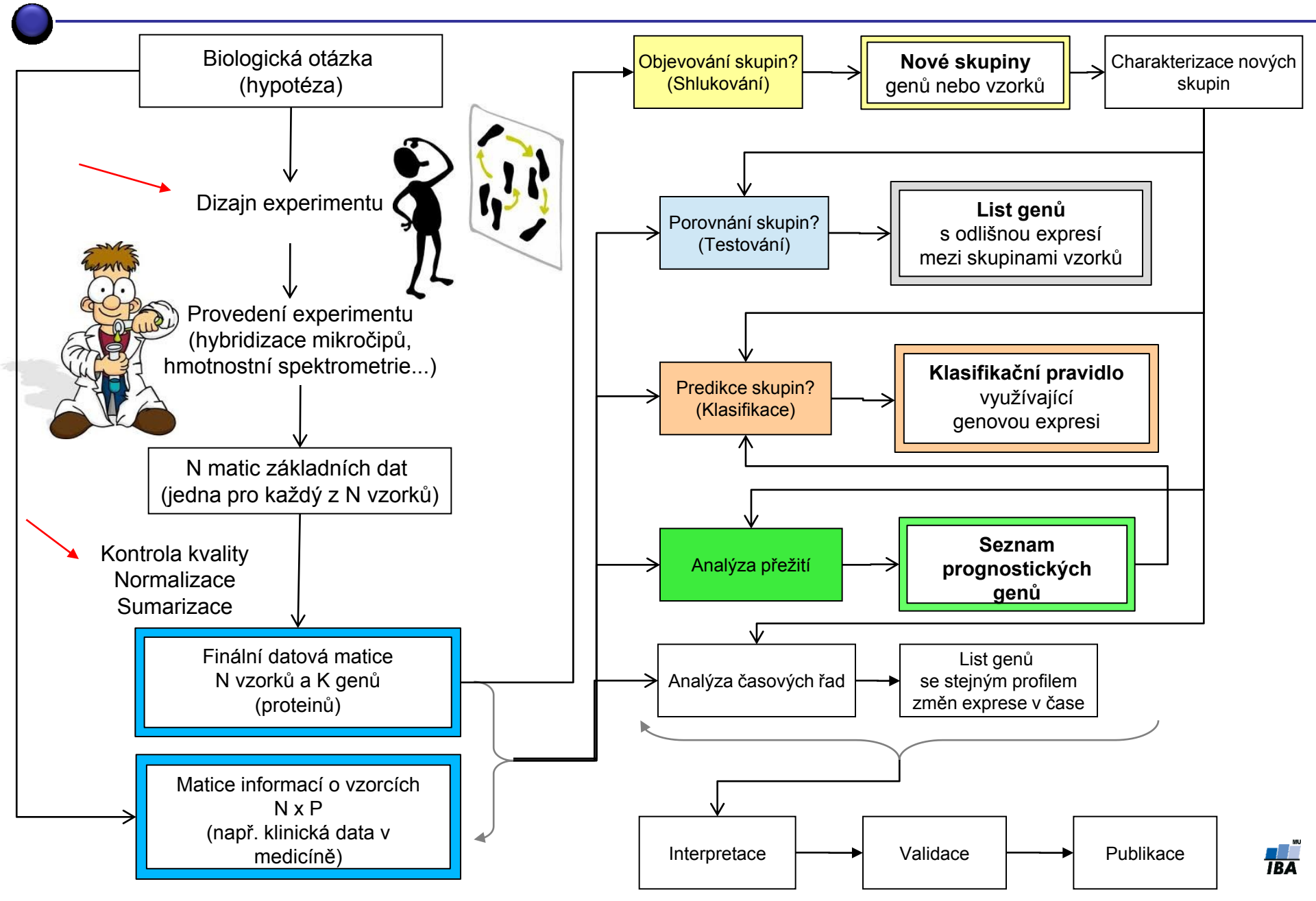
Kapitola VI

Analýza génových sád (pathway analýza)

Společná schéma analýzy dat



Motivácia

- Gény a proteíny sú navzájom prepojené vo veľkej spleti rôznych signálnych, metabolických a rôznych iných dráh
- Ako odhaliť tieto závislosti?
  1. Gény, ktoré nájdeme odlišne exprimované medzi skupinami (porovnanie skupín) môžeme ad-hoc vložiť do databázy a pozrieť sa kam patria (KEGG, MsigDB....)
     - nevýhoda – nemáme štatistickú významnosť, ktorá z dráh je zastúpená najviac
  2. Môžeme priamo porovnávať všetky gény so skupinami génov v jednotlivých dráhach
- Predpoklad týchto analýz: operujú s už definovanými skupinami génov jednotlivých dráhach

Génová sada vs dráha

- Génový sada je akákoľvek množina génov, napríklad
  - všetky gény patriace do jednej dráhy
  - všetky gény ktoré majú podobnú funkciu
  - ...

- Sada génov nie je dráha – je to všeobecnejší a menej špecifický pojem

Analýza dráh/génových sád

- Cieľ je priradiť každej génovej sade, prípadne dráhe jedno číslo - skóre, alebo p-hodnotu, aby sme mohli odpovedať na otázku
  - Koľko génov v dráhe je odlišne exprimovaných a je to dostatočne štatisticky významné, aby sme mohli povedať, že táto dráha je špecifická pre naše porovnávané skupiny?
- Osnova:
  1. Kde hľadať informácie o dráhach / génových sadách
  2. Všeobecné rozdiely medzi nástrojmi pre analýzu génových sád
  3. Niektoré z metód popíšeme detailnejšie

Databáze génových sád/pathways

- Gene Ontology (GO) databáza
  - http://www.geneontology.org/
  - Hierarchická databáza
  - Rodičovské uzly: obecnejšie termíny
  - Potomkovia uzly: viac špecifické
  - Na konci hierarchie sú gény/proteíny
  - Na vrchole sú 3 rodičovské uzly:
    1. Biologické procesy
    2. Molekulárna funkcia
    3. Bunkové zložky

Gene Ontology

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

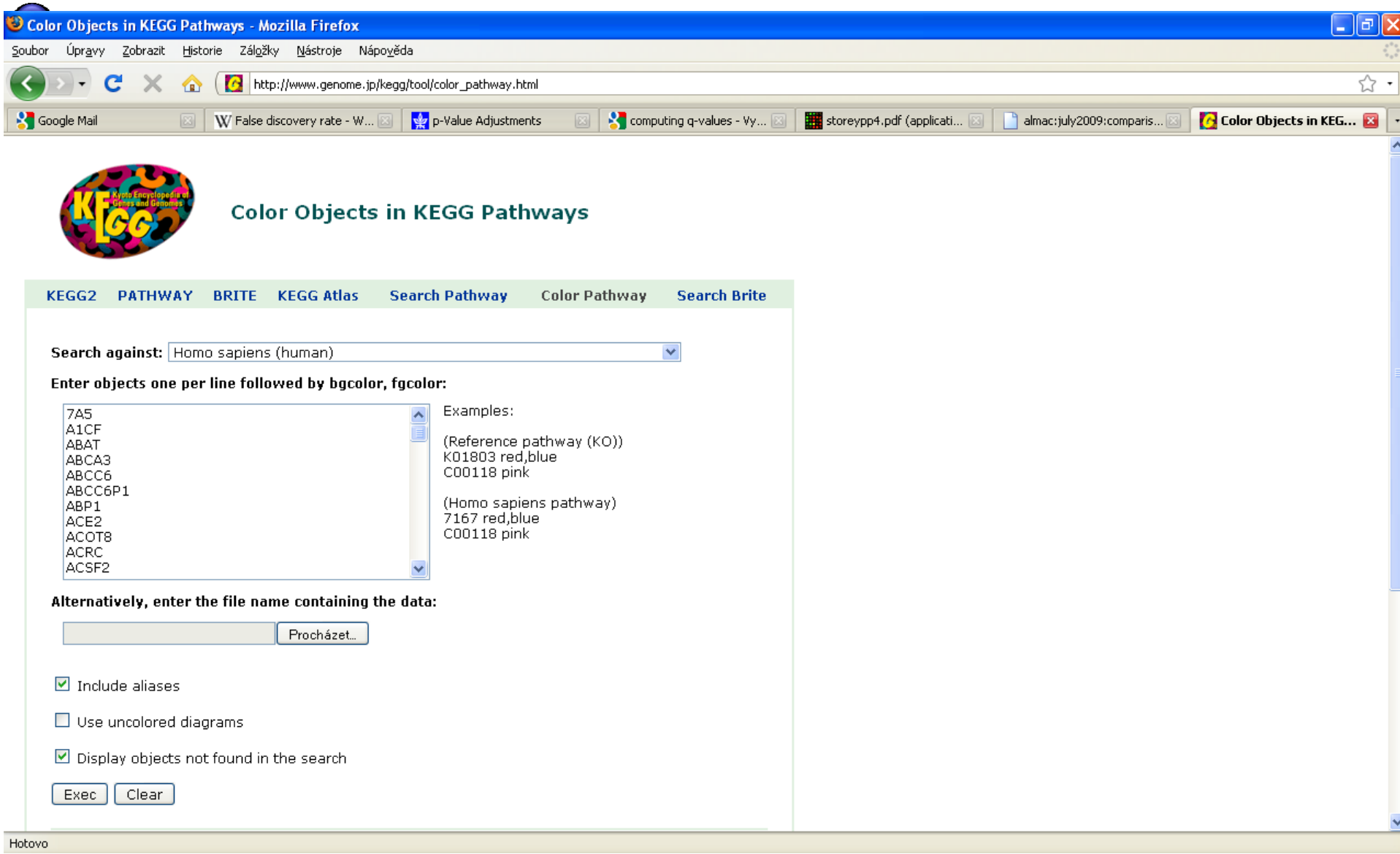
Remove all filters

- 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]

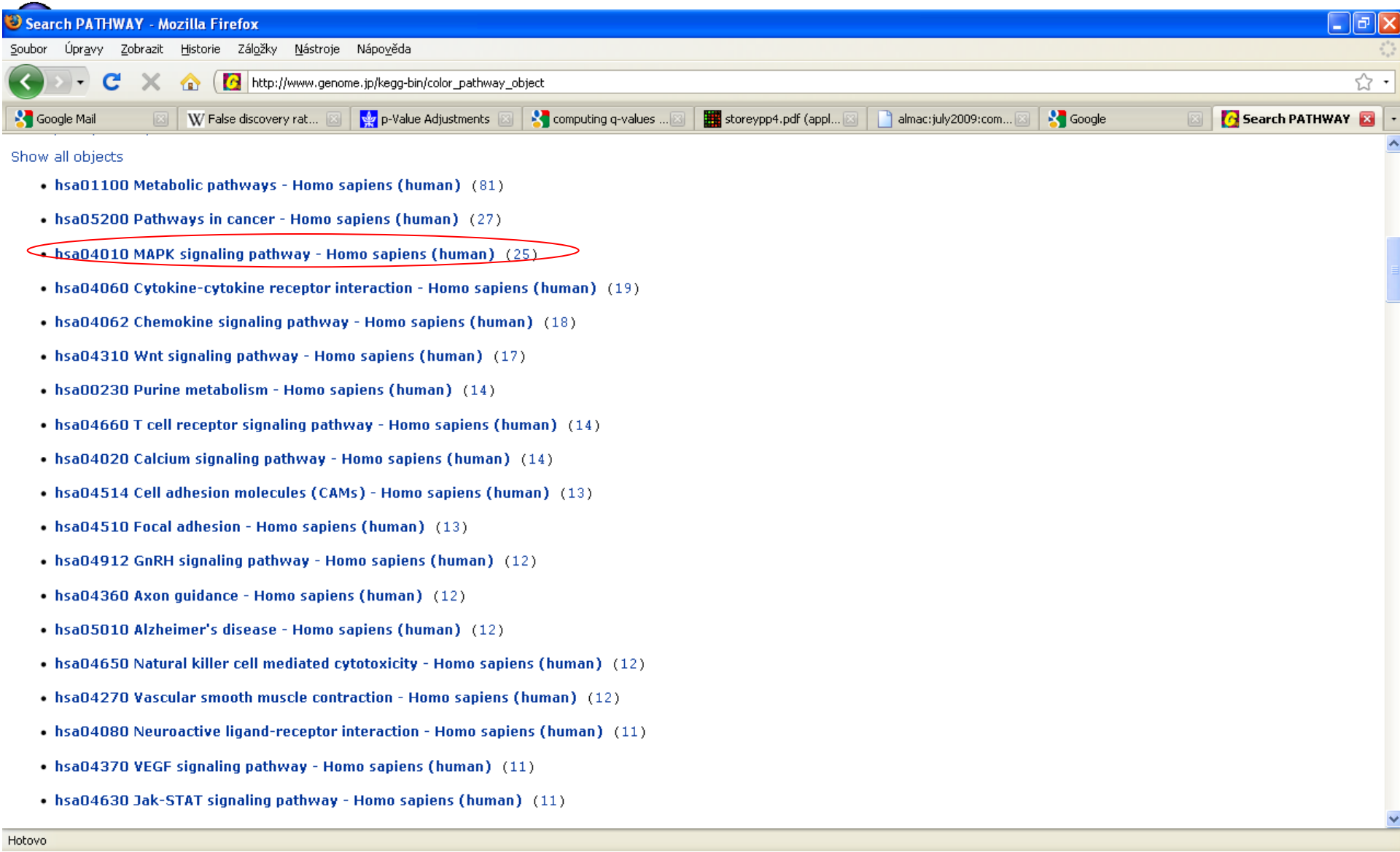
KEGG pathway databáza

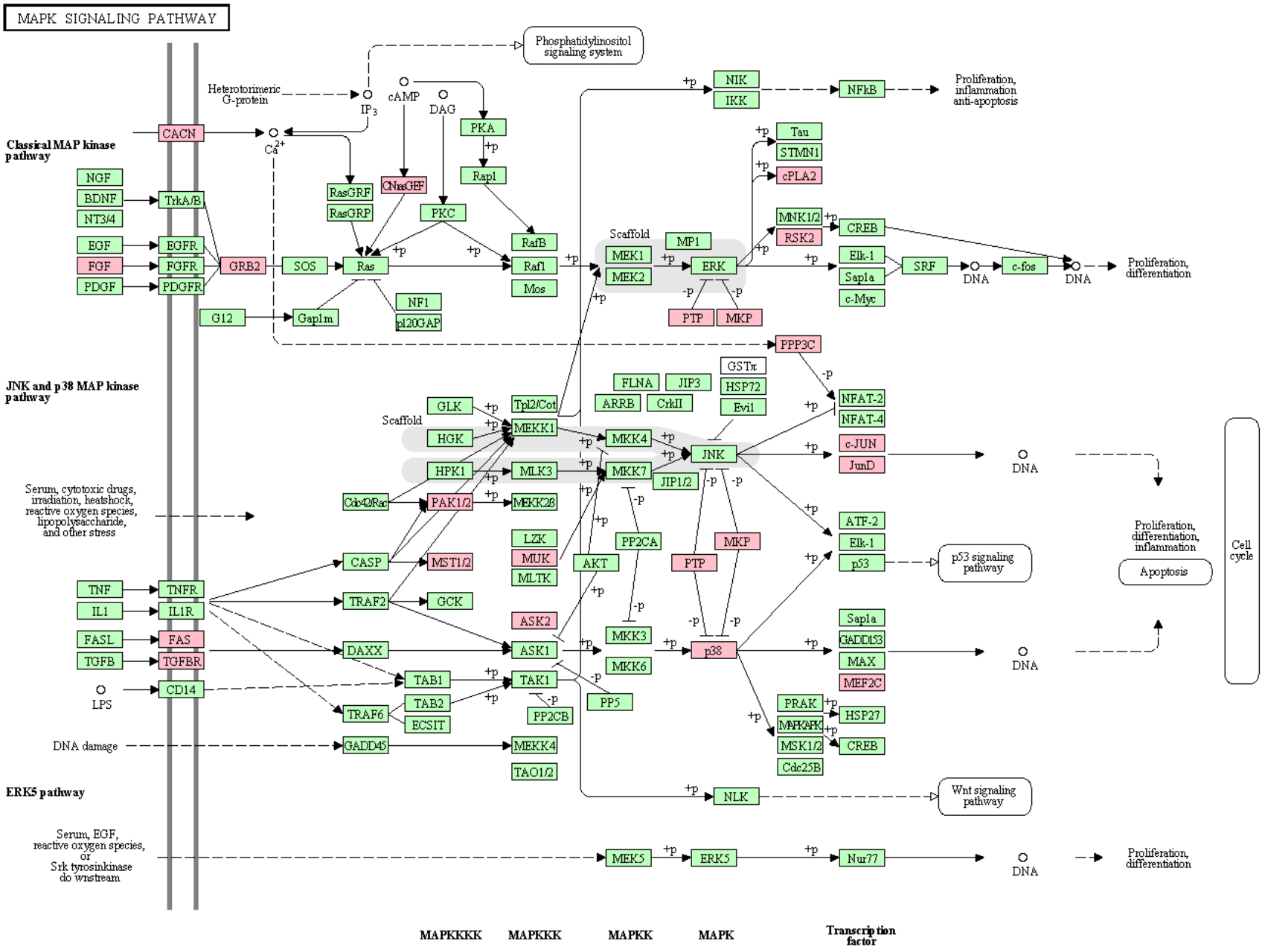
- KEGG = Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
  - Viac informácií než GO, máme tu už vzťahy medzi génmi a génovými produktami
  - Detailná informácia len pre niektoré organizmy a procesy
  - Využíva hlavne overené poznatky, nemôže ju meniť ktokoľvek
  - Preto sa tu nenachádzajú všetky gény (obvykle tak tretina až polovica z hľadaných)
  - Aktualizovaná databáza nie je voľne pristupná

KEGG


Color Objects in KEGG Pathways - Mozilla Firefox
Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda
http://www.genome.jp/kegg/tool/color_pathway.html
Google Mail
False discovery rate - W...
p-Value Adjustments
computing q-values - Vy...
storeypp4.pdf (applicati...
almac:july2009:comparis...
Color Objects in KEG...
Color Objects in KEGG Pathways
KEGG2 PATHWAY BRITE KEGG Atlas Search Pathway Color Pathway Search Brite
Search against: Homo sapiens (human)
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:
Procházet...
Include aliases
Use uncolored diagrams
Display objects not found in the search
Exec
Clear
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



KEGG pathway databáza

- Poklikanie na jednotlivé uzly zobrazí viac informácie o jednotlivých génoch:
  - Všetky ostatné dráhy do ktorých patrí gén
  - Identifikátory daného génu v rôznych iných databázach
  - Odkaz na literatúru z ktorej boli informácie čerpané, prípadne ďalšie dôležité články
  - Informáciu o sekvencii

- Je možné zafarbiť jednotlivé gény podľa rozličných farieb

Nástroje pre analýzu génových sád

- Podľa toho s akou informáciou pracujú na
  - *metódy deliacej hranice* – berú do úvahy len informáciu "významný" vs "nevýznamný" gén
  - *metódy celého zoznamu génov* – pracujú priamo so všetkými *p*-hodnotami (i nevýznamnými!) a teda s poradím

- Nové metódy pracujú aj s topológiou dráhy

- Rozdeľujeme podľa skupiny génov ktoré analyzujú na:
  - *uzavreté* – analýza len v rámci génov v sade
  - *kompetitívne* – porovnanie so všetkými génmi experimentu

Uzavreté vs kompetitívne I.

- Uzavretá metóda používa len hodnoty génov z danej množiny:
  - H_0 : “Žiadne gény z génovej množiny nie sú odlišne exprimované”

- Kompetitívny test porovnáva gény v génovej množine s ostatnými génmi v experimente
  - H_0 : “Gény v génovej množine nie sú viac odlišne exprimované než ostatné gény v experimente”

Príklad, metódy deliacej hranice

- Dátový súbor 12 639 génov. Z nich p<0.05 má 1272 génov
- 96 génov v génovej sade, z toho 8 má p-hodnoty < 5%
- Koľko odlišne exprimovaných génov očakávame náhodne?
- Uzavretá metóda
  - Náhodne očakávame 96 x 5% = 4.8 významných génov
  - Pomocou binomického testu vypočítame pravdepodobnosť spozorovania 8 a viac významných génov: p = 0.1079, teda nie významné

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

- Kompetitívny test
  - 1272 z 12639 génov je odlišne exprimovaných v tomto dátovom súbore (to je zhruba 10%)
  - V množine náhodne vybraných 96 génov očakávame teda 96 x 10% = 9.6 významných génov
  - p-hodnotu vypočítame z kontingenčnej tabuľky pomocou Fisherovho či Chi-kvadrát testu

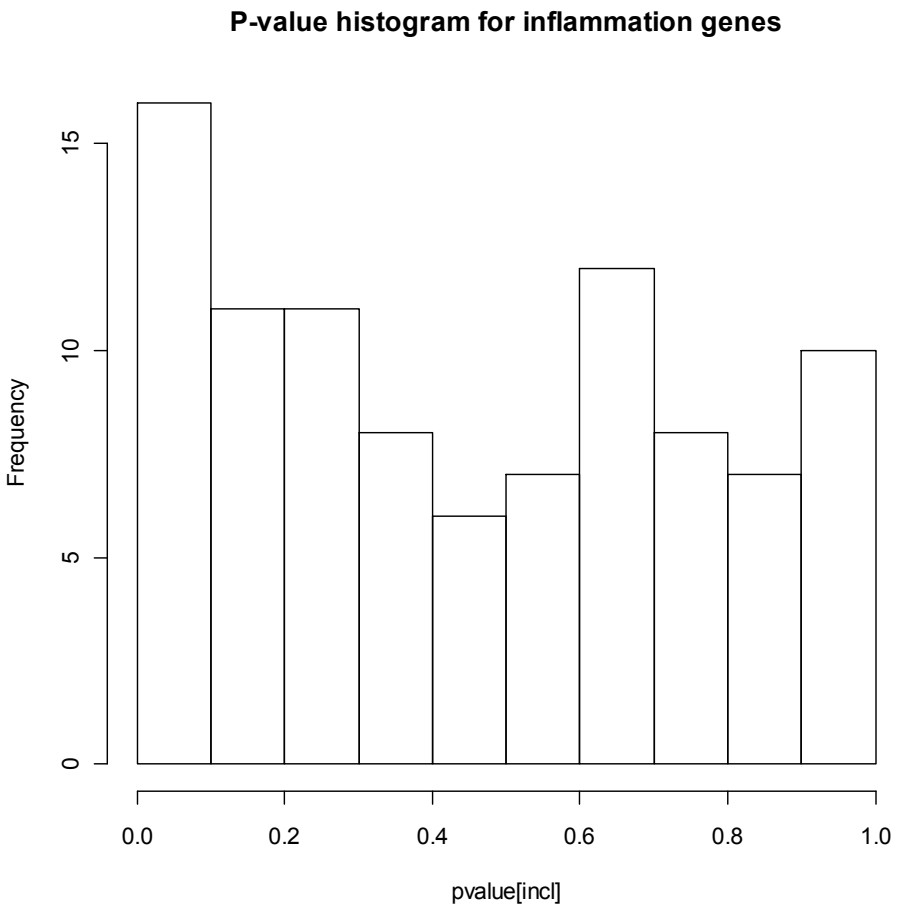
	V GS	Nie je v GS
Význ	8	1264
Nevýzn	88	11279

p = 0.73 Fisherovho testu (jednostranného):, teda nevýznamná

Metódy deliacej hranice vs. metódy celého zoznamu

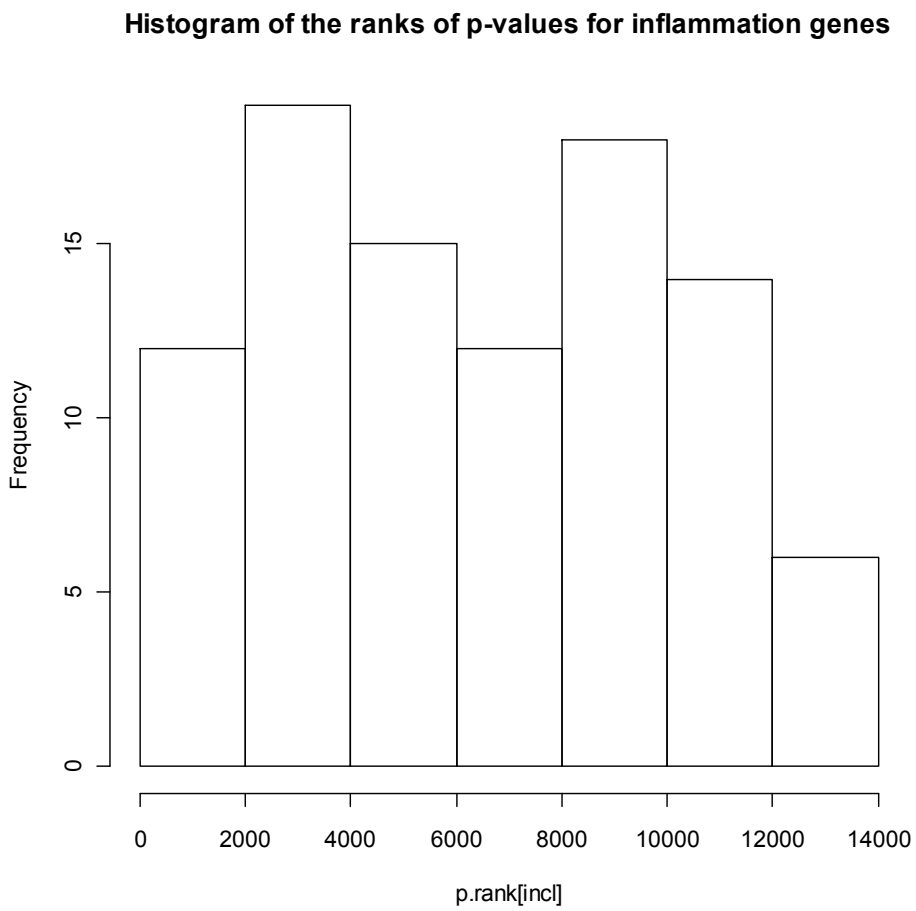
- Dve predchádzajúce metódy sú závislé na deliacich hraniciach – cut-offs
- V prípade, že povieme, že gén je pre nás významný už na 10% FDR, výsledok sa zmení
- Ďalej strácame informáciu tým, že redukujeme p-hodnotu na binárne premenné (významné/nevýznamné)
- Je rozdiel vedieť či štatisticky nevýznamné gény v našej množine sú takmer signifikantné na hranici významnosti alebo vôbec nie

Metóda celého zoznamu génov: *uzavretá*



- Môžeme študovať rozloženie p-hodnôt v množine génov
- V prípade že žiadne gény nie sú odlišne exprimované, malo by sa jednať o uniformné rozloženie
- Pík vľavo indikuje významnosť niektorých génov
- Aplikujeme Kolmogorov-Smirnov-Test pre porovnanie rozložení
- p = 8.2%, nie veľmi významné
- Je to uzavretá metóda, lebo používame len gény z génovej sady

Metóda celého zoznamu génov: *kompetitívna*



- Alternatívne sa môžeme dívať na rozloženie PORADÍ p-hodnôt
- Toto by bola kompetitívna metóda, pretože porovnávame našu génovú sadu s ostatnými génmi v experimente
- Zas môžeme aplikovať KS test
- p= 85.1%, veľmi nevýznamné

Uzavreté vs kompetitívne II.

- Výsledky kompetitívnych testov závisia na počte testovaných génov (napr. génov na microarray sklíčku a predchádzajúcom filtrovaní)
  - Na malom mikročipovom sklíčku, kde sú zmenené všetky gény, kompetitívna metóda nenájde žiadne odlišne exprimované množiny génov.

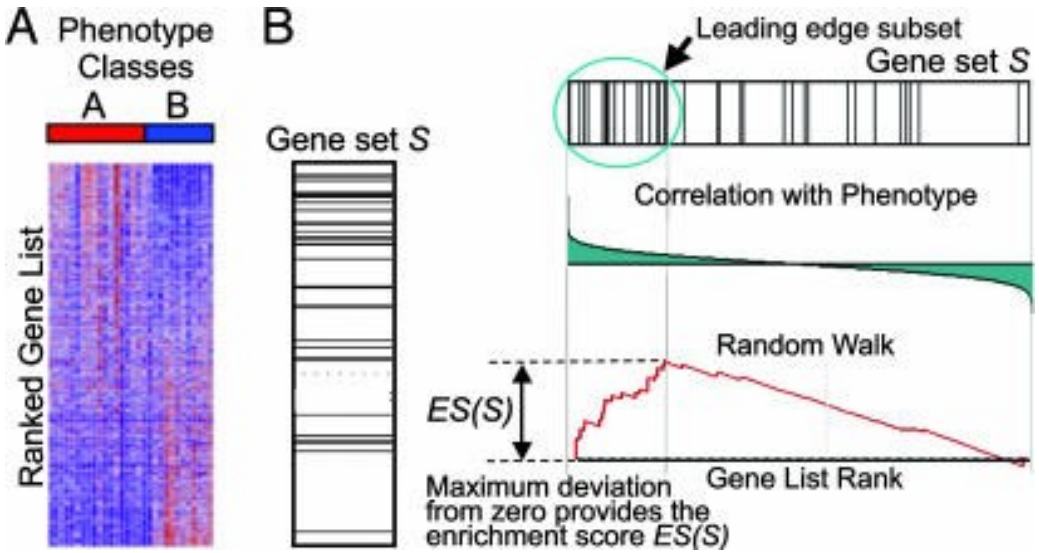
- Kompetitívne testy dávajú menej významných génov než uzavreté

Zmiešané metódy

- Najznámejšia je GSEA – gene set enrichment analysis (analýza obohatenia génovej sady)
- Počíta sa na zoradených p-hodnotách a sleduje sa,či sa gény z génovej sady sú náhodne rozložené v tomto zoradenom liste, alebo sa vyskytujú v horných,významných pozíciách
- Postup: 1. Výpočet skóre obohatenia (ES)

  2. Odhad významnosti ES (p hodnota) na základe permutačného testu

  3. Upravenie p-hodnôt na problém mnohonásobného porovnávania

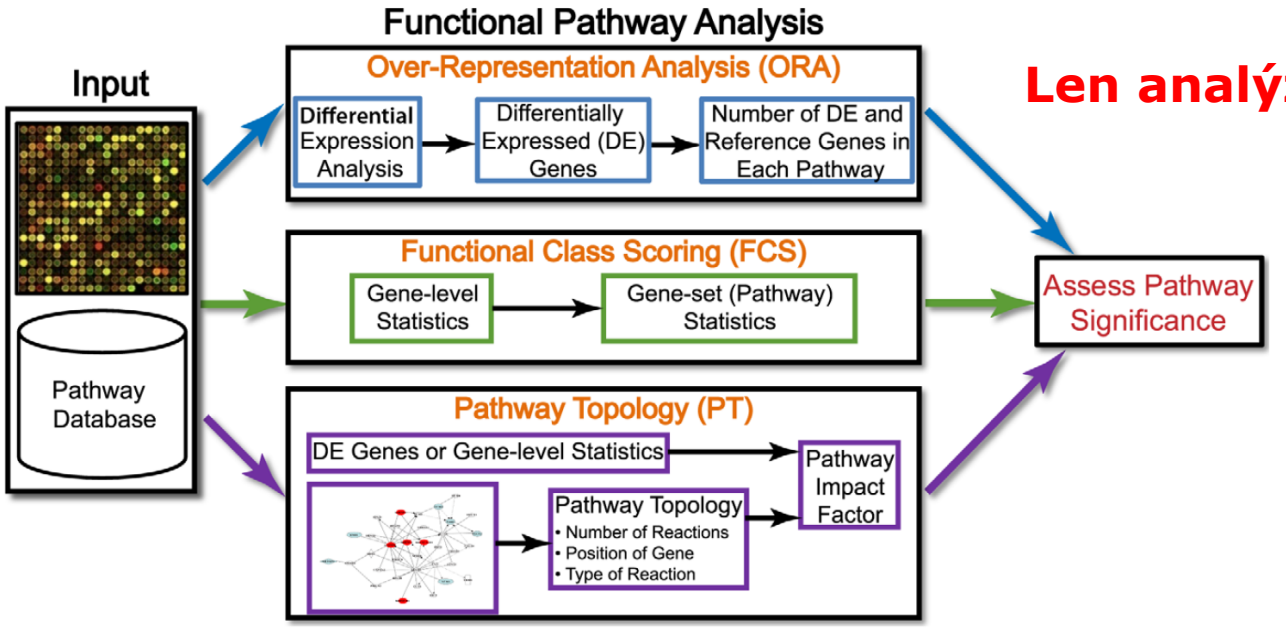


Ďalšie aspekty

- Smer zmeny
  - Ak chceme zistiť smer zmeny, musíme zopakovať analýzu pre jednostranný test
    - len up-regulované
    - len down-regulované

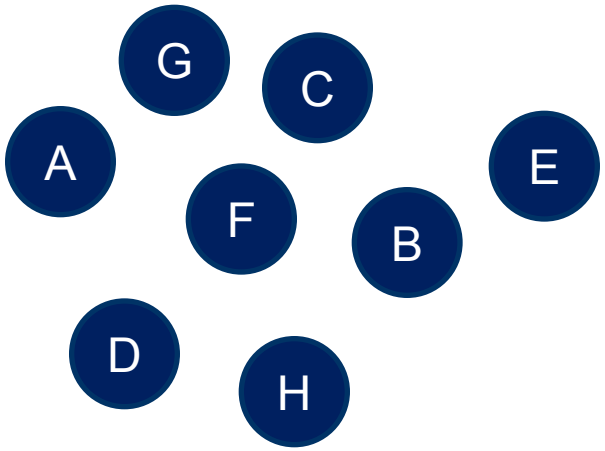
- Mnohonásobné testovanie
  - Takisto ako u testovania hypotéz na génoch medzi skupinami, aj tu ak máme veľký počet génových sád!
  - FDR je trochu komplikované, pretože génové množiny sa prekrývajú
  - Bonferroniho korekcia tu vždy funguje

Metódy využivajúce topológie

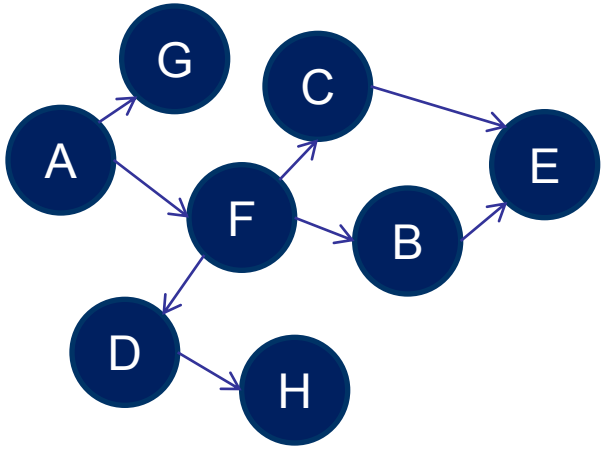


Len analýza dráh!

Bez topológie



S topológiou



Topológia využívaná rôzne

- Cieľ:
  - zmena priemernej expresie, korelácie, topológie
- Jednotka záujmu:
  - dráha, modul, cesta, gény
- Topológia známa vopred alebo odhadovaná z dát
- Celková sieť alebo individuálne dráhy

Topológia využívaná rôzne

- Cieľ:
  - zmena priemernej expresie, korelácie, topológie
- Jednotka záujmu:
  - dráha, modul, cesta, gény
- Topológia známa vopred alebo odhadovaná z dát
- Celková sieť alebo individuálne dráhy

Všeobecné princípy I.

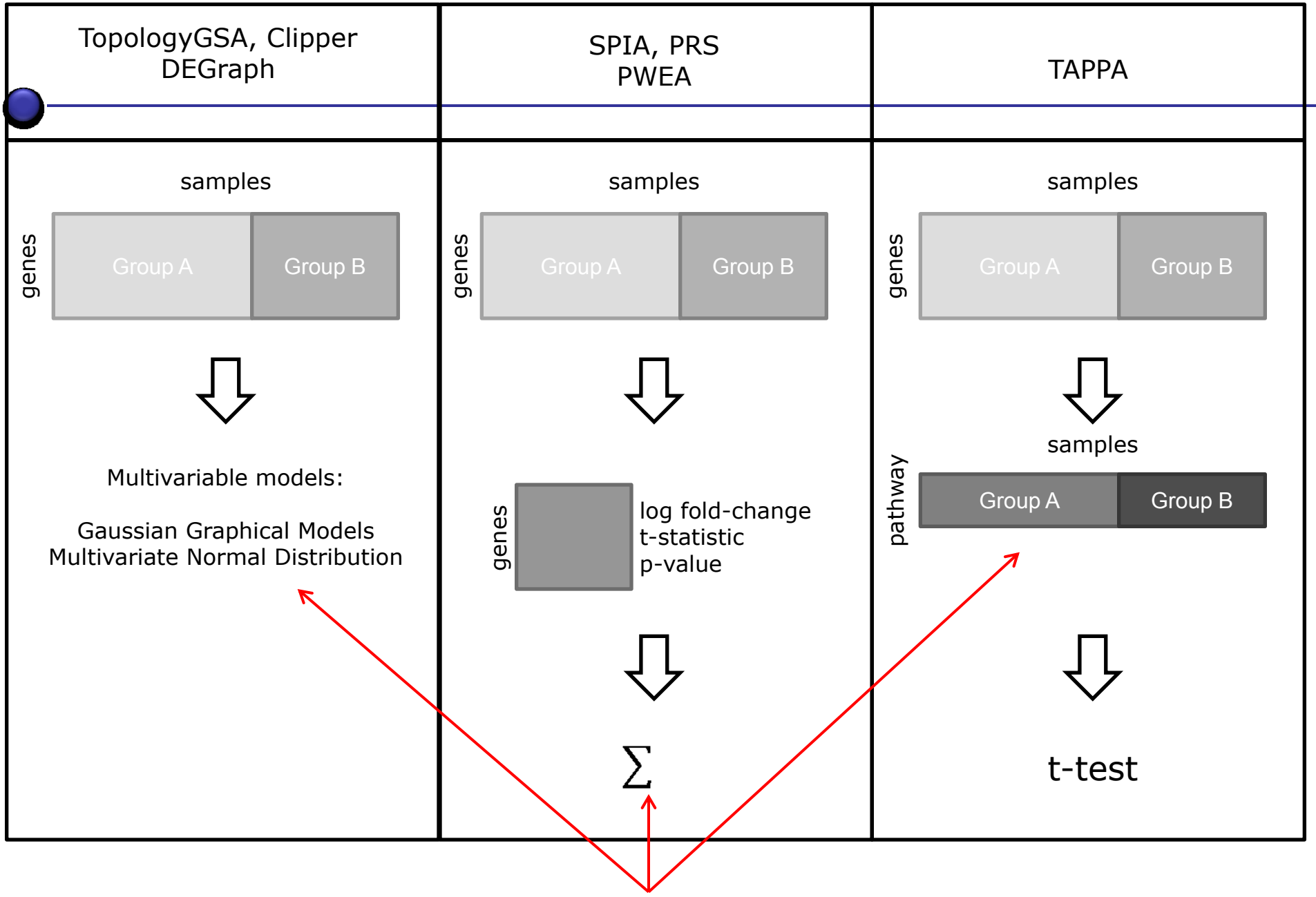
- Mnohorozmené metódy (TopologyGSA, clipper, DEGraph):
  - Grafové Gausovské modely (Graphical Gausian Models)
  - Analýza topológie + mnohorozmený test

Všeobecné princípy II.

- Jednorozmerné metódy (SPIA, PRS, PWEA, CePa):
  1. Analýza zmeny expresie génov
  2. Výber významných génov (voliteľné)
  3. Váhy podľa pozície génov v dráhe
  4. Sumarizácia
  5. Permutačný test

Všeobecné princípy III.

- Transformácia (TAPPA, PathOlogist):
  - Transformácia génového profilu na dráhový
  - Jednorozmerný test


TopologyGSA, Clipper
DEGraph
SPIA, PRS
PWEA
TAPPA
samples
genes
Group A
Group B
Multivariable models:
Gaussian Graphical Models
Multivariate Normal Distribution
log fold-change
t-statistic
p-value
Σ
pathway
t-test
Pathway topology

Pozor na korelácie medzi génmi !

- Všetky testy ktoré sme preberali predpokladali, že gény vnútri skupín sú nezávislé
  - To je ale veľmi nepravdepodobné!

- Ak sú gény korelovené, tak p-hodnoty jednotlivých testov (napr. Fisherov test) budú nesprávne
  - Vyriešime permutačnými metódami
    - Poprehadzujeme skupiny **vzoriek**
    - Zopakujeme analýzu
    - Porovnáme hodnoty s pozorovanými dátami

Pozor na prieniky medzi dráhami

- 250 KEGG dráh pre H. Sapiens
  - najčastejšie zastúpene gény

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

Študijný 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
```

- MSigDB - web

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

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

https://david.ncifcrf.gov/

Úloha [1 bod]

- Data ALL z balíka ALL (Bioconductor)
- Nájsť nasýtene (overrepresented) GO pojmy v sade génov odlišne exprimovaných medzi pacientami s fúziou BCR/ABL a bez tejto fúzie (pData(ALL)[,“mol.biol“])

- Odlišne exprimované gény: FDR = 5%
- balík GOstats
- fisher.test() pre GO:0005886 - plasma membrane, bez ohľadu na úroveň dôkazu