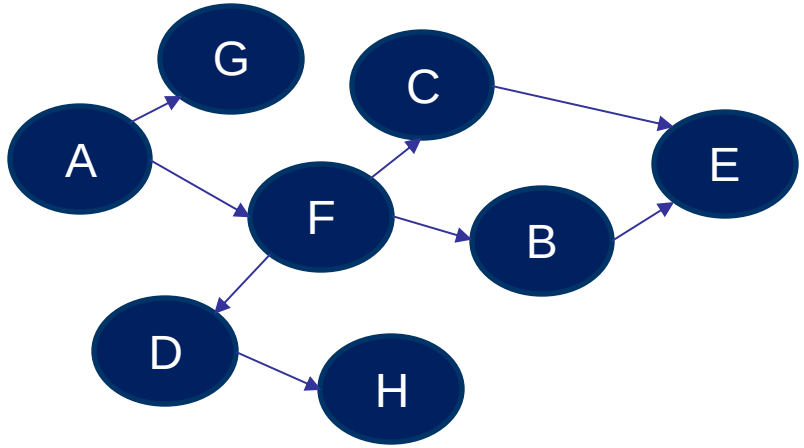
- Stejně jako u testování hypotéz na genech mezi skupinami, i pokud máme velký počet genových sad!
- FDR je trochu komplikované, protože genové množiny se překrývají
- Bonferroniho korekce vždy funguje

Topologie

Bez topologie



S topologií



Topologie využívaná různě

- Cíl:
  - změna průměrné exprese, korelace, topologie
- Jednotka zájmu:
  - dráha, modul, cesta, geny
- Topologie známá dopředu a nebo odhadovaná z dat
- Celková síť a nebo individuální dráhy

TopologyGSA, Clipper
DEGraph

Vzorky

gény

Skupina A | Skupina B

Mnohorozměrné modely:

Gaussian Graphical Models
Multivariate Normal Distribution

SPIA, PRS
PWEA

Vzorky

gény

Skupina A | Skupina B

gény

Změna exprese
t-statistika
p-hodnota

$\sum$

TAPPA

Vzorky

gény

Skupina A | Skupina B

Vzorky

dráhy

Skupina A | Skupina B

t-test

MUNI | RECETOX

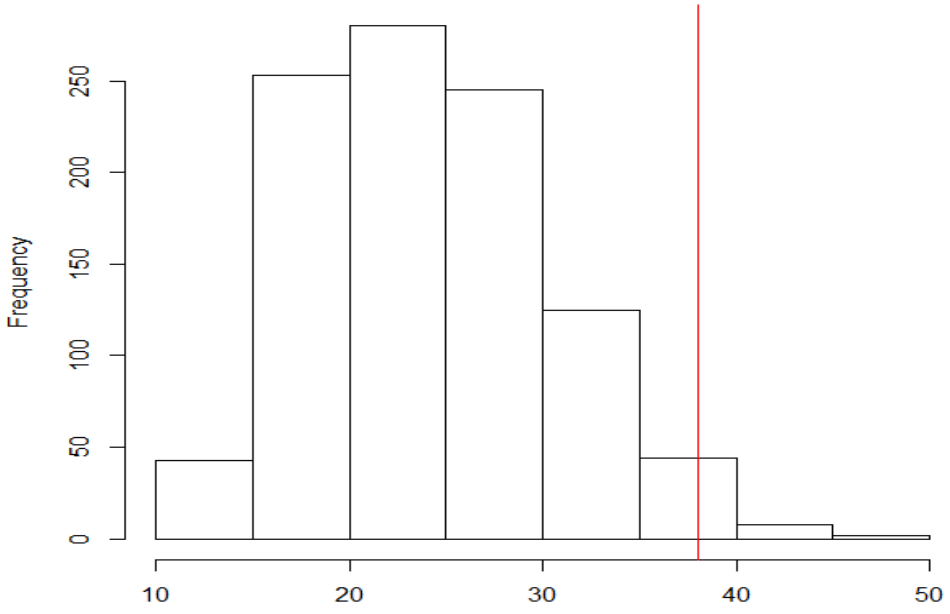
Příklad – uzavřená metoda dělící hranice

- 96 genů v dráze, z toho 8 má p-hodnoty < 5%
- Je exprese dráhy změněná?

- Využití topologické informace:
  - Definujeme statistiku
  - s = $\sum_{i=1}^{n} w_i d_i$
  - n – počet genů v dráze
  - i – index pro gen
  - w_i – počet interakcí genu *i*
  - d_i – 1 – pokud je gen *i* odlišně exprimovaný, 0 jinak

Příklad – uzavřená metoda dělící hranice

- Z 8 odlišně exprimovaných genů:
  - 2 interagují s 10 geny v dráze
  - 3 interagují s 5 geny v dráze
  - 3 interagují s jedním genem v dráze
- s = 2*10 + 3*5 + 3*1 = 38
- Opakovaně v dráze náhodně vybíráme 8 genů a získáme rozdělení statistik, které porovnáme s první statistikou.



- $p = \sum_{i=1}^{N} (s_{náhodne} \geq s_{pozorované})/N$
- N=počet náhodných výberov
- p=0.028, významné

- Z 8 odlišně exprimovaných genů:
  - 2 interagují s 10 geny v dráze
  - 3 interagují s 5 geny v dráze
  - 3 interagují s jedním genem v dráze
- s = 2*10 + 3*5 + 3*1 = 38
- Opakovaně v dráze náhodně vybíráme 8 genů a získáme rozdělení statistik, které porovnáme s první statistikou.

PLOS ONE | PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A critical comparison of topology-based pathway analysis methods

Ivana Ihnatova, Vlad Popovici, Eva Budinska

Published: January 25, 2018 • https://doi.org/10.1371/journal.pone.0191154

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191154

MUNI | RECETOX

Pozor na korelace mezi geny!

- Všechny testy, které jsme probírali předpokládají, že geny uvnitř skupin jsou nezávislé
  - To je ale velmi nepravděpodobné!

- Pokud jsou geny korelované, tak p-hodnoty jednotlivých testů (např. Fisherův test) budou nesprávné
  - Vyřešíme permutačními metodami
    - Popřehazujeme skupiny **vzorků**
    - Zopakujeme analýzu
    - Porovnáme hodnoty s pozorovanými daty

Pozor na průniky mezi dráhami

- 250 KEGG drah pro H. Sapiens
  - najčastěji zastoupené geny

PIK3CD	PIK3CG	PIK3R2	PIK3CA	MAPK3	MAPK1
70	70	70	71	78	79

Další studijní materiály a SW

- Hana Imrichová*: Možnosti propojení výsledku genomických experimentů s gene ontology online databázemi pro tvorbu metabolických sítí*, Masarykova Univerzita,2010,Bakalárska práca
- Ihnatova et al. A critical comparison of topology-based pathway analysis methods, PLoS One, 2018
- https://biodatamining.biomedcentral.com/articles/10.1186/s13040-018-0166-8 (kritické review klasických GS metod)
- R balíky: PGSEA, GSA,ToPASeq, gage, DOSE, phenoTest, limma, GOstats
- MSigDB – web http://www.broadinstitute.org/gsea/msigdb/index.jsp
- Gorilla: http://cbl-gorilla.cs.technion.ac.il/
- DAVID: https://david.ncifcrf.gov/