

## Detekce biomarkerů z omics experimentů

- Mgr. Eva Budinská, PhD
- RECETOX
- [budinska@recetox.muni.cz](mailto:budinska@recetox.muni.cz)
- Experimentální onkologie, podzim 2019



# Analýza genových sad

(pathway analýza)

# Motivace

- Geny, proteiny a další molekuly jsou navzájem propojené ve velké spleti různých signálních, metabolických a různých 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

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/>

- 

- 

- 

- 

- 

- 

- 

-

# GO databáze

## Term Lineage

[Switch to viewing term parents, siblings and children](#)

### ▼ Filter tree view ?

Filter Gene Product Counts

Data source

All  
AspGD  
CGD  
dictyBase

Species

All  
Anaplasma phagocy...  
Arabidopsis thaliana  
Bacillus anthraci...

View Options

Tree view  Full  Compact

Set filters

Remove all filters

all : all [377382 gene products]

GO:0008150 : biological\_process [270820 gene products]

GO:0050896 : response to stimulus [30457 gene products]

GO:0009605 : response to external stimulus [5585 gene products]

GO:0009611 : response to wounding [2289 gene products]

GO:0006954 : inflammatory response [1173 gene products]

GO:0002526 : acute inflammatory response [427 gene products]

**GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]**

GO:0006950 : response to stress [16147 gene products]

GO:0006952 : defense response [4501 gene products]

GO:0006954 : inflammatory response [1173 gene products]

GO:0002526 : acute inflammatory response [427 gene products]

**GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]**

GO:0009611 : response to wounding [2289 gene products]

GO:0006954 : inflammatory response [1173 gene products]

GO:0002526 : acute inflammatory response [427 gene products]

**GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]**



# KEGG pathway databáze

- 
- <http://www.genome.jp/kegg/pathway.html>
- 
- 
- 
- 
-

# KEGG

The screenshot shows a web browser window titled "Color Objects in KEGG Pathways - Mozilla Firefox". The address bar contains the URL "http://www.genome.jp/kegg/tool/color\_pathway.html". The browser's tab bar shows several open tabs, including "Google Mail", "False discovery rate - W...", "p-Value Adjustments", "computing q-values - Vy...", "storeypp4.pdf (applicati...", "almac:july2009:comparis...", and "Color Objects in KEG...".

The main content area features the KEGG logo and the title "Color Objects in KEGG Pathways". Below the title is a navigation menu with links for "KEGG2", "PATHWAY", "BRITE", "KEGG Atlas", "Search Pathway", "Color Pathway", and "Search Brite".

The "Search Pathway" section includes a dropdown menu for "Search against:" set to "Homo sapiens (human)". Below this is a text input field for "Enter objects one per line followed by bgcolor, fgcolor:". The input field contains a list of pathway IDs: 7A5, A1CF, ABAT, ABCA3, ABCC6, ABCC6P1, ABP1, ACE2, ACOT8, ACRC, and ACSF2. To the right of the input field, there are "Examples:" showing "(Reference pathway (KO)) K01803 red,blue C00118 pink" and "(Homo sapiens pathway) 7167 red,blue C00118 pink".

Below the input field is a section titled "Alternatively, enter the file name containing the data:" with a text input field and a "Procházet..." button. There are three checkboxes: "Include aliases" (checked), "Use uncolored diagrams" (unchecked), and "Display objects not found in the search" (checked). At the bottom of this section are "Exec" and "Clear" buttons.

The status bar at the bottom left of the browser window displays the word "Hotovo".

# KEGG

Search PATHWAY - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje nápověda

http://www.genome.jp/kegg-bin/color\_pathway\_object

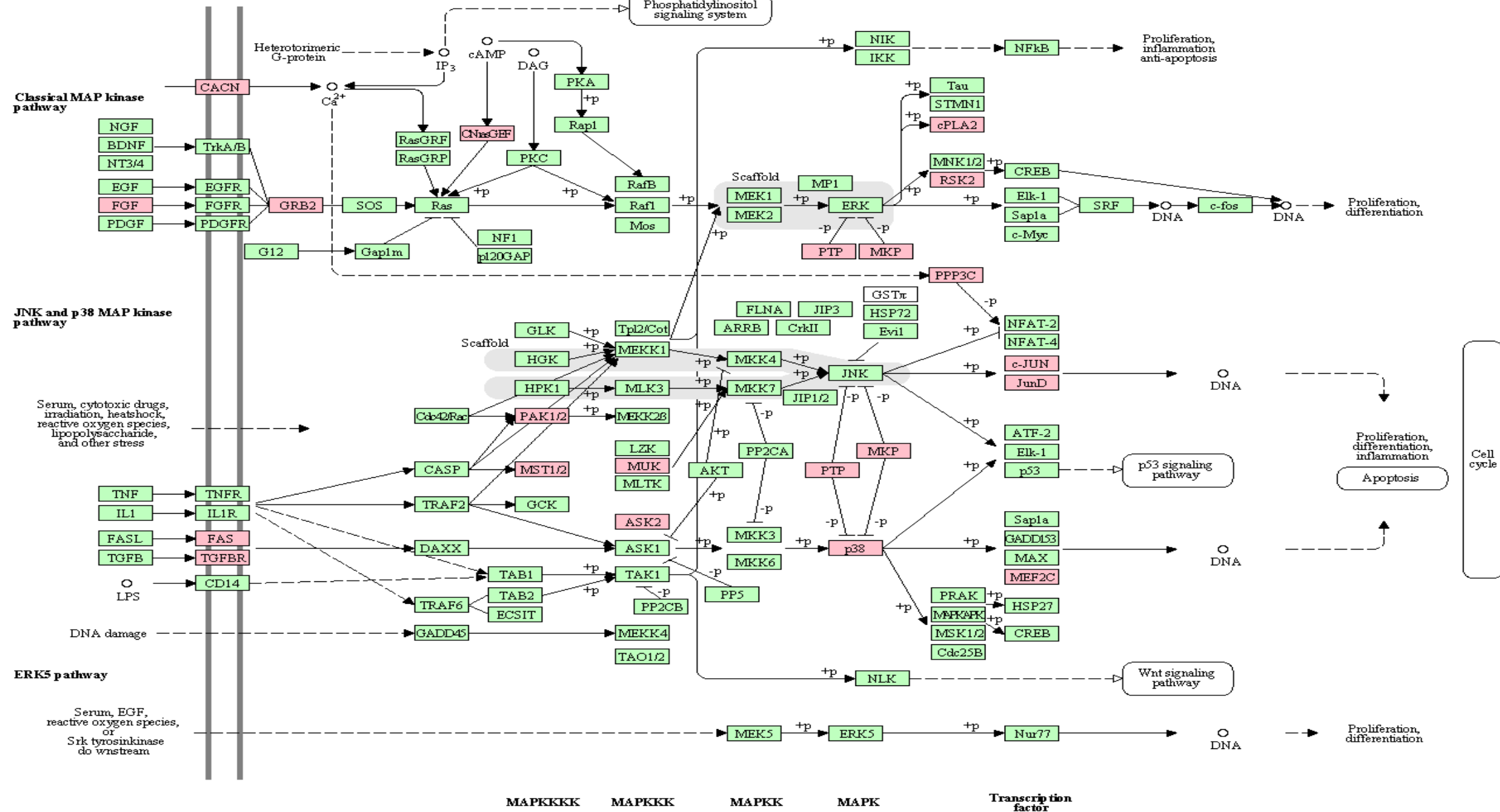
Google Mail False discovery rat... p-Value Adjustments computing q-values ... storeypp4.pdf (appl... almac:july2009:com... Google Search PATHWAY

Show all objects

- [hsa01100 Metabolic pathways - Homo sapiens \(human\)](#) (81)
- [hsa05200 Pathways in cancer - Homo sapiens \(human\)](#) (27)
- [hsa04010 MAPK signaling pathway - Homo sapiens \(human\)](#) (25)
- [hsa04060 Cytokine-cytokine receptor interaction - Homo sapiens \(human\)](#) (19)
- [hsa04062 Chemokine signaling pathway - Homo sapiens \(human\)](#) (18)
- [hsa04310 Wnt signaling pathway - Homo sapiens \(human\)](#) (17)
- [hsa00230 Purine metabolism - Homo sapiens \(human\)](#) (14)
- [hsa04660 T cell receptor signaling pathway - Homo sapiens \(human\)](#) (14)
- [hsa04020 Calcium signaling pathway - Homo sapiens \(human\)](#) (14)
- [hsa04514 Cell adhesion molecules \(CAMs\) - Homo sapiens \(human\)](#) (13)
- [hsa04510 Focal adhesion - Homo sapiens \(human\)](#) (13)
- [hsa04912 GnRH signaling pathway - Homo sapiens \(human\)](#) (12)
- [hsa04360 Axon guidance - Homo sapiens \(human\)](#) (12)
- [hsa05010 Alzheimer's disease - Homo sapiens \(human\)](#) (12)
- [hsa04650 Natural killer cell mediated cytotoxicity - Homo sapiens \(human\)](#) (12)
- [hsa04270 Vascular smooth muscle contraction - Homo sapiens \(human\)](#) (12)
- [hsa04080 Neuroactive ligand-receptor interaction - Homo sapiens \(human\)](#) (11)
- [hsa04370 VEGF signaling pathway - Homo sapiens \(human\)](#) (11)
- [hsa04630 Jak-STAT signaling pathway - Homo sapiens \(human\)](#) (11)

Hotovo

MAPK SIGNALING PATHWAY



# KEGG pathway databáze

Poklikání na jednotlivé uzly zobrazí víc  
informací o jednotlivých genech:

Všechny ostatní  
dráhy do kterých  
patří gen

Identifikátory  
daného genu v  
různých jiných  
databázích

Odkaz na literaturu  
z které byly  
informace čerpané,  
případně další  
důležité články

Informaci o sekvenci

Je možné zabarvit jednotlivé geny podle  
rozdílných barev

# Metody analýzy genových sad

# Rozdělení metod

Podle toho s jakou informací pracují na

- *metody dělicí hranice* – berou do úvahy jen informaci "významný" vs. "nevýznamný" gen
- *metody celého seznamu genů* – pracují přímo se všemi  $p$ -hodnotami (i nevýznamnými!) a teda 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 I.

Uzavřená metoda používá jen hodnoty genů z dané množiny:

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

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

- $H_0$  : “Geny v genové množině nejsou víc odlišně exprimované než ostatní geny v experimentu”

## Příklad

---

Datový soubor 12 639 genů.  
Z nich  $p < 0.05$  má 1272 genů

---

96 genů v genové sadě, z  
toho 8 má  $p$ -hodnoty  $< 5\%$

---

Kolik odlišně exprimovaných  
genů očekáváme náhodně?

1. Náhodně očekáváme  $96 \times 5\% = 4.8$  významných genů
2. Pomocí binomického testu vypočteme pravděpodobnost pozorování **8** a více významných genů:  $p = 0.1079$ , teda není významné
3. `binom.test(x=8,n=96,p=0.05, alternative="greater")`

**Příklad, uzavřená metoda  
dělicí 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  $96 \times 10\% = 9.6$  významných genů
- $p$ -hodnotu vypočítáme z kontingenční tabulky pomocí Fisherova nebo Chi-kvadrát testu

**Příklad, kompetitivní  
metoda dělicí hranice**

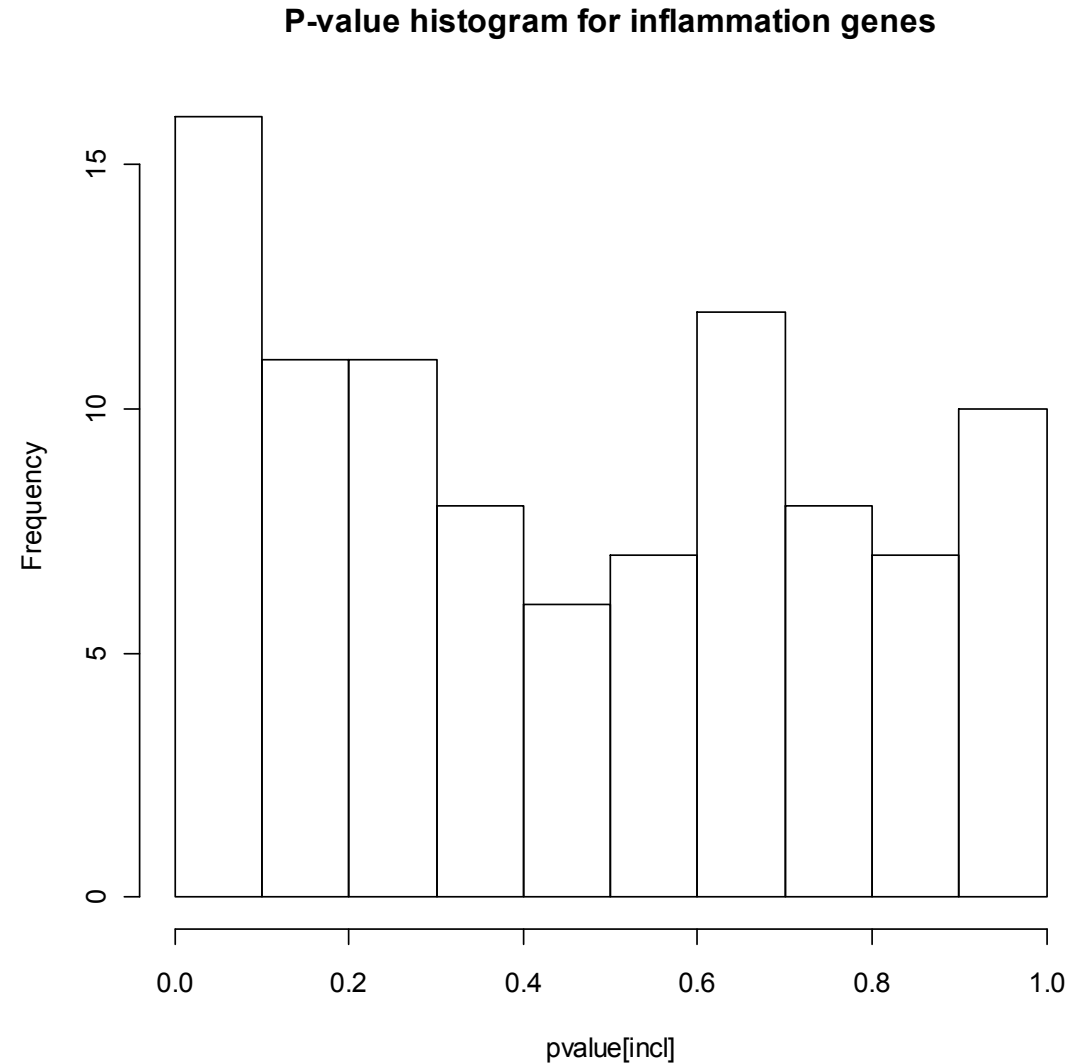
Dělení metod podle toho s  
jakou informací pracují

# Metody dělicí hranice vs. metody celého seznamu

- Dvě předchozí metody byly závislé na dělicích hranicích – cut-offs a tedy závislé na  $N$
- V případě, že řekneme, že gen je pro nás významný již na 10% FDR, výsledek se změní!
- Dále ztrácíme informaci tím, že redukuje p-hodnotu na binární proměnné (významné/nevýznamné)
- Je rozdíl vědět 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, mělo by se jednat o uniformní rozložení
- Pík vlevo indikuje významnost některých genů
- Aplikujeme Kolmogorův-Smirnovův test pro porovnání rozložení
- $p = 8.2\%$ , není velmi významné
- Je to **uzavřená** metoda, protože používáme jen geny z genové sady

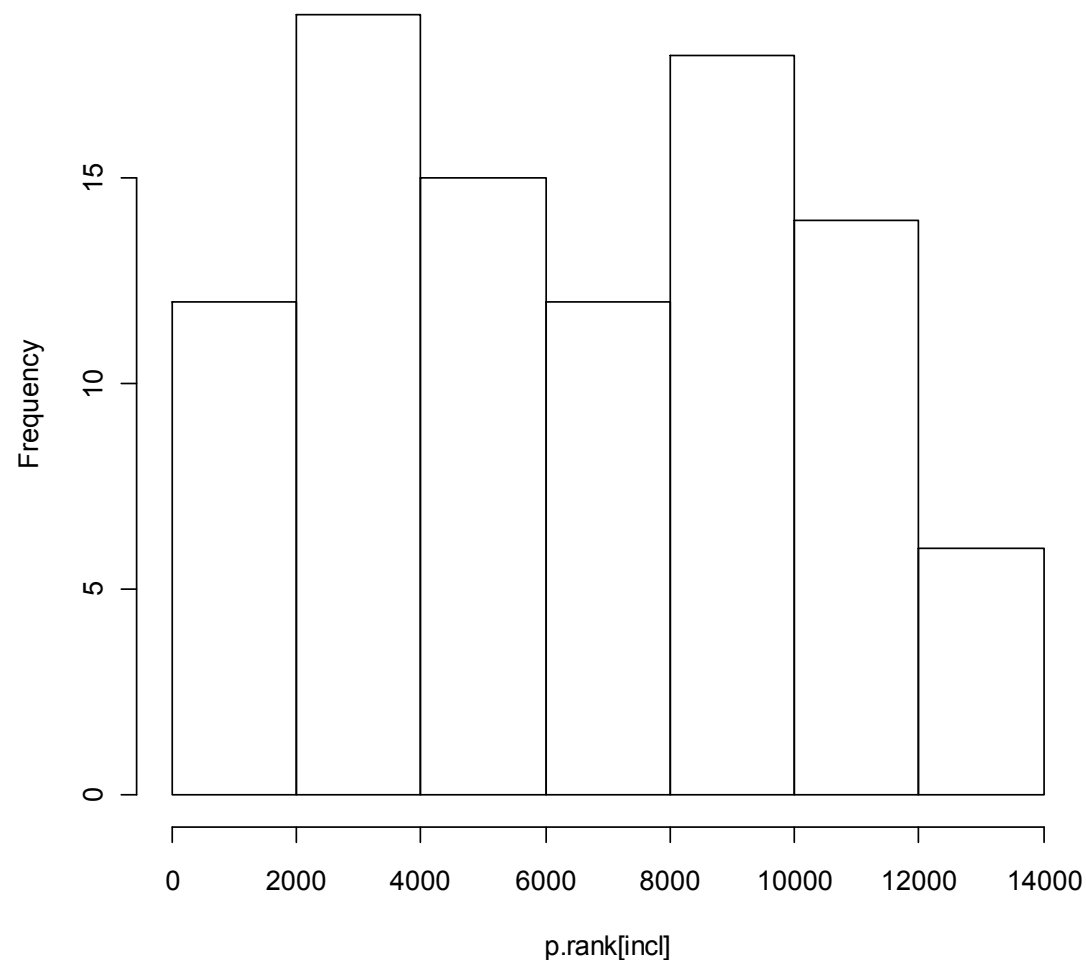




# Metoda celého seznamu genů: *kompetitivní*

- 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

Histogram of the ranks of p-values for inflammation genes

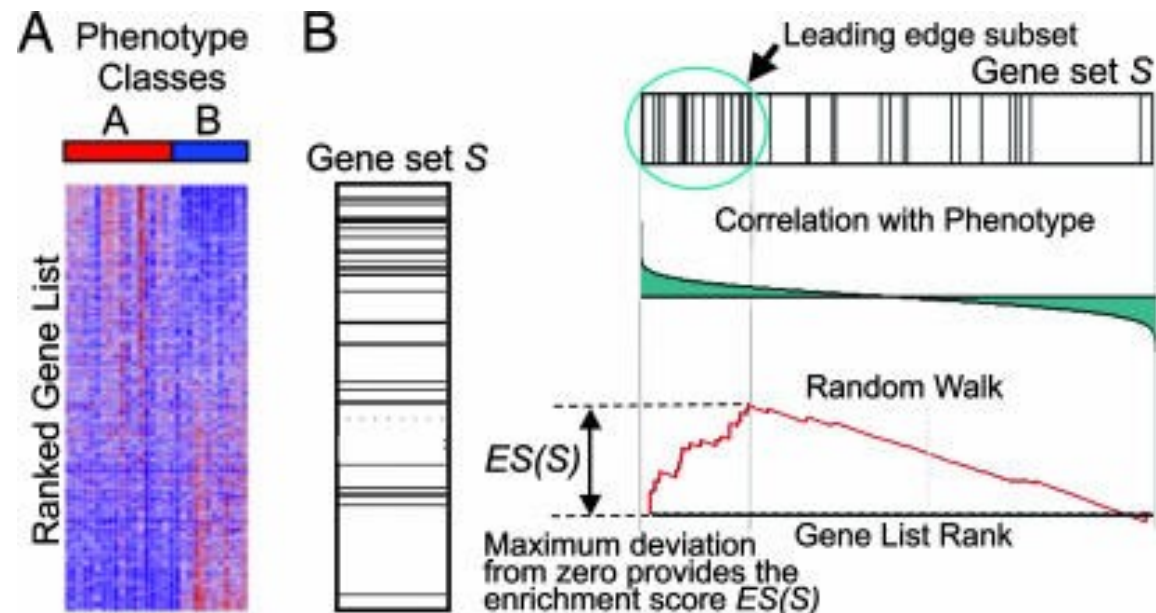


# Uzavřené vs. kompetitivní II.

- Výsledky kompetitivních testů závisí na počtu testovaných genů (např. genů na microarray sklíčku a předcházejícím 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

# Smíšené metody

- Najznámější je GSEA – gene set enrichment analysis (analýza obohacení genové sady)
- 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



# Další aspekty

## Směr změny

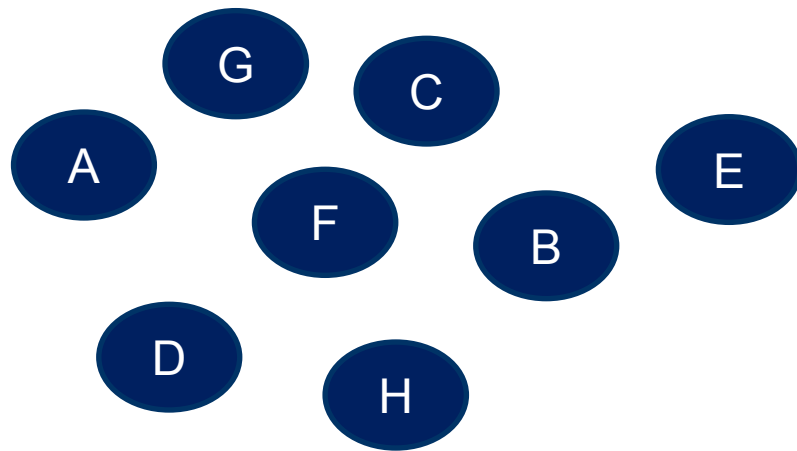
- Pokud chceme zjistit **směr** změny, musíme zopakovat analýzu pro jednostranný test
  - jen up-regulované
  - jen down-regulované

## Mnohonásobné testování

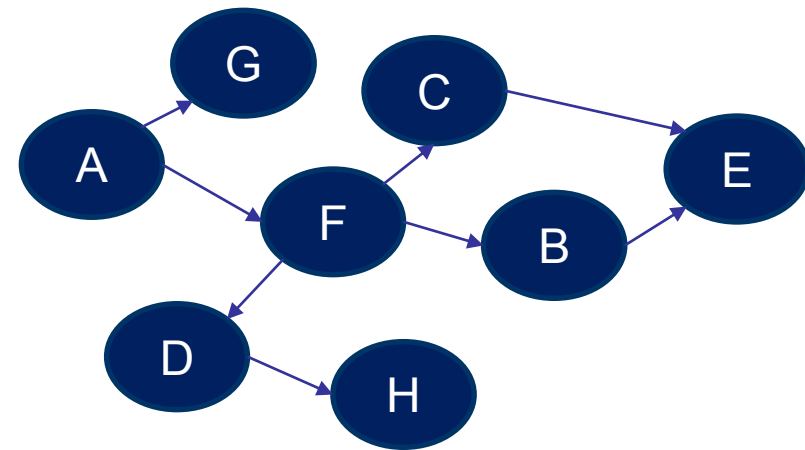
- Stejně jako u testování hypotéz na genech mezi skupinami, i pokud máme velký počet genových sad!
- FDR je trochu komplikované, protože genové množiny se překrývají
- Bonferroniho korekce vždy funguje

# 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

SPIA, PRS  
PWEA

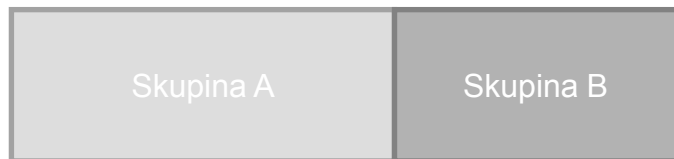
TAPPA

Vzorky

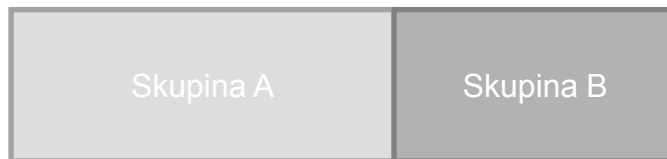
Vzorky

Vzorky

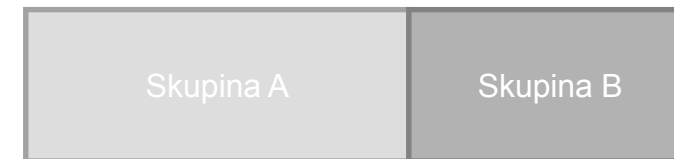
gény



gény



gény



Mnohorozměrné modely:

Gaussian Graphical Models  
Multivariate Normal Distribution

gény

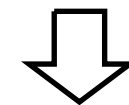


Změna exprese  
t-statistika  
p-hodnota

dráhy



$\Sigma$



t-test

Topologie dráhy

# Příklad – uzavřená metoda dělicí hranice

---

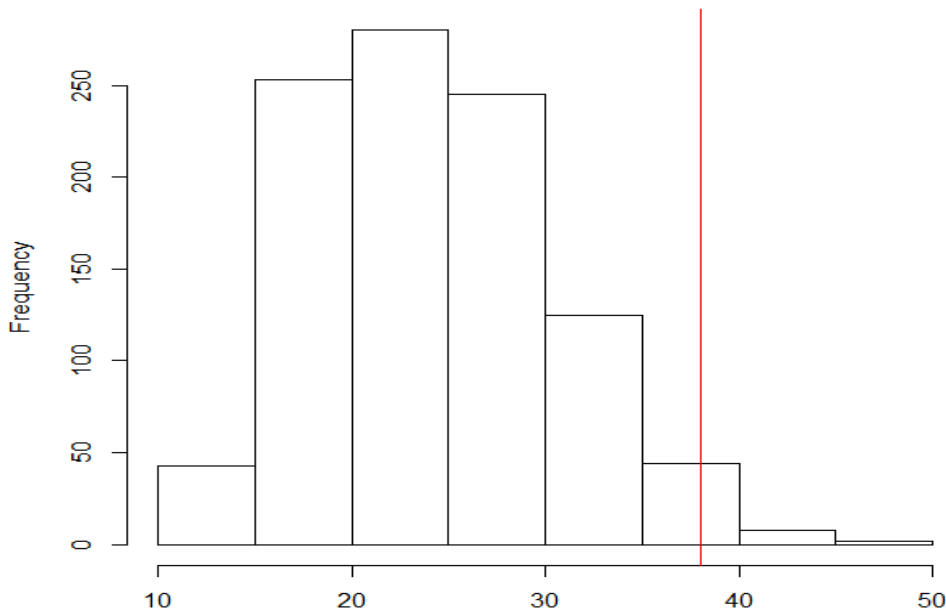
- 96 genů v dráze, z toho 8 má p-hodnoty < 5%
- Je exprese dráhy změněná?
- Využití topologické informace:
  - Definujeme statistiku
  - $s = \sum_{i=1}^n w_i d_i$
  - n – počet genů v dráze
  - i – index pro gen
  - $w_i$  – počet interakcí genu  $i$
  - $d_i$  – 1 – pokud je gen  $i$  odlišně exprimovaný, 0 jinak



# Příklad – uzavřená metoda dělicí 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.



- $p = \sum_{i=1}^N (s_{náhodné} \geq s_{pozorované}) / N$
- 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
- $s = 2 \cdot 10 + 3 \cdot 5 + 3 \cdot 1 = 38$
- Opakovaně v dráze náhodně vybíráme 8 genů a získáme rozdělení statistik, které porovnáme s první statistikou.

 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>

# 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
  - nejčastěji 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, Bakalářská práce
- Ihnatova et al. A critical comparison of topology-based pathway analysis methods, PLoS One, 2018
- 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/>