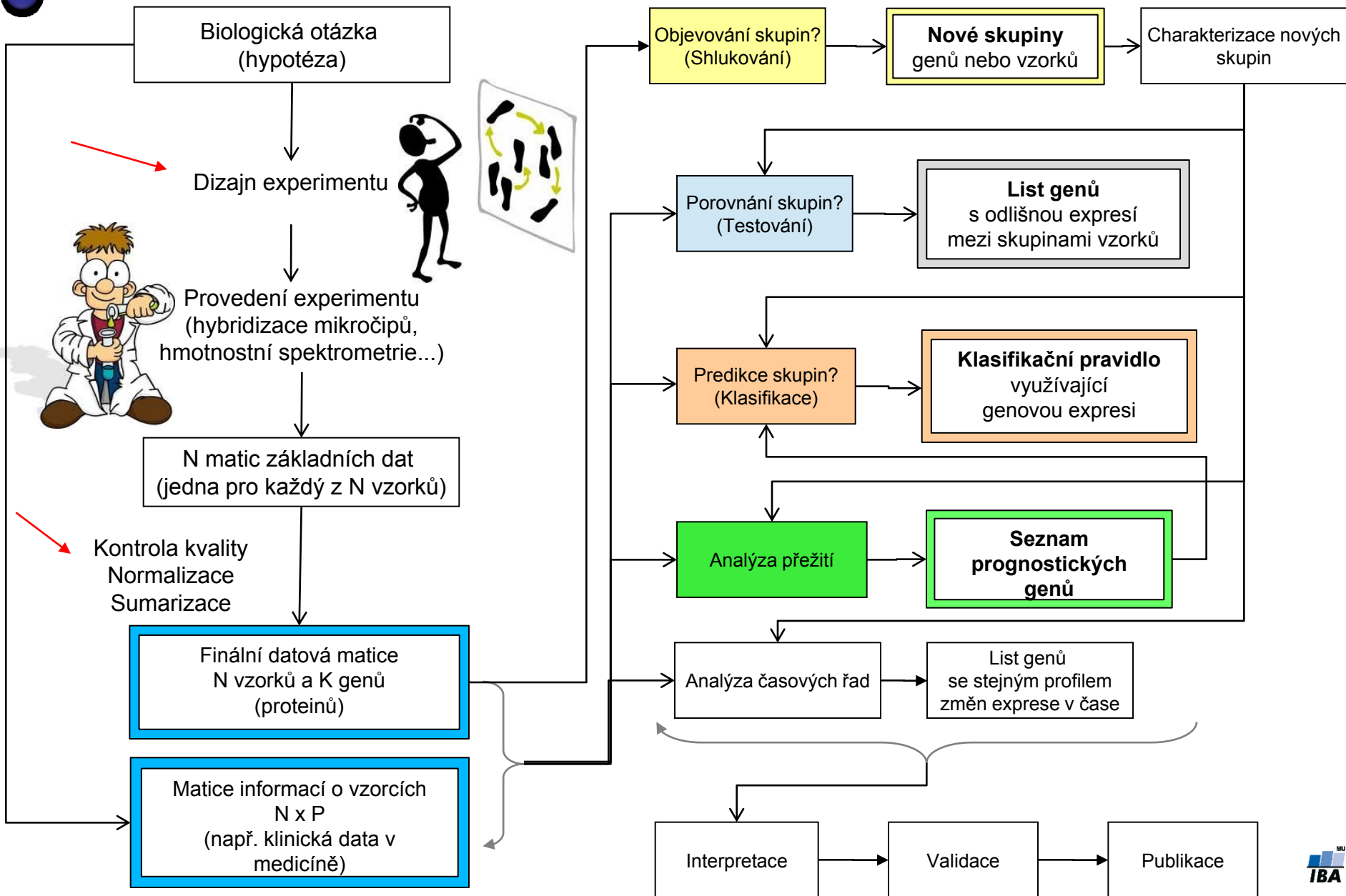


Kapitola VI

Analýza genových sad (pathway analýza)

Společné schéma analýzy dat



Motivace

- Geny a proteiny jsou navzájem propojené ve velké spleti různých signálních, metabolických a různých jiných drah
- Jak odhalit tyto závislosti?
 1. 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
 2. Můžeme přímo porovnávat všechny geny se skupinami genů v jednotlivých dráhách
- Předpoklad těchto analýz: operují s už definovanými skupinami genů jednotlivých drah

Genová sada vs dráha

- Genová sada je jakákoliv množina genů, například
 - všechny geny patřící do jedné dráhy
 - všechny geny které mají podobnou funkci
 - ...
- Sada genů není dráha – je to všeobecnější a méně specifický pojem

Analýza drah/genových sad

- Cíl je přiřadit každé genové sadě, případně dráze jedno číslo - skóre, a nebo p-hodnotu, abychom mohli odpovědět na otázku
 - Kolik genů je v dráze odlišně exprimovaných a je to dostatečně statisticky významné, abychom mohli říct, že je tato dráha specifická jen pro naše porovnávané skupiny?
- Osnova:
 1. Kde hledat informace o drahách/genových sadách
 2. Všeobecné rozdíly mezi nástroji pro analýzu genových sad
 3. Některé z metod popíšeme detailněji

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 geny/proteiny
 - Na vrcholu jsou 3 rodičovské uzly:
 1. Biologické procesy
 2. Molekulární funkce
 3. Buněčné složky

Gene Ontology

Term Lineage

Switch to viewing term parents, siblings and children

▼ Filter tree view ?

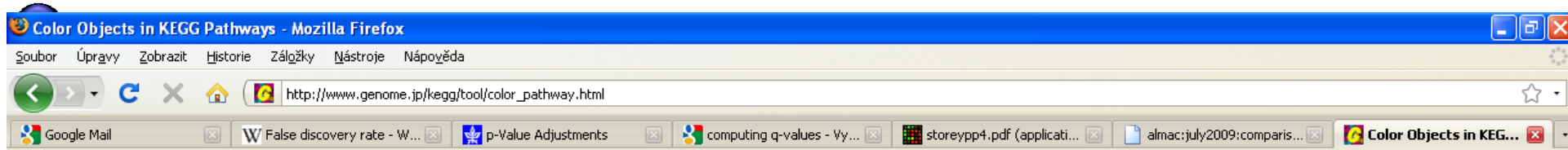
Filter Gene Product Counts	View Options	Buttons										
<table border="1"><thead><tr><th>Data source</th><th>Species</th></tr></thead><tbody><tr><td>All</td><td>All</td></tr><tr><td>AspGD</td><td>Anaplasma phagocy...</td></tr><tr><td>CGD</td><td>Arabidopsis thaliana</td></tr><tr><td>dictyBase</td><td>Bacillus anthraci...</td></tr></tbody></table>	Data source	Species	All	All	AspGD	Anaplasma phagocy...	CGD	Arabidopsis thaliana	dictyBase	Bacillus anthraci...	Tree view <input checked="" type="radio"/> Full <input type="radio"/> Compact	<input type="button" value="Set filters"/> <input type="button" value="Remove all filters"/>
Data source	Species											
All	All											
AspGD	Anaplasma phagocy...											
CGD	Arabidopsis thaliana											
dictyBase	Bacillus anthraci...											

- ▣ 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

- KEGG = Kyoto Encyclopedia of Genes and Genomes
 - <http://www.genome.jp/kegg/pathway.html>
 - Více informací než GO, máme tu už vztahy mezi geny a genovými produkty
 - Detailní informace 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

KEGG



Color Objects in KEGG Pathways

[KEGG2](#) [PATHWAY](#) [BRITE](#) [KEGG Atlas](#) [Search Pathway](#) [Color Pathway](#) [Search Brite](#)

Search against:

Enter objects one per line followed by bgcolor, fgcolor:

```
7A5
A1CF
ABAT
ABCA3
ABCC6
ABCC6P1
ABP1
ACE2
ACOT8
ACRC
ACSF2
```

Examples:

```
(Reference pathway (KO))
K01803 red,blue
C00118 pink
```

```
(Homo sapiens pathway)
7167 red,blue
C00118 pink
```

Alternatively, enter the file name containing the data:

- Include aliases
- Use uncolored diagrams
- Display objects not found in the search

Hotovo

Search PATHWAY - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje nápověda

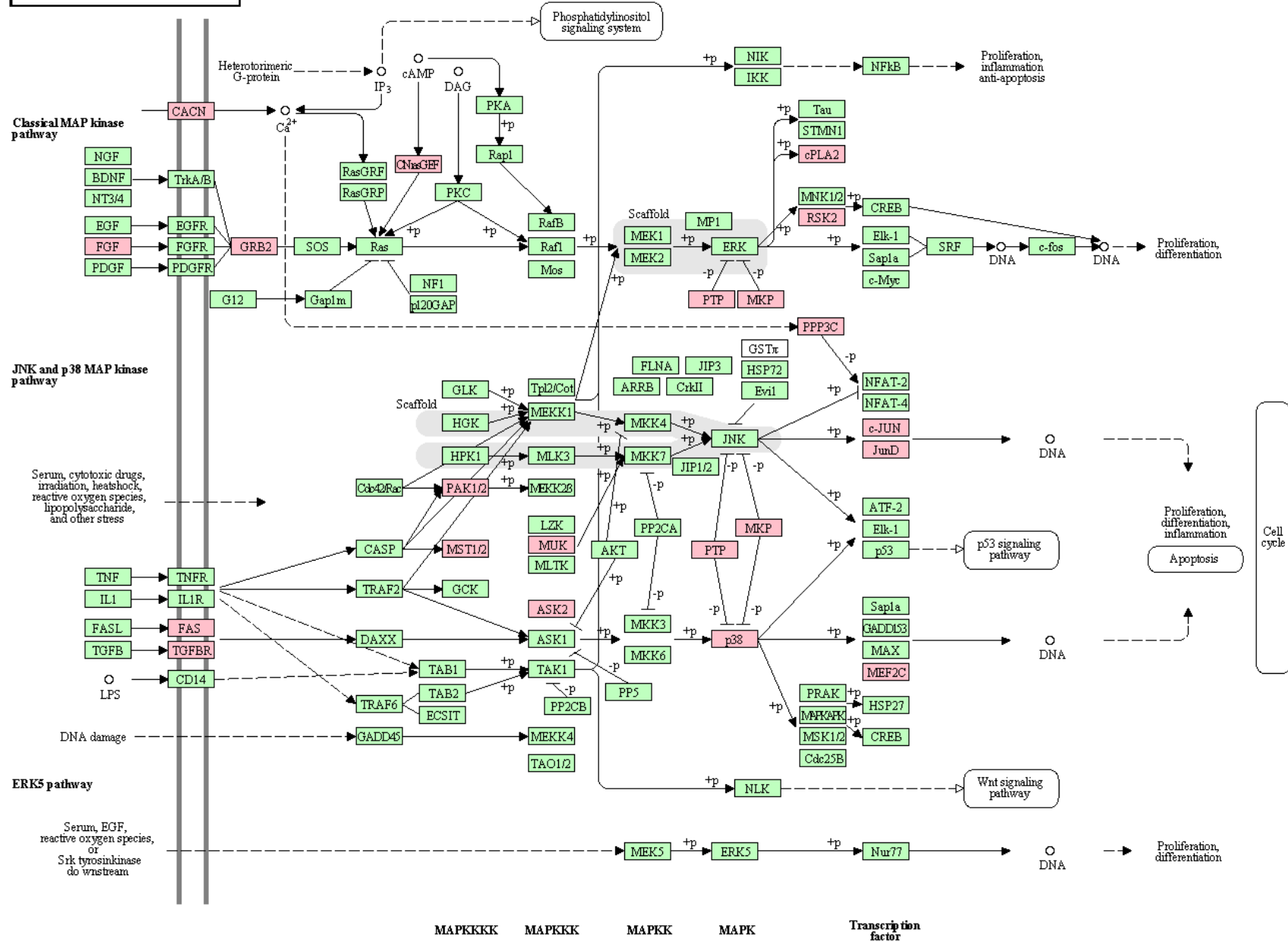
http://www.genome.jp/kegg-bin/color_pathway_object

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)

MAPK SIGNALING PATHWAY



MAPKKKK MAPKKK MAPKK MAPK Transcription factor

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

Nástroje pro analýzu genových sad

- Podle toho s jakou informací pracují na
 - *metody dělicí hranice* – berou do úvahy jen informáci "významný" vs. "nevýznamný" gen
 - *metody celého seznamu genů* – pracují přímo se všemi p -hodnotami (i nevýznamnými!) a teda s pořadím
- Nové metody pracují i s topologií dráhy
- Rozdělujeme podle skupiny genů které analyzují na:
 - *uzavřené* – analýza jen v rámci genů v sadě
 - *kompetitivní* – porovnání se všemi geny experimentu

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?

Příklad, uzavřená metoda dělicí hranice

- 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?

- Uzavřená metoda
 - Náhodně očekáváme $96 \times 5\% = 4.8$ významných genů
 - Pomocí binomického testu vypočítáme 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, kompetitivní metoda dělicí hranice

- 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?
- Kompetitivní test
 - 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

	V GS	Není v GS
Význ	8	1264
Nevýzn	88	11279

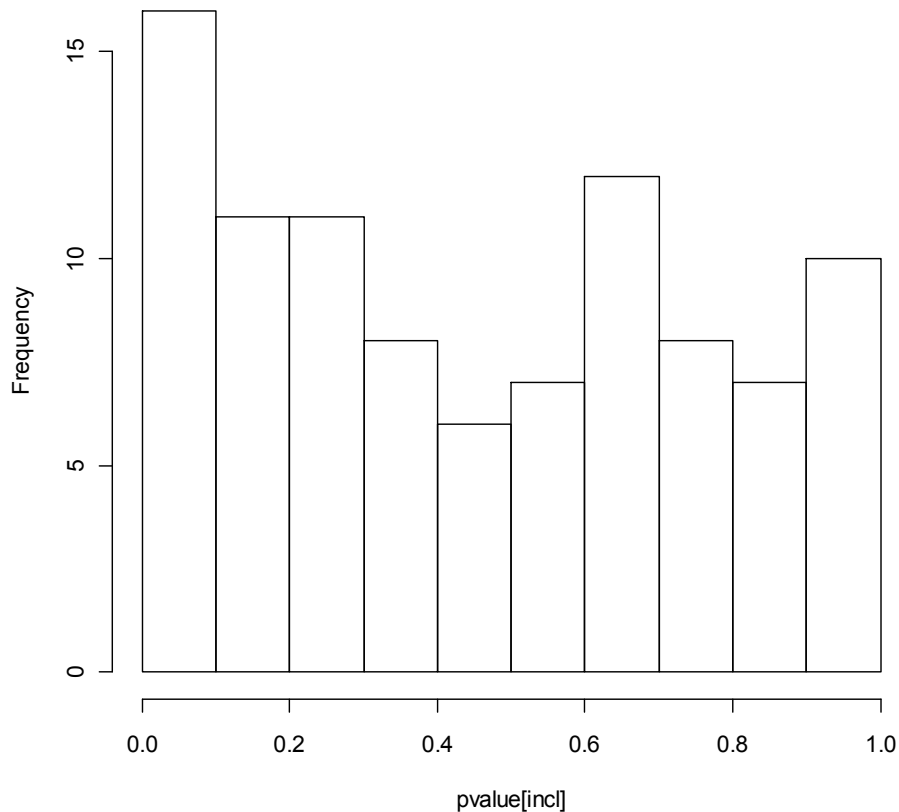
$p = 0.73$ (Fisherův test – jednostranný)

Metody dělicí hranice vs. metody celého seznamu

- Dvě předcházející metody jsou závislé na dělicích hranicích – cut-offs a tedy závislé na N
- V případě, že povíme, že gen je pro nás významný už 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*

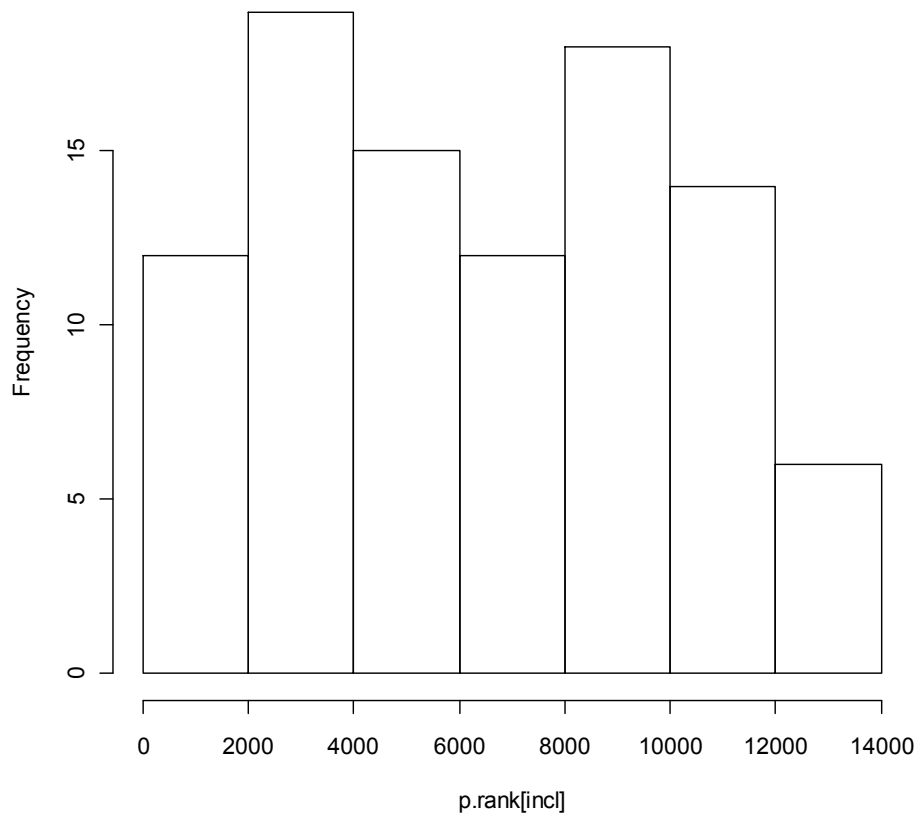
P-value histogram for inflammation genes



- 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*

Histogram of the ranks of p-values for inflammation genes



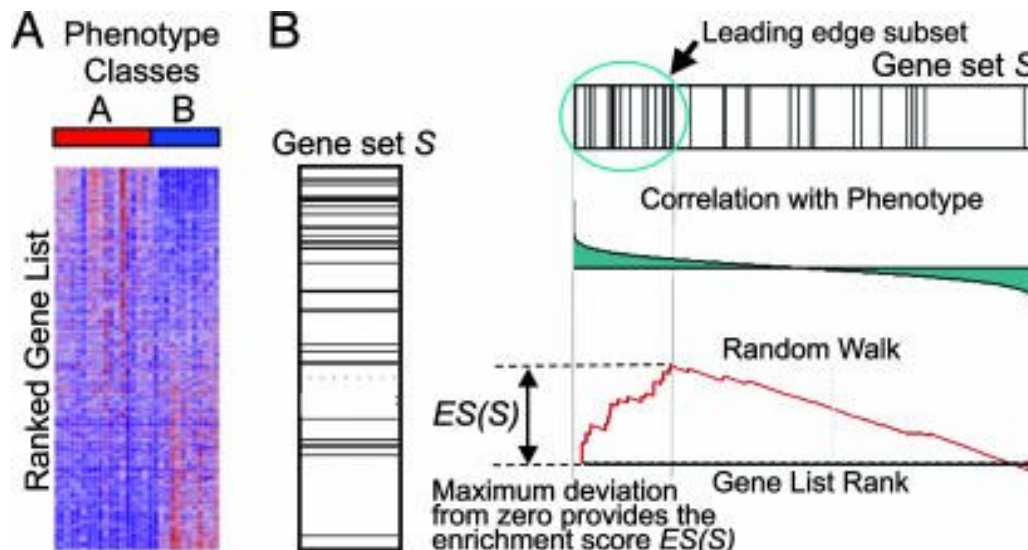
- Alternativně se můžeme dívat na rozložení **pořadí** p-hodnot
- Toto by byla kompetitivní metoda, protože porovnáváme naši genovou sadu s ostatními geny v experimentu
- Opět můžeme aplikovat KS test
- $p=85.1\%$, velmi nevýznamné

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 změřeny 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

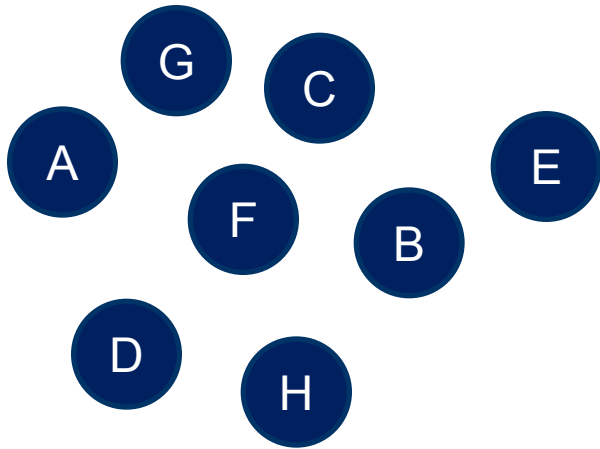


Další aspekty

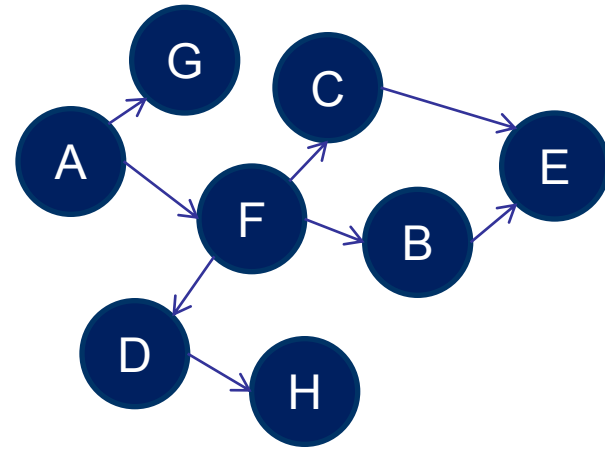
- Směr změny
 - Pokud chceme zjistit směr změny, musíme zopakovat analýzu pro jednostranný test
 - jen up-regulované
 - jen down-regulované
- Mnohonásobné testování
 - Stejně jako u testování hypotéz na genech mezi skupinami, i pokud máme velký počet genových sad!
 - FDR je trochu komplikované, protože genové množiny se překrývají
 - Bonferroniho korekce vždy funguje

S topologickou informací vs. bez

Bez topologie



S topologií



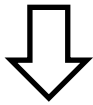
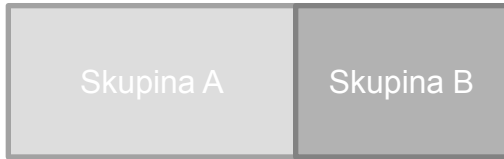
TopologyGSA, Clipper
DEGraph

SPIA, PRS
PWEA

TAPPA

Vzorky

gény

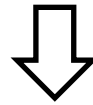
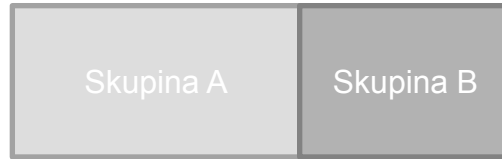


Mnohorozměrné modely:

Gaussian Graphical Models
Multivariate Normal Distribution

Vzorky

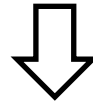
gény



gény



Změna exprese
t-statistika
p-hodnota



Σ

Vzorky

gény



Vzorky

dráhy



t-test

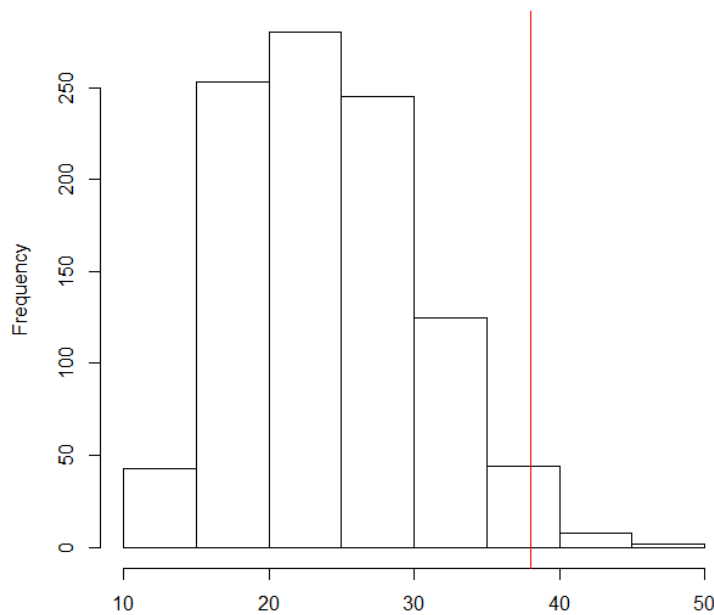
Topologie dráhy

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

- 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.



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

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

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

Studijní materiál a SW

- Hana Imrichová: *Možnosti propojení výsledku genomických experimentů s gene ontology online databázemi pro tvorbu metabolických sítí*, Masarykova Univerzita, 2010, Bakalárska práca

- R balíky

```
source("http://www.bioconductor.org/biocLite.R")
```

```
biocLite("PGSEA")
```

```
biocLite("GSA") # http://statweb.stanford.edu/~tibs/GSA/
```

```
biocLite("ToPASeq")
```

```
gage, DOSE, phenoTest, limma, GOstats
```

- MSigDB - web

```
http://www.broadinstitute.org/gsea/msigdb/index.jsp
```

```
http://cbl-gorilla.cs.technion.ac.il/
```

```
https://david.ncifcrf.gov/
```