

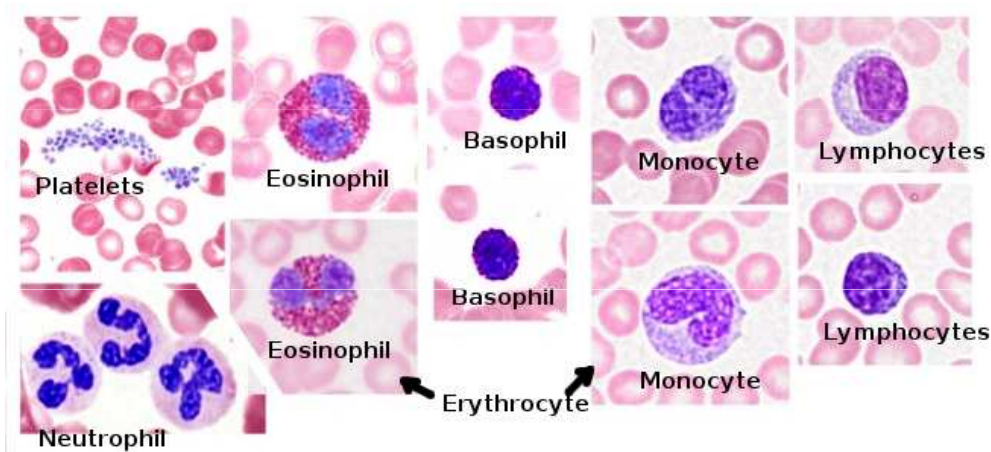
## Detekce biomarkerů z omics experimentů

- Mgr. Eva Budinská, PhD
- RECETOX
- [eva.budinska@recetox.muni.cz](mailto:eva.budinska@recetox.muni.cz)
- podzim 2024

# Odhad proporce různých typů buněk ve vzorku

(dekonvoluce)

# Motivace



## Cíl:

- Zjistit proporci **různých typů buněk** ve vzorku (např. imunitní buňky, epiteliální buňky, stromální buňky).
- Získat náhled na **heterogenitu** vzorků a **biologické procesy**, které ovlivňují genovou expresi.

## Význam:

- Pochopení mikroprostředí tkání (např. nádorových tkání).
- Analýza imunitní odpovědi (např. podíl T-buněk nebo makrofágů).
- Interpretace výsledků genové exprese ve směsných vzorcích (bulk RNA-seq).

## Příklad:

- Identifikace nádorových infiltrujících lymfocytů (TILs) z RNA-seq dat.

# Princip dekonvoluce genové exprese

Hlavní koncept:

Každý typ buňky má unikátní vzor genové exprese (tzv. signatura).

Genová exprese směsného vzorku je kombinací expresí z jednotlivých typů buněk podle jejich proporcí.

Základní model:

$$E = C \cdot S$$

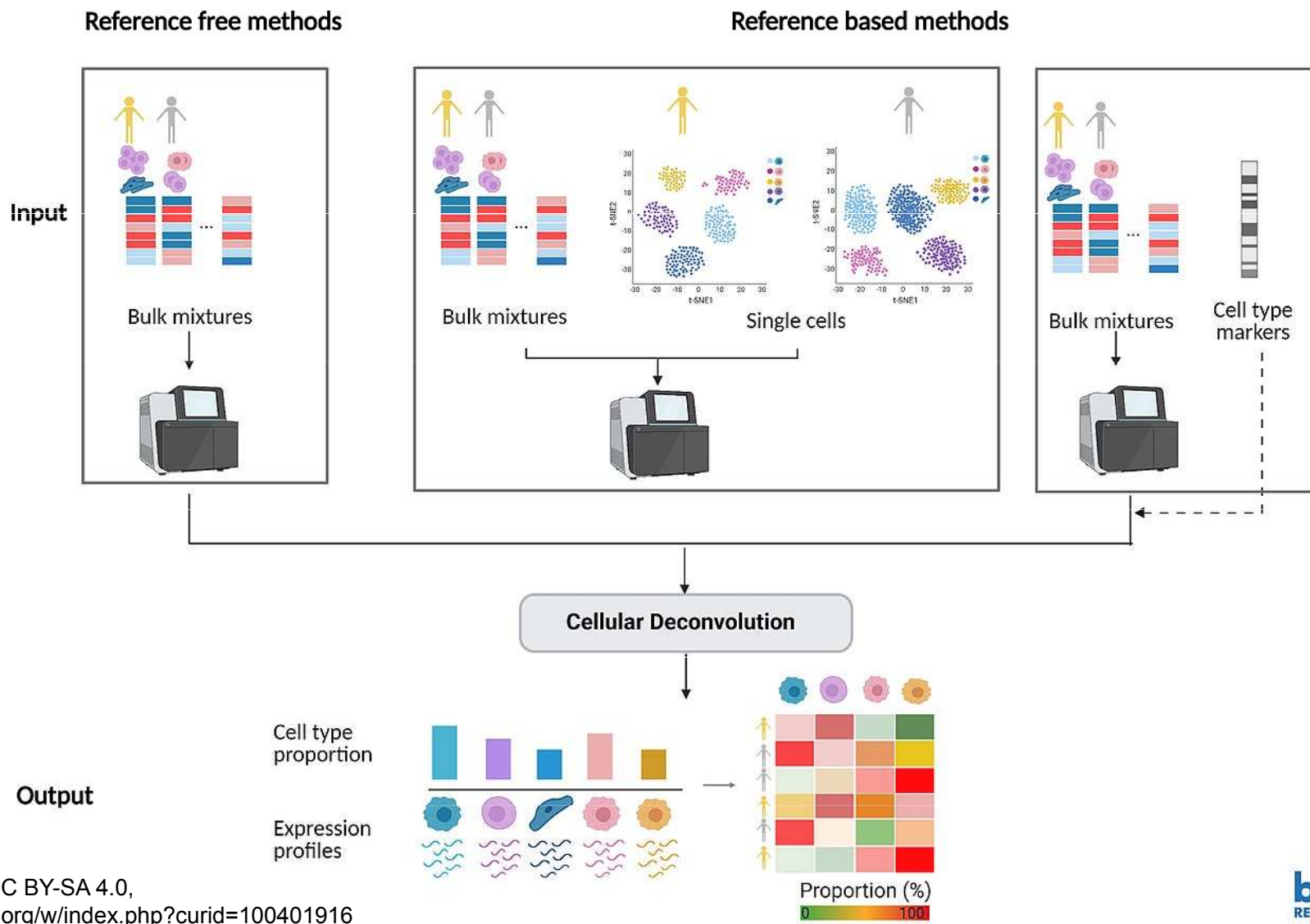
Kde:

*E*: Matice naměřených expresních dat (bulk RNA-seq).

*C*: Matice proporcí buněčných typů.

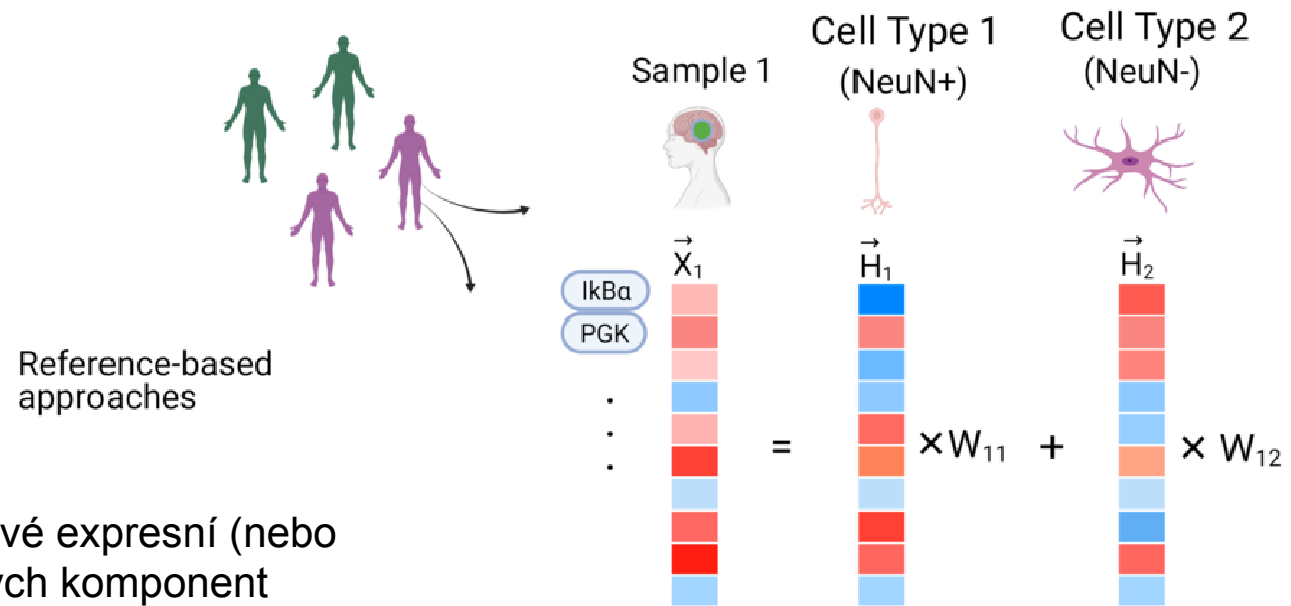
*S*: Matice signatur genové exprese specifických pro buněčné typy

# Metody



By Momur17 - Own work, CC BY-SA 4.0,  
<https://commons.wikimedia.org/w/index.php?curid=100401916>

# Metody založené na referenci (supervised)

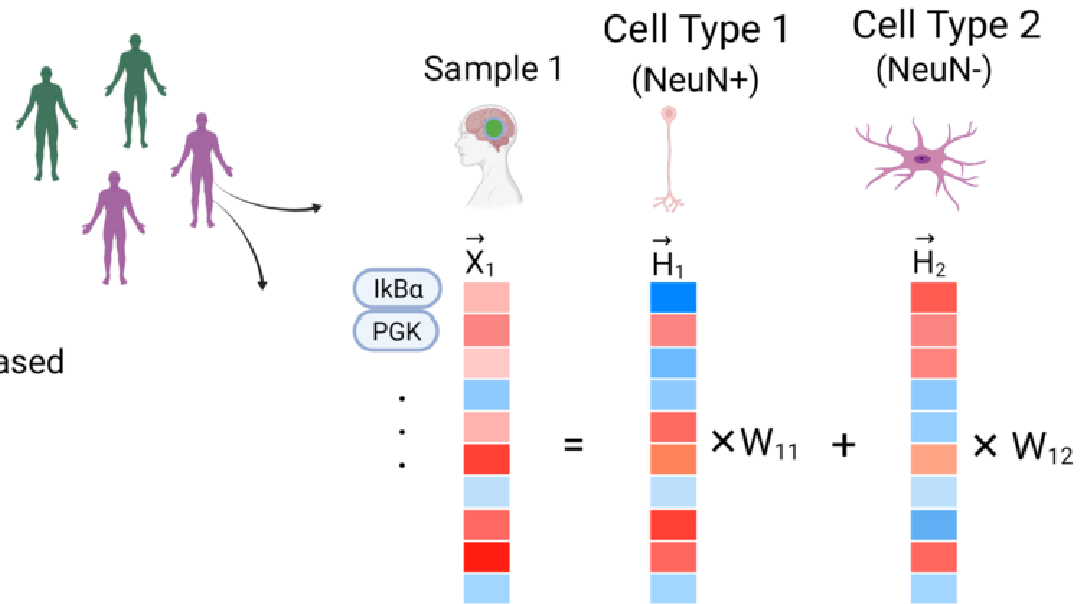


- Vyžadují **předem známé** genové expresní (nebo metylační atd) profily jednotlivých komponent
- Obvykle jsou známé hlavně **markerové geny**

## Příklady markerových genů:

- Imunitní buňky: **CD3E** (T-buňky), **CD19** (B-buňky), **CD68** (makrofágy)
- Stromální buňky: **COL1A1** (fibroblasty)
- Epiteliální buňky: **EPCAM**, **KRT18**

# Metody založené na referenci (supervised)



- Obvykle metody **regrese s omezením**

## Nevýhody

- Výsledek závisí na kvalitě markerových genů a znalostní databáze
- Celý signál se rozloží pouze do buněk, které máme ve slovníku

# Jak vznikají referenční signatury buněk

Referenční signatury jsou **vzorové profily genové exprese (metylace)**, které reprezentují unikátní expresní charakteristiky specifické pro jednotlivé buněčné typy.

Typicky jde o seznam genů a jejich úrovní exprese (metylace)

## 1. Izolace jednotlivých buněk:

- Fluorescenční třídění buněk (FACS)
- Mikrodisekce
- Jednobuněčná RNA-seq (scRNA-seq)
- Buněčné kultury

## 2. Analýza profilů

- Jednobuněčná RNA-seq (scRNA-seq)
- RNAseq, microarray, ....

## 3. Derivace markerových genů



# Jak vznikají referenční signatury buněk

Referenční signatury jsou **vzorové profily genové exprese (metylace)**, které reprezentují unikátní expresní charakteristiky specifické pro jednotlivé buněčné typy.

Typicky jde o seznam genů a jejich úrovní exprese (metylace)

## 1. Izolace jednotlivých buněk:

- Fluorescenční třídění buněk (FACS)
- Mikrodisekce
- Jednobuněčná RNA-seq (scRNA-seq)
- Buněčné kultury

## 2. Analýza profilů

- Jednobuněčná RNA-seq (scRNA-seq)
- RNAseq, microarray, ....

## 3. Derivace markerových genů

## 4. Validace

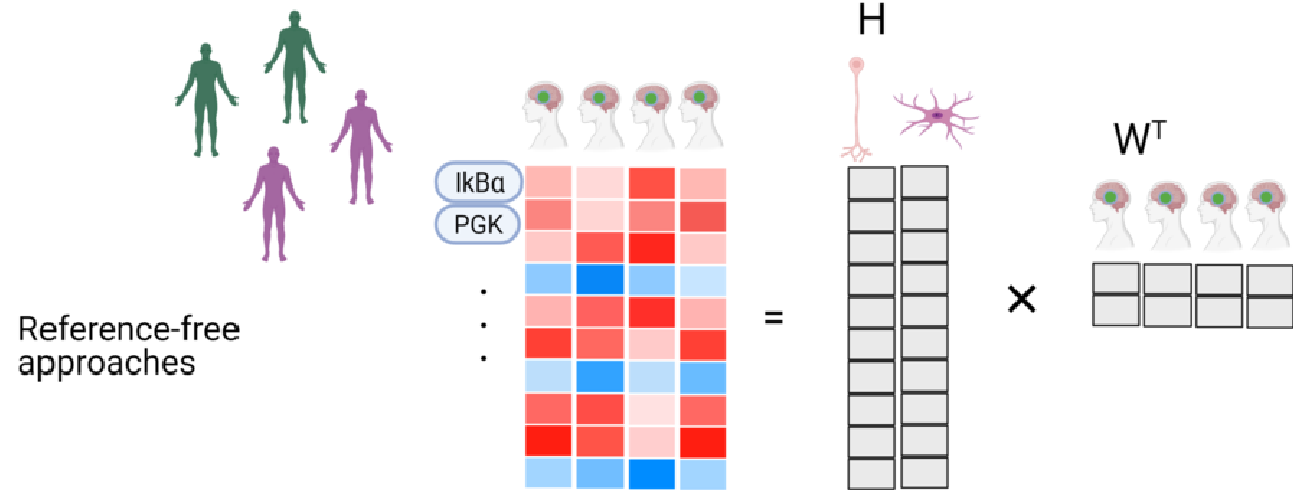
### **PROBLÉMY:**

- **Heterogenita buněk:** Buněčné typy nejsou homogenní a mohou mít různé stavy (např. aktivované vs. klidové buňky).

- **Specifičnost tkáně:** Některé signatury se mohou měnit podle kontextu (např. makrofágy v plicích vs. ve slezině).

- **Technické artefakty:** Rozdíly mezi platformami (scRNA-seq vs. bulk RNA-seq) mohou ovlivnit přesnost.

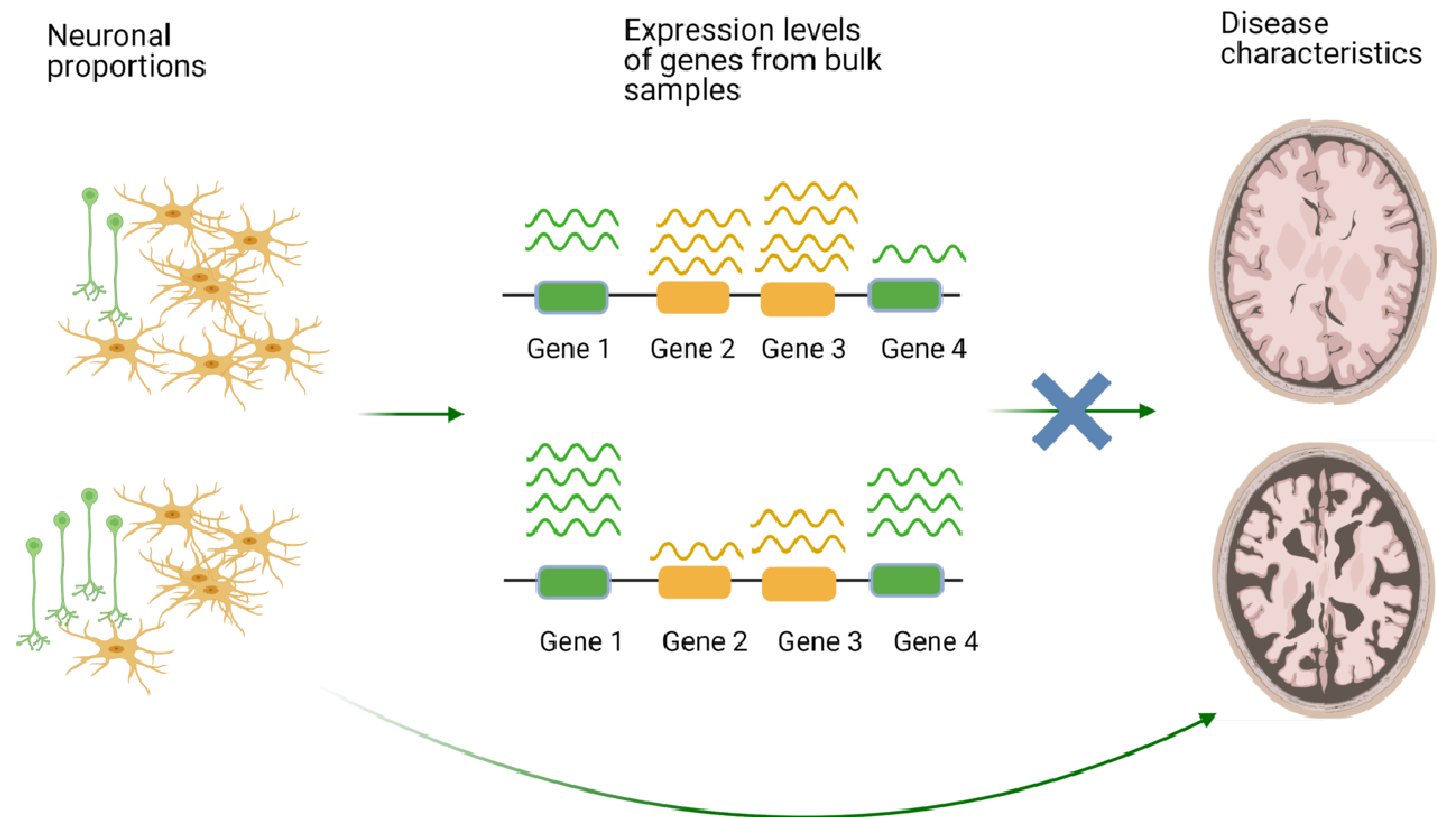
# Metody bez referenčních profilů (unsupervised)



- Nemají referenční signatury odhadují tyto signatury i složení zaráz
- Nejčastější přístupy:
  - **Non-negative matrix factorization**
  - **Bayesovské metody**

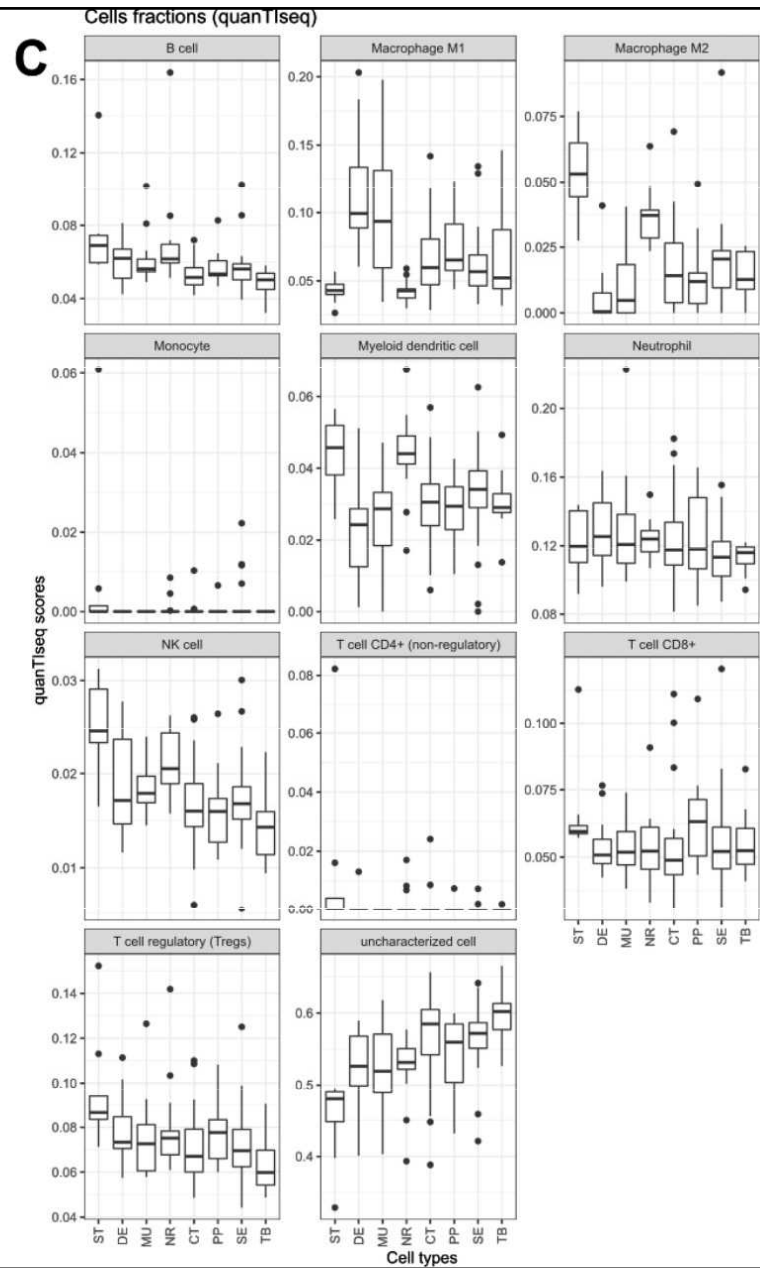
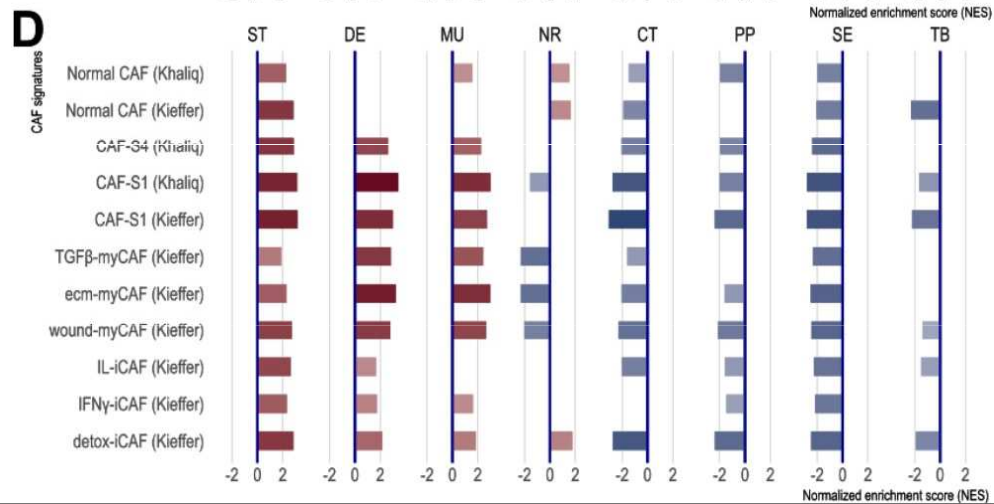
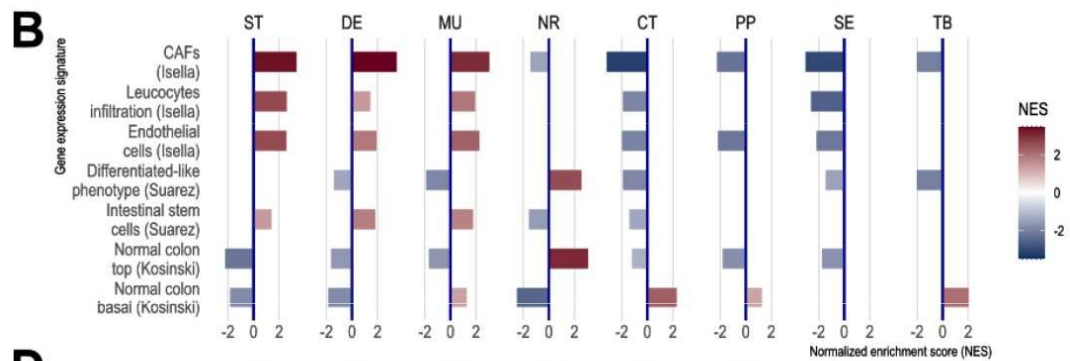
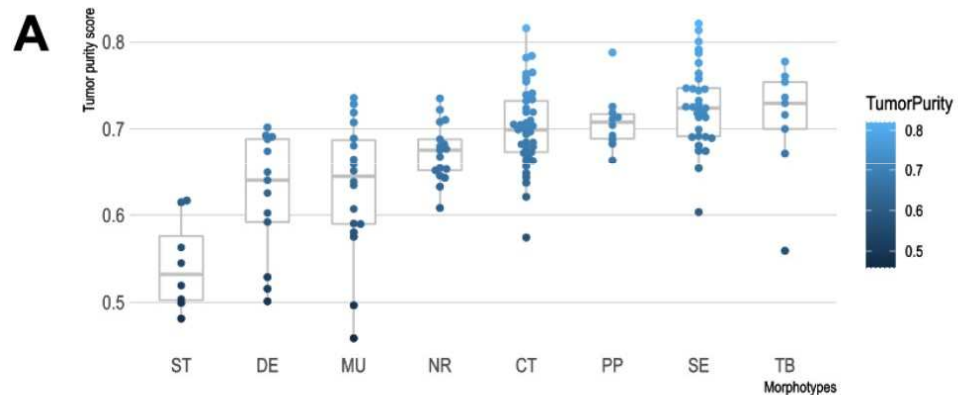
# Příklad

Matoucí vliv podílu buněčných typů může vést k chybným asociacím mezi genovou expresí v mozkové kůře a klinickou patologií Alzheimerovy choroby.



Patrick E, Taga M, Ergun A, Ng B, Casazza W, Cimpean M, et al. (August 2020). ["Deconvolving the contributions of cell-type heterogeneity on cortical gene expression"](#). *PLOS Computational Biology*. 16 (8): e1008120.

**MUNI | RECETOX**



TOX

# Nejčastější metody

Zkratka metody	Název metody	Typ metody	Vstupní data	Rok publikace
<a href="#">CIBERSORT<sup>[23]</sup></a>	Robust enumeration of cell subsets from tissue expression profiles	Reference based	Gene expression	2018
<a href="#">CDSeq<sup>[24]</sup></a>	A complete deconvolution method for dissecting tissue heterogeneity	Reference free	Gene expression	2019
<a href="#">FARDEEP<sup>[25]</sup></a>	Fast and robust deconvolution of expression profiles	Reference based	Gene expression	2019
<a href="#">UNDO<sup>[26]</sup></a>	Unsupervised deconvolution of tumor-stromal mixed expressions	Reference free	Gene expression	2015
<a href="#">dtangle<sup>[27]</sup></a>	Accurate and robust cell type deconvolution	Reference based	Gene expression	2019
<a href="#">EPIC<sup>[28]</sup></a>	Estimating the proportions of different cell types from bulk gene expression data	Reference based	Gene expression	2017
<a href="#">BSEQ-sc<sup>[29]</sup></a>	Deconvolution of bulk sequencing experiments using single cell data	Reference based	Gene expression	2016
<a href="#">MuSiC<sup>[18]</sup></a>	Cell-type Identification by estimating relative subsets of RNA transcripts	Reference based	Gene expression	2019
<a href="#">SCDC<sup>[30]</sup></a>	Bulk gene expression deconvolution by multiple single-Cell RNA sequencing references	Reference based	Gene expression	2020
<a href="#">DWLS<sup>[31]</sup></a>	Gene expression deconvolution using dampened weighted least squares	Reference based	Gene expression	2019
<a href="#">deconvSeq<sup>[32]</sup></a>	Deconvolution of cell mixture distribution in sequencing data	Reference based	Gene expression	2019
<a href="#">Bisque<sup>[19]</sup></a>	Decomposition of bulk expression with single-cell sequencing	Reference based	Gene expression	2020
<a href="#">TOAST<sup>[33]</sup></a>	Tools for the analysis of heterogeneous tissues	Reference free	DNA methylation	2019
<a href="#">Houseman<sup>[9]</sup></a>	Reference-free deconvolution of DNA methylation data and mediation by cell composition effects	Reference based	DNA methylation	2016
<a href="#">methyCC<sup>[34]</sup></a>	Technology-independent estimation of cell type composition using differentially methylated regions	Reference based	DNA methylation	2019
<a href="#">BayesCCE<sup>[35]</sup></a>	Bayesian framework for estimating cell-type composition from DNA methylation	Reference free	DNA methylation	2018

# Srovnání metod

**Figure 3.** Key characteristics and technical evaluation of cellular deconvolution methods. (A) Method characterization according to implementation, input, output, embedded reference and the underlying algorithm.

(B) Performance assessment based on five criteria:  
 the accuracy of the predicted cell type proportions,  
 the scalability in analyzing large input sizes,  
 the stability (opposite of crash rate and other errors),  
 the consistency of the predicted cell type proportions using  
 different initializations  
 usability as code quality and ease of use.

\*Abbreviations: S: signature matrix; F: full cell-type expression matrix; PCA: principal component analysis; NMF: non-negative matrix factorization; CLS: constrained least squares; SVR: support vector regression; MLE: maximum likelihood estimation; DNN: deep neural network; ensemble: combination of multiple methods; scoring: enrichment using marker sets. W prefix: weighted. R prefix: regularized. \*\*\*BisqueRef requires scRNA data of at least two subjects as input. TICPE requires cancer cell expression, normal cell expression, immune cell expression and marker gene sets as input.

Method	Input						Technical Evaluation								
	Platform	#Cell types	scRNA-seq	CT expr**	Markers**	Embedded Ref.	Rel. Proportion	Signature	Main Technique**	Overall	Accuracy	Scalability	Consistency	Stability	Usability
<b>Reference-based</b>															
MuSiC	R	✓						W-CLS							
DWLS	R	✓	S				✓	W-CLS							
LinDeconSeq	R	✓	S				✓	W-CLS							
AdRoit	R	✓					✓	R-CLS							
RNA-Sieve		✓					✓	MLE							
Scaden		✓						DNN							
spatialDWLS	R	✓	S	✓			✓	W-CLS							
AutoGeneS		✓	S				✓	CLS/SVR							
DecOT		✓	S				✓	Ensemble							
BayesPrism	R	✓					✓	Bayesian							
DigitalDLSorter		✓						DNN							
BayICE	R	✓					✓	Bayesian							
DeconPeaker		✓	S				✓	CLS							
CPM	R	✓						SVR							
BisqueRef***	R	✓						CLS							
SCDC	R	✓	S				✓	Ensemble							
DAISM-DNN		✓				✓		DNN							
MCMF	R	✓					✓	NMF							
DeMixT	R	✓						MLE							
deconvSeq	R	✓	S				✓	MLE							
<b>Reference-free</b>															
CIBERSORT	R		S	✓				v-SVR							
MethylResolver	R		S	✓				CLS							
MIXTURE	R		S					v-SVR							
FARDEEP	R		S					CLS							
MySort			S	✓				v-SVR							
NITUMID	R		S	✓			✓	NMF							
quanTiseq			S	✓				CLS							
DeconRNASeq	R		S					CLS							
Bseq-SC	R		S	✓			✓	v-SVR							
DCQ			S	✓				R-CLS							
DESeq2's unmix	R		F					CLS							
dtangle	R		F	✓				Scoring							
ARIC			F					W-SVR							
PREDE	R		F				✓	NMF							
EMeth	R		F					MLE							
ImmuCellAI	R		F	✓	✓	✓		CLS							
EPIC	R		F	✓	✓	✓		W-CLS							
DeCompress	R	✓	F				✓	Ensemble							
TICPE***	R		F	✓				Scoring							
<b>Reference-free</b>															
Linseed	R	✓						Scoring							
TOAST	R	✓					✓	NMF/PCA							
CellDistinguisher	R	✓					✓	NMF							
DeconICA	R	✓					✓	NMF							
deBGAM	R	✓					✓	NMF							
BayesCCE		✓						Bayesian							
deconf	R	✓					✓	NMF							
ReFACTor	R	✓					✓	PCA							
BayCount	R	✓					✓	Bayesian							
SMC		✓					✓	Bayesian							
<b>Semi-reference-free</b>															
MCP-counter	R	✓		✓	✓	✓		Scoring							
Deblender		✓		✓	✓	✓		NMF							
BisqueMarker	R	✓		✓	✓	✓		PCA							
DSA	R	✓		✓	✓	✓		Scoring							

# Další čtení

- [Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges | Nucleic Acids Research | Oxford Academic](#)
- [Comprehensive evaluation of deconvolution methods for human brain gene expression | Nature Communications](#)

# Analýza genových sad

(pathway analýza)



**Jak se hledá potenciální biomarker v omics datech**

Biologická otázka (hypotéza)

Dizajn experimentu

Provedení experimentu (hybridizace mikročipů, hmotnostní spektrometrie...)

N matic základních dat (jedna pro každý z N vzorků)

Kontrola kvality  
Normalizace  
Sumarizace

Finální datová matice N vzorků a K genů (proteinů)

Matice informací o vzorcích N x P (např. klinická data v medicíně)

Objevování skupin? (Shlukování)

**Nové skupiny genů nebo vzorků**

Charakterizace nových skupin

Porovnání skupin? (Testování)

**List genů s odlišnou expresí mezi skupinami vzorků**

Predikce skupin? (Klasifikace)

**Klasifikační pravidlo využívající genovou expresi**

**Analýza přežití**

**Seznam prognostických genů**

Analýza časových řad

List genů se stejným profilem změn exprese v čase

**Pathway analýza**

Interpretace

**Validace**

Publikace

# Motivace

- Geny, proteiny a další molekuly jsou navzájem propojené ve velké spleti různých signálních, metabolických a různých jiných drah
- Potřebujeme zjistit, jaké dráhy jsou zasažené naším experimentálním protokolem (liší se v mezi skupinami)

# Jak na to?

- Seznam molekul můžeme **ad-hoc vložit** do existující databáze drah a podívat se kam patří (KEGG, MsigDB....)
  - nevýhoda – nemáme statistickou významnost
- **Provedeme analýzu genových sad** (pathway analýzu)
  
- **Předpoklad všech těchto analýz:** operují s již definovanými skupinami genů

# 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 nemusí být dráha –  
je to všeobecnější a méně  
specifický pojem

# Cíl analýzy 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 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

# GO databáze

**Term Lineage**

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

▼ Filter tree view ?

Filter Gene Product Counts	View Options
Data source	Tree view <input checked="" type="radio"/> Full <input type="radio"/> Compact
Species	<input type="button" value="Set filters"/>
All	<input type="button" value="Remove all filters"/>
AspGD	
CGD	
dictyBase	

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]**

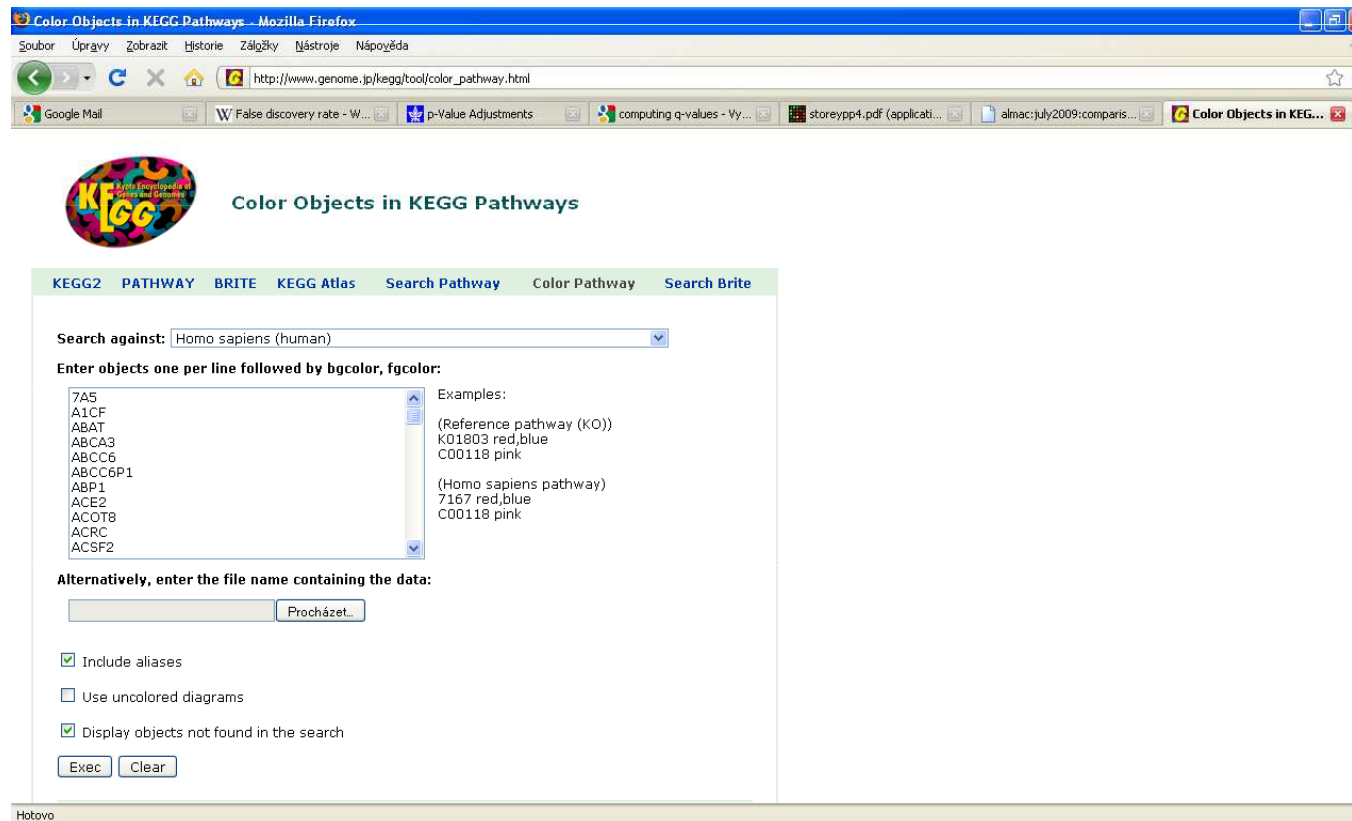
- GO:0009611 : response to wounding [2289 gene products]
  - GO:0006954 : inflammatory response [1173 gene products]
    - GO:0002526 : acute inflammatory response [427 gene products]
      - GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]**



# KEGG pathway databáze

- KEGG = Kyoto Encyclopedia of Genes and Genomes
- <http://www.genome.jp/kegg/pathway.html>
- Více informací než GO, máme tu již 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



The screenshot shows a web browser window titled "Color Objects in KEGG Pathways - Mozilla Firefox". The address bar displays the URL "http://www.genome.jp/kegg/tool/color\_pathway.html". The browser's menu bar includes "Soubor", "Úpravy", "Zobrazit", "Historie", "Záložky", "Nástroje", and "Nápožďeda". The browser's toolbar shows icons for back, forward, home, and search, along with several open tabs: "Google Mail", "W 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 tabs: "KEGG2", "PATHWAY", "BRITE", "KEGG Atlas", "Search Pathway", "Color Pathway", and "Search Brite".

The "Search Pathway" tab is active, showing a search interface. The "Search against:" dropdown menu is set to "Homo sapiens (human)". Below this, the instruction "Enter objects one per line followed by bgcolor, fgcolor:" is displayed. A text input field contains the following list of objects:

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

To the right of the input field, under the heading "Examples:", two examples are provided:

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

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

Below the input field, the instruction "Alternatively, enter the file name containing the data:" is shown, followed by 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 the form are "Exec" and "Clear" buttons.

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

# KEGG

Search PATHWAY - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápožeda

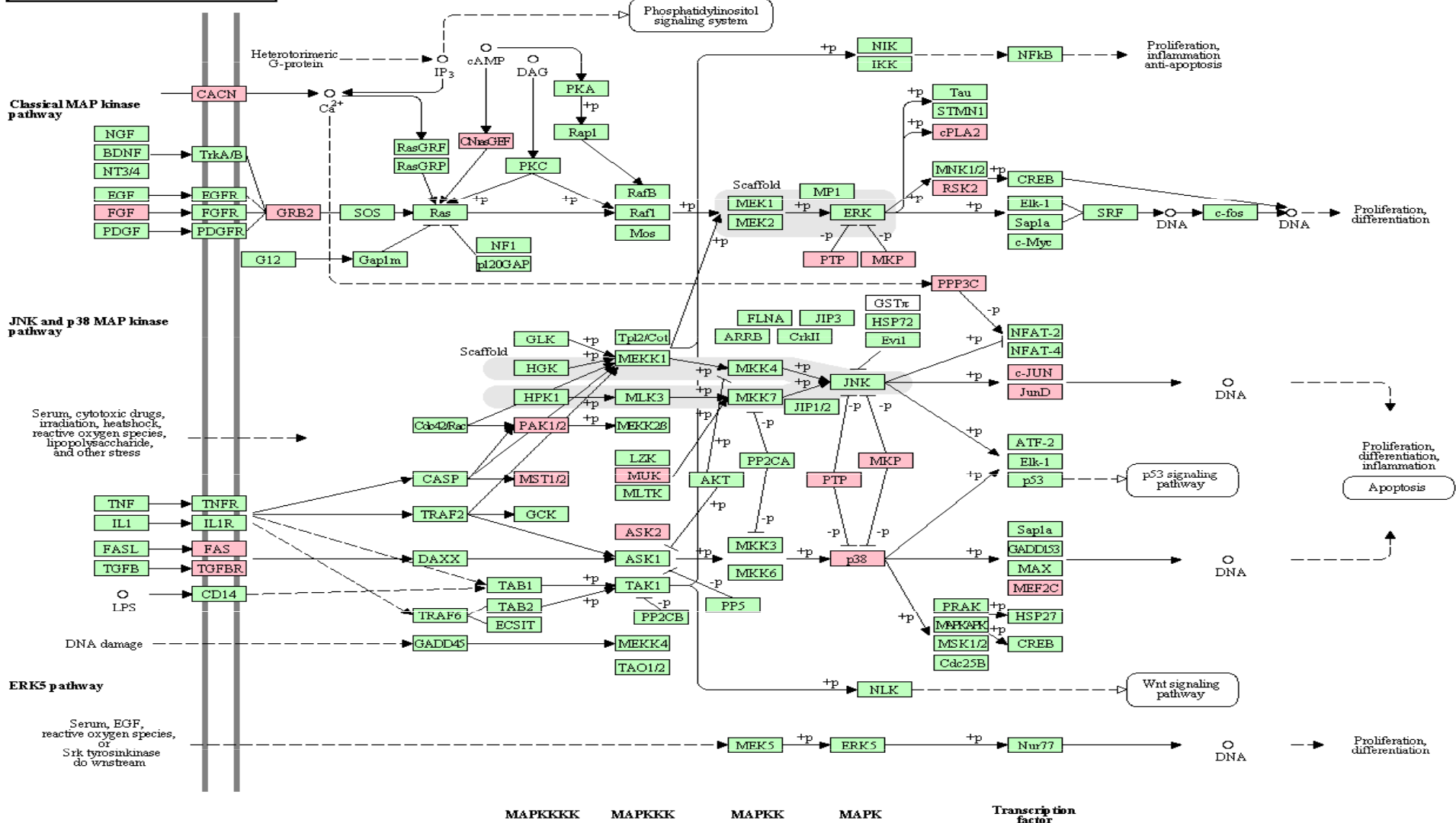
http://www.genome.jp/kegg-bin/color\_pathway\_object

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**



MAPK K K K K    MAPK K K K    MAPK K    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 zbarvit jednotlivé geny  
podle rozdílných barev

# MsigDB databáze

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

**GSEA**  
Gene Set Enrichment Analysis

login  
register

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

MSigDB Home  
About Collections  
Browse Gene Sets  
Search Gene Sets  
Investigate Gene Sets  
View Gene Families  
Help

**MSigDB**  
Molecular Signatures Database

Molecular Signatures Database v7.4

**Overview**

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the **HALLMARK\_APOPTOSIS** gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEX**

**License Terms**

GSEA and MSigDB are available for use under these license terms.

**Collections**

The MSigDB gene sets are divided into 9 major collections:

- H hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
- C1 positional gene sets** for each human chromosome and cytogenetic band.
- C2 curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- C3 regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.
- C4 computational gene sets** defined by mining large collections of cancer-oriented microarray data.
- C5 ontology gene sets** consist of genes annotated by the same ontology term.

# 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?

---

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?

Náhodně očekáváme  $96 \times 5\% = 4.8$   
významných genů

Jaká je pravděpodobnost pozorování **8** a více  
významných genů?

**Vhodné testy:**

*Binomický test* ( $p = 0.1079$ )

**Příklad, uzavřená metoda  
dělicí hranice**

# Hypergeometrický nebo binomický test?

Kritérium	Binomický test	Hypergeometrický test
Reálné použití	Opakované pokusy s neměnnou šancí na výsledek	Biologická data s konečným počtem vzorků
Příklad aplikace	Určení podílu konkrétních výsledků v pokusech	Výběr genů bez vrácení, analýza nadměrného výskytu
Velikost populace	Teoreticky neomezená populace	<b>Konečná</b> populace
Typ vzorkování	S náhradou	<b>Bez náhrady</b>
Zohlednění závislosti vzorků	Vzorky jsou nezávislé	Zohledňuje vzájemnou závislost vzorků
Typický kontext použití	Opakované nezávislé pokusy (např. házení mincí)	<b>Obohacení genových sad</b> (gene set enrichment analysis)
Pravděpodobnost	Zůstává konstantní	<b>Mění se s každým výběrem</b>

**Příklad, uzavřená metoda dělicí hranice**

# Hypergeometrický test... pravděpodobnost výběru každého dalšího genu se mění s každým dalším výběrem

	V GS	Není v GS
Význ	8	1264
Nevýzn	88	11279

```
x <- 8 # Počet úspěchů ve vzorku (geny s p < 0,05 v sadě)
m <- 1272 # Celkový počet úspěchů v populaci (všechny geny s p < 0,05)
n <- 11367 # Celkový počet neúspěchů v populaci (všechny ostatní geny)
k <- 96 # Velikost vzorku (genová sada)
```

Výpočet hypergeometrické pravděpodobnosti

```
p_value <- phyper(q = x - 1, m = m, n = n, k = k, lower.tail = FALSE)
```

p\_value **0.7627**

**Příklad, kompetitivní metoda dělicí hranice**

# Hypergeometrický nebo Fisherův test?

	V GS	Není v GS
Význ	8	1264
Nevýzn	88	11279

$p = 0.7627$  (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

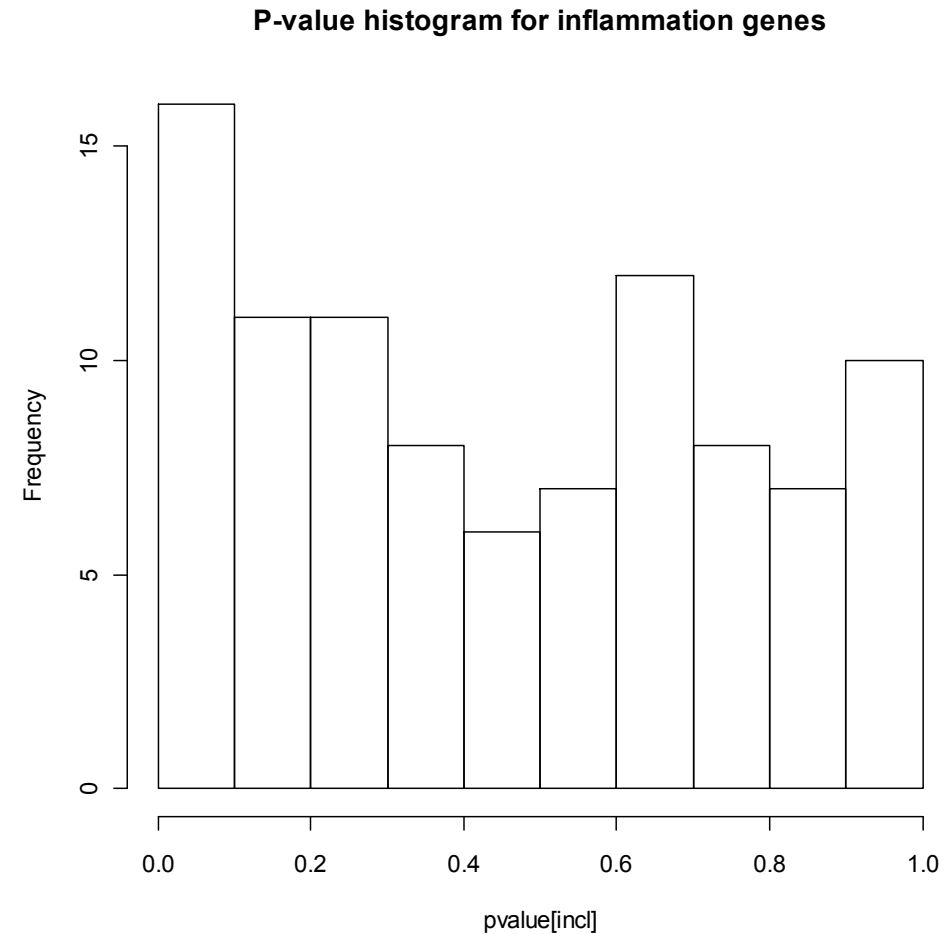


# Metody dělicí hranice vs. metody celého seznamu

- 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*

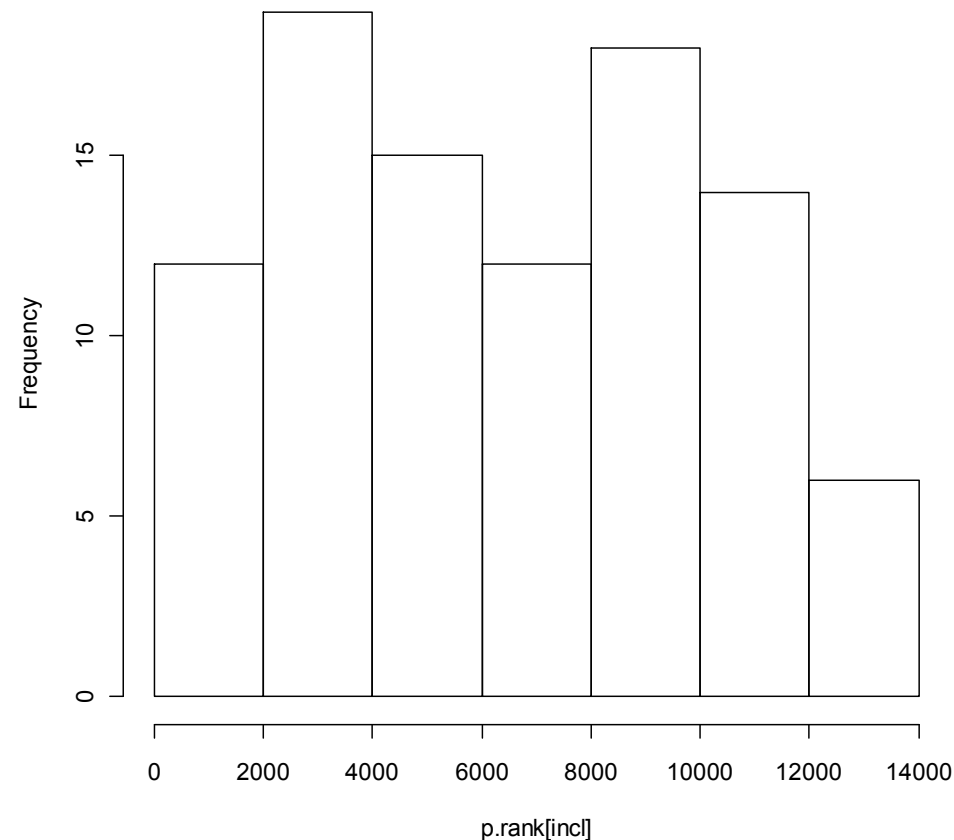
- Můžeme studovat rozložení p-hodnot v genové sadě
- V případě, že žádné geny nejsou odlišně exprimované, mělo by se jednat o uniformní rozložení
- Pík vlevo indikuje významnost některých genů
- Aplikujeme Kolmogorův-Smirnovův test pro porovnání rozložení
- $p = 8.2\%$ , není velmi významné
- Je to **uzavřená** 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



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

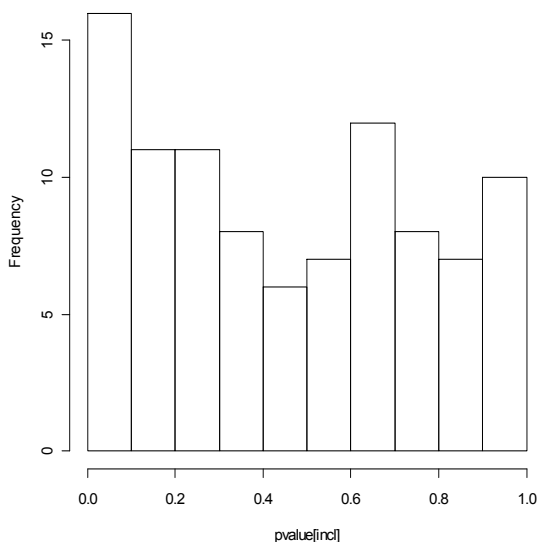
# Metoda celého seznamu genů: uzavřená

p-hodnota pořadí

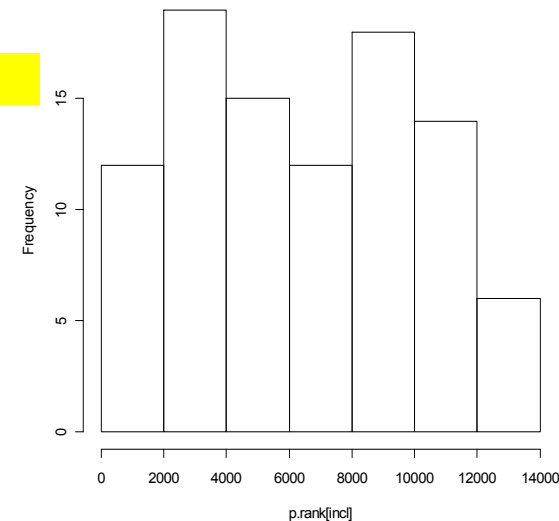
Gen A	0.001	1
Gen H	0.001	2
Gen Z	0.031	3
Gen G	0.024	4
.	.	.
Gen M	0.024	62
.	.	.
Gen O	0.049	1272
.	.	.
Gen J	0.351	5843
.	.	.
Gen L	0.454	7390
.	.	.
Gen B	0.752	10287
.	.	.
.	.	.
Gen C	0.989	12639

# Metoda celého seznamu genů: kompetitivní

P-value histogram for inflammation genes



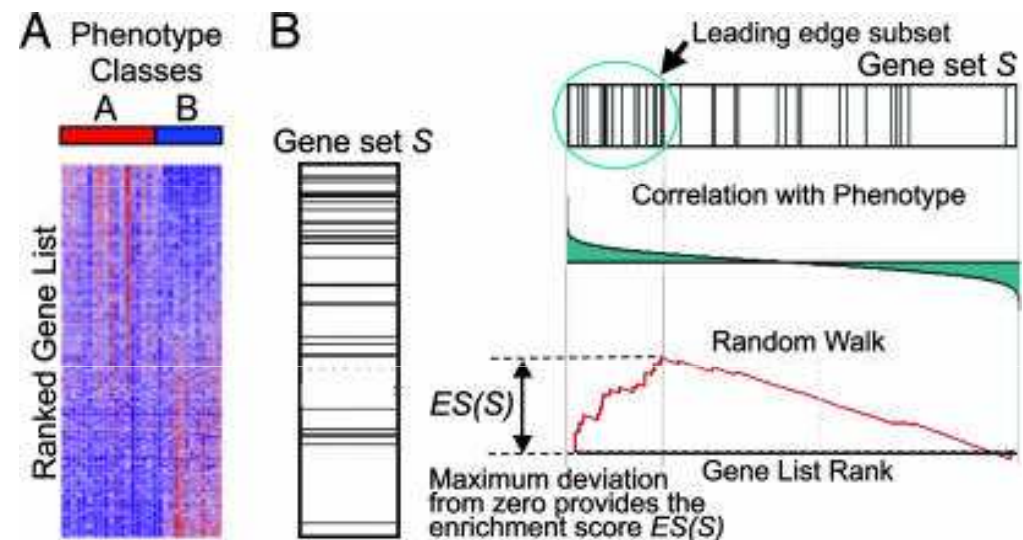
Histogram of the ranks of p-values for inflammation genes



rozložení p-hodnot v genové sadě

# GSEA

- 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 rozloženy 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



# 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

## Další aspekty

### Směr změny

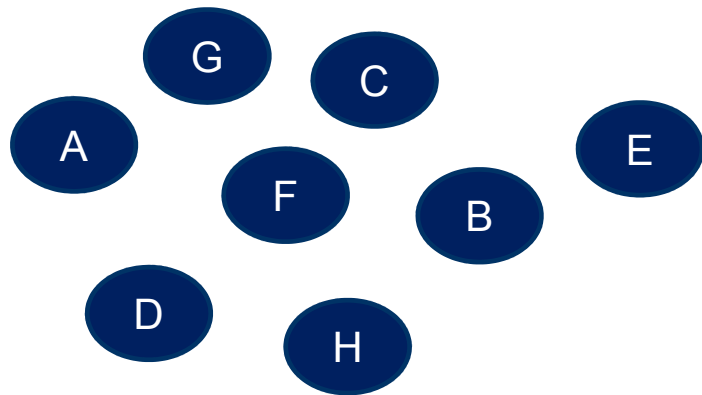
- Pokud chceme zjistit **směr** změny, musíme zopakovat analýzu pro jednostranný test
  - jen up-regulované
  - jen down-regulované

### Mnohonásobné testování

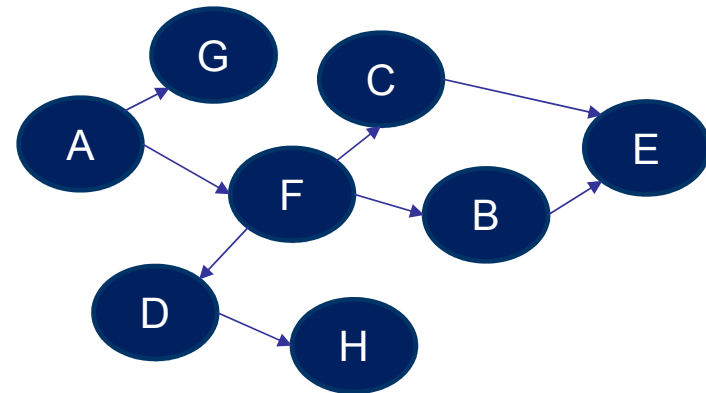
- Stejně jako u testování hypotéz na genech mezi skupinami, i pokud máme velký počet genových sad!
- FDR je trochu komplikované, protože genové množiny se překrývají
- Bonferroniho korekce vždy funguje

# Topologie

Bez topologie



S topologií





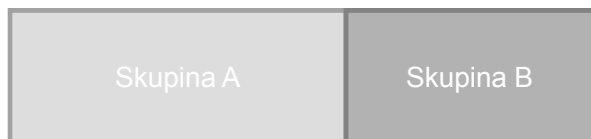
# Topologie využívaná různě

- 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



Mnohorozměrné modely:

Gaussian Graphical Models  
Multivariate Normal Distribution

SPIA, PRS  
PWEA

Vzorky

gény



gény



Změna exprese  
t-statistika  
p-hodnota

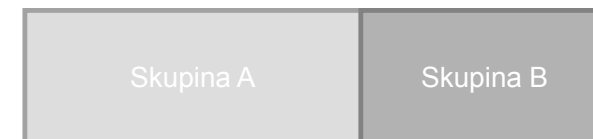


$\Sigma$

TAPPA

Vzorky

gény



Vzorky

dráhy



t-test

# Příklad – topologie uzavřená metoda dělicí hranice

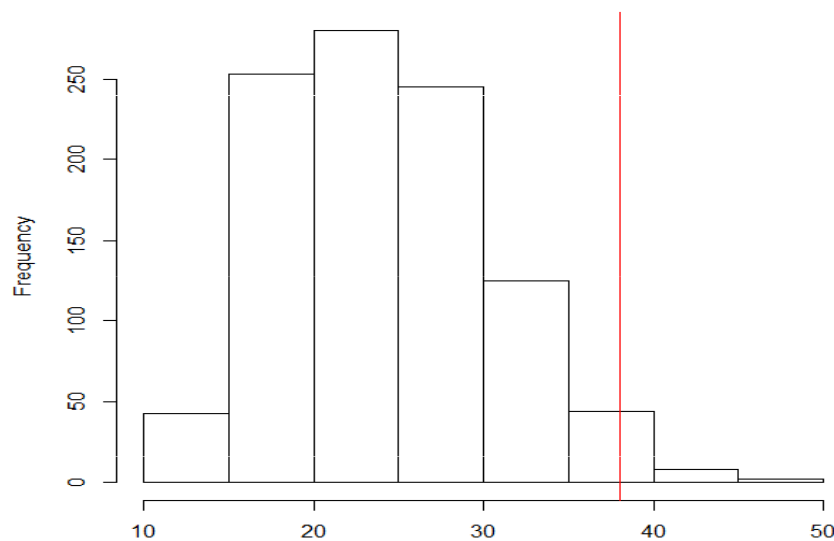
---

- 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 – topologie 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.



- $p = \sum_{i=1}^N (s_{\text{náhodne}} \geq s_{\text{pozorované}}) / N$
- 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>

**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
  - 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/>