Načrtovanje in analiza podatkov DNA mikromrež z BRB-ArrayTools

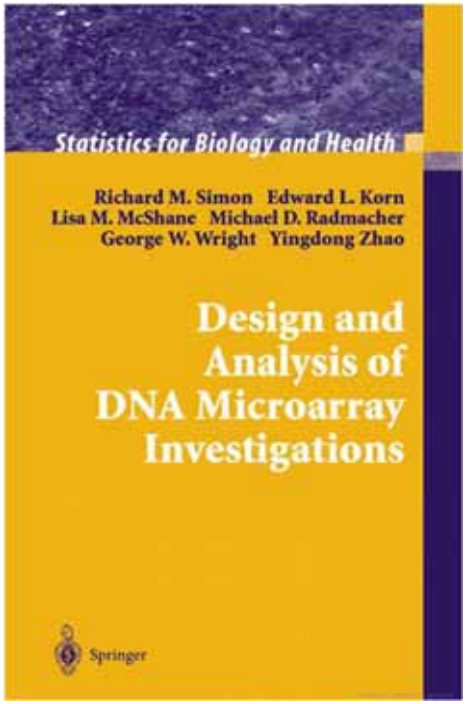
Peter Juvan
UL MF
e-mail: peter.juvan@mf.uni-lj.si

Ljubljana, 31.3.2014

Literatura

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., Zhao, Y.: Design and Analysis of DNA Microarray Investigations. Springer, 2004.

- http://linus.nci.nih.gov/~brb/book.html
- http://books.google.si/books?id=bEJitvcC338C



Program BRB-ArrayTools

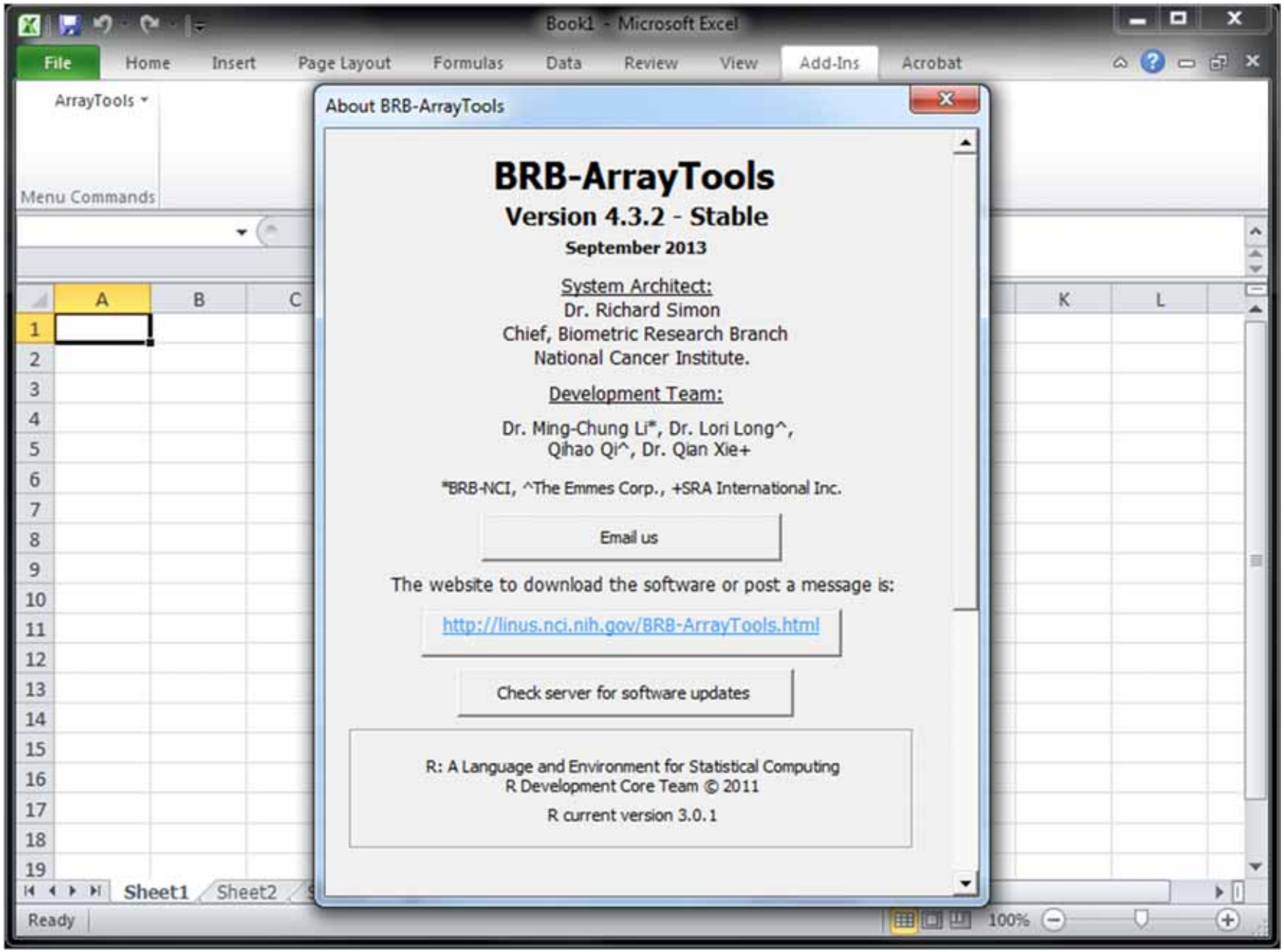
http://linus.nci.nih.gov/BRB-ArrayTools.html

- **analiza** in **vizualizacija** podatkov DNA mikromrež
- dodatek za **Microsoft Excel**
- z Visual Basic for Applications (VBA) povezan z analitičnimi orodji

- **R Project** for Statistical Computing
  http://www.r-project.org/

- **Bioconductor**
  http://www.bioconductor.org/

- programski jeziki C, Fortran in Java


Book1 - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-Ins
Acrobat
ArrayTools
Menu Commands
About BRB-ArrayTools
BRB-ArrayTools
Version 4.3.2 - Stable
September 2013
System Architect:
Dr. Richard Simon
Chief, Biometric Research Branch
National Cancer Institute.
Development Team:
Dr. Ming-Chung Li*, Dr. Lori Long^,
Qihao Qi^, Dr. Qian Xie+
*BRB-NCI, ^The Emmes Corp., +SRA International Inc.
Email us
The website to download the software or post a message is:
http://linus.nci.nih.gov/BRB-ArrayTools.html
Check server for software updates
R: A Language and Environment for Statistical Computing
R Development Core Team © 2011
R current version 3.0.1
Sheet1
Sheet2
Ready
100%

Poskus z DNA mikromrežami

Na primeru analize izražanja genov

Analiza izražanja genov

- nivo **RNA**
- detekcija **razlik** v količini mRNA **med vzorci**
  - intenziteta signala je v sorazmerna količini mRNA

- izražanje = ekspresija
  - **relativna** vrednost (vzorec VS referenca)
  - **logaritmska** transformacija

$$e = log_2(I_{vzorec}/I_{referenca})$$

Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

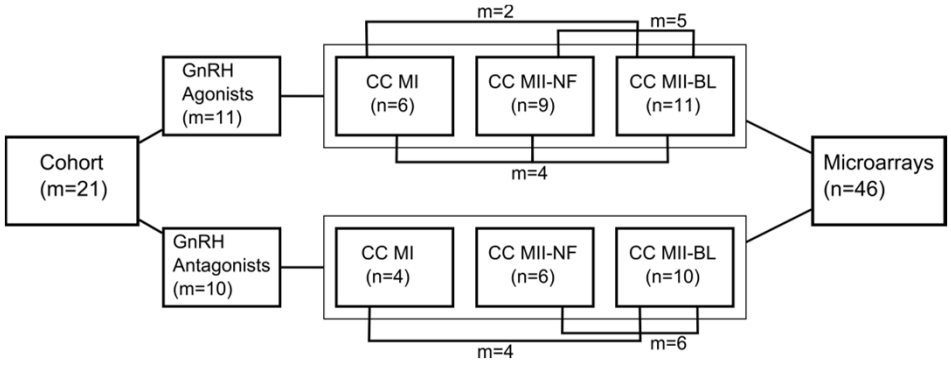
Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Načrtovanje poskusa

- opredelitev **biološkega vprašanja**
  - identifikacija neodvisnih spremenljivk
  - izbira vzorca

- določitev **ciljev**
  - primerjava razredov
  - napovedovanje razreda
  - oblikovanje razredov



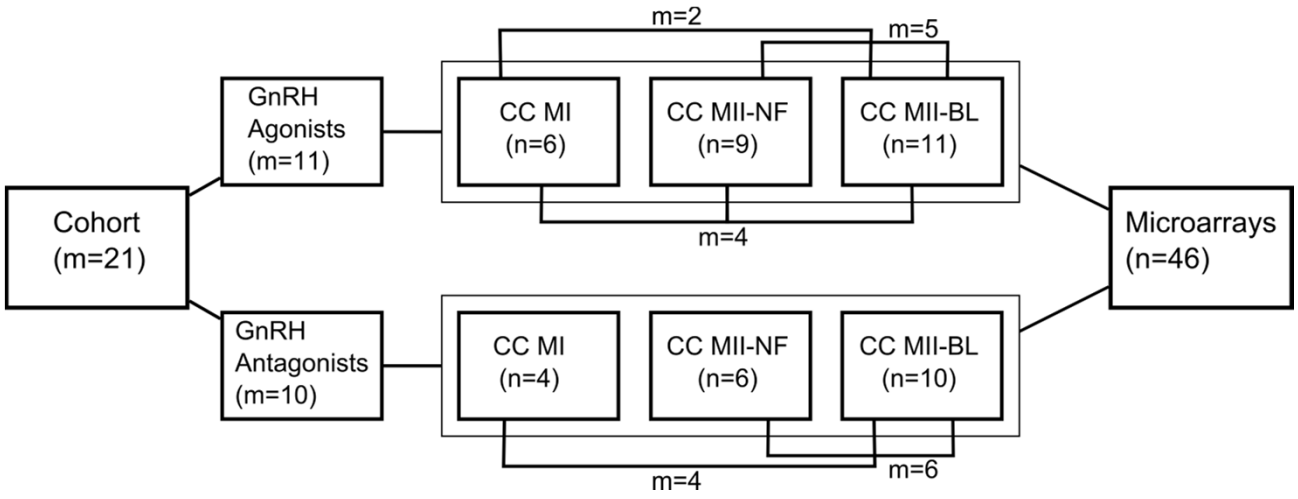
- izbira **metod** za analizo podatkov
  - specifične za uporabljeno platformo mikromrež
  - splošne (statistika, strojno učenje)

OPEN ACCESS Freely available online

PLOS | ONE

Cumulus Cells Gene Expression Profiling in Terms of Oocyte Maturity in Controlled Ovarian Hyperstimulation Using GnRH Agonist or GnRH Antagonist

Rok Devjak[1], Klementina Fon Tacer[2,3], Peter Juvan[2], Irma Virant Klun[1], Damjana Rozman[2], Eda Vrtačnik Bokal[1]*



Experimental design. Number of **patients** (**m=21**) and **samples** (**n=46**) included in the study. 11 patients were administered **GnRH agonist** and 10 **GnRH antagonist**. CC from MI, MII-NF and MII-BL oocytes were collected from 4 GnRH agonists treated patients. CC from MII-NF and MII-BL oocytes were collected from 5 GnRH agonist and 6 GnRH antagonist treated patients; and CC from MI and MII-BL were collected from 2 GnRH agonist and 4 GnRH antagonist treated patients. Altogether, **10** CC samples from **MI** oocytes, **15** from **MII-NF** oocytes and **21** from **MII-BL** oocytes were collected and considered in transcriptome analysis.

Devjak R, Tacer K, Juvan P, et al. (2012) Cumulus Cells Gene Expression Profiling in Terms of Oocyte Maturity in Controlled Ovarian Hyperstimulation Using GnRH Agonist or GnRH Antagonist. *PLoS ONE* 7(10).

Viri variabilnosti

- **razlike v izražanju** genov so posledica
  - razlik med **pogoji poskusa**
  - razlik med **osebki**
  - razlik med **vzorci** in izolirano RNA
  - razlik med DNA **čipi**
  - razlik med **probami** znotraj čipa

  } tehnična variabilnost
- neželene razlike odpravimo s **ponovitvami** meritev
  - **biološke ponovitve**
    - več osebkov pod enakimi pogoji poskusa
  - tehnične ponovitve
    - ponovitve meritev na istih vzorcih
- **združevanje** vzorcev (pooling)
  - zgolj v primeru premajhne količine RNA

Eno/dvo-barvne mikromreže

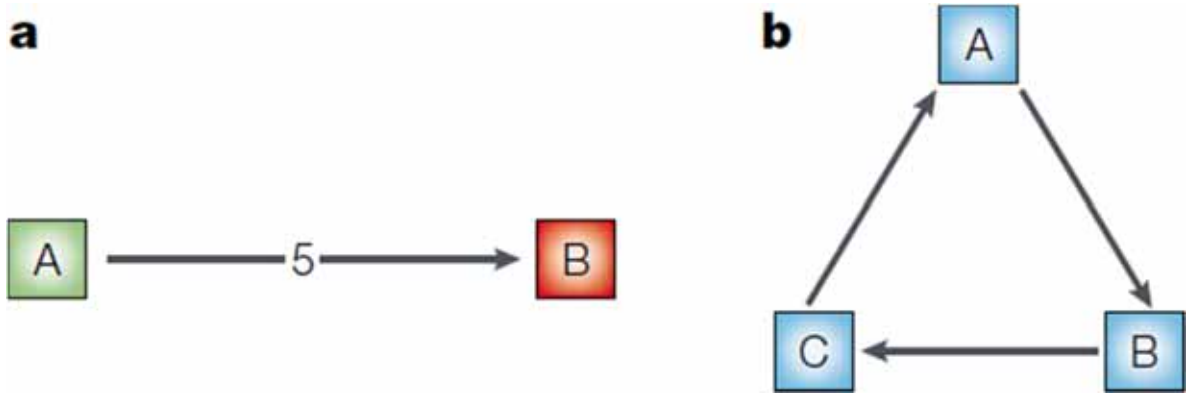
- **enobarvne**
  - vsak vzorec na svoj DNA čip
  - npr. Affymetrix GeneChip

- **dvobarvne**
  - 2 vzorca, označena z različnimi barvili (**Cy3**, **Cy5**)
  - zmanjšamo problem razlik med mikromrežami
  - dodatna tehnična variabilnost zaradi razlik med barvili
  - problem izbire parov vzorcev, ki jih hibridiziramo skupaj

- večbarvne
  - redkeje v uporabi

Združevanje vzorcev pri dvobarvnih mikromrežah

- načrt poskusa predstavimo z grafom
  - vozlišča: vzorci
  - povezave: hibridizacije
  - oznake povezav: št. ponovitev (hibridizacij)
  - določimo barvila, npr.
    - **Cy3** pri izvoru
    - **Cy5** pri ponoru



Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nature reviews. Genetics* 2002; 3(8): 579-588.

neposredna / posredna primerjava

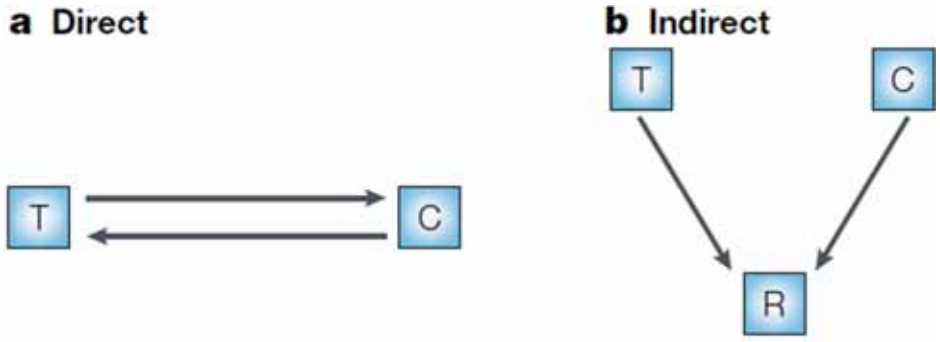
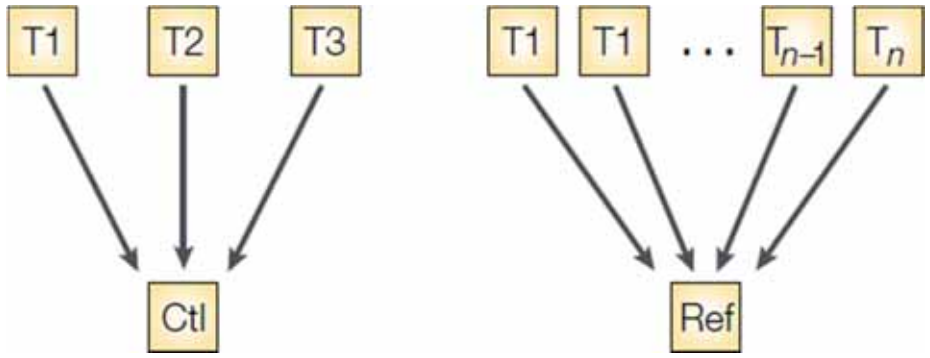


Figure 1 | **Direct versus indirect designs.** Two possible designs that compare gene expression in two cell-population samples T and C. **a** | In a direct comparison, the differential expression of the genes in samples T and C is measured directly on the same slide (in a single experiment). **b** | In an indirect comparison, expression levels of samples T and C are measured separately on two different slides. The log ratio $\log_2(T/C)$ is estimated by the difference $\log_2(T/R) - \log_2(C/R)$. R, reference.

Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nature reviews. Genetics* 2002; 3(8): 579-588.

skupna referenca

- ciljni vzorec hibridiziramo skupaj z enotnim **referenčnim** vzorcem
- **najpogosteje** uporabljen
- preprost, možnost širitve poskusa
- problem izbire **referenčnega vzorca**
  - količina
  - izraženost genov
- priprava reference
  - kontrola
  - univerzalna (komercialna)
  - pool vseh vzorcev (problem širitve poskusov)
- **slaba izkoriščenost**: ½ hibridizacij odpade na referenčni vzorec



Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nature reviews. Genetics* 2002; 3(8): 579-588.

krožna zasnova

- primerna za **manjše št. vzorcev**
- več vzorcev -> več indirektnih primerjav
- neuspešna hibridizacija prekine krog -> vse primerjave niso možne

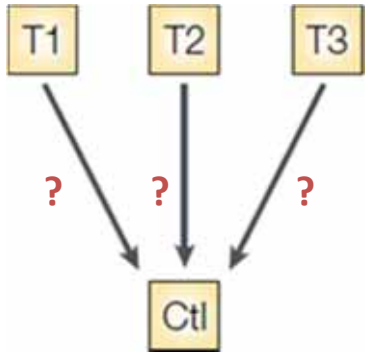
menjava barvil (angl. *dye swap*)

- odpravlja razlike v intenziteti barvil (Cy3, Cy5, ...)
- **neučinkovite** (tehnične) ponovitve
  - bolje: več vzorcev (bioloških ponovitev)

Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nature reviews. Genetics* 2002; 3(8): 579-588.

Št. bioloških ponovitev

- lahko **ocenimo**
  - za 2 razreda in specifične načrte
  - z ozirom na FP, FN, std. dev, pričakovanih razlik v ekspresiji
- težje oceniti pri več faktorjih / razredih
- **v praksi** pogojeno z
  - biološkim sistemom (človek, miš, kvasovka, ...)
  - pogoji poskusa
  - *financami ?!*



Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

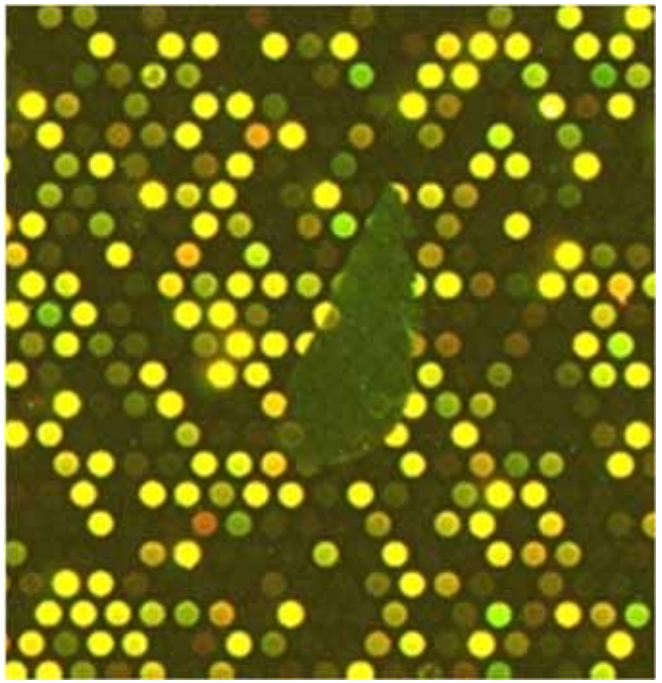
Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Kontrola kvalitete

- obravnavanje šuma
  - **izračunaj** mero kvalitete
  - **izloči** meritev, če mera pod pragom
  - **obravnavaj/imputiraj** izločene vrednosti
- kontrola kvalitete na **nivoju**
  1. sond
  2. genov
  3. čipov
- **komercialni** čipi bistveno bolj kvalitetni od „**doma narejenih**“
  - nanos cDNA/oligonukleotidov z robotom
  - sinteza oligonukleotidov *„in situ“*

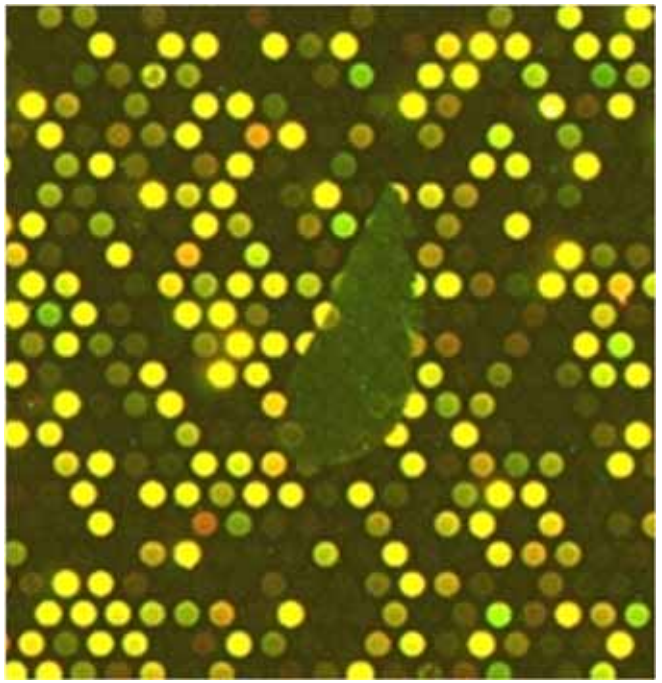
1. Nivo sond



- nizka kvaliteta kot posledica
  - **tehnične napake**
    - pri izdelavi DNA mikromreže
    - nečistoče pri hibridizaciji
    - pomanjkljivo spiranje po hibridizaciji
  - **šibkega signala**
    - nizek SNR (angl. *signal-to-noise ratio*)
- **vizualni** pregled slike
  - umazanija -> visok signal
  - zračni mehurček med hibridizacijo -> nizek signal
  - meglice -> nizko razmerje med signalom in ozadjem
- št. slikovnih točk (px)
  - min. 25-50 px (radij = 3-4 px)
  - ustrezna **resolucija** skeniranja

izločimo sonde, pri katerih:

- **nizek signal v R in G hkrati**
  - zaradi visokega vpliva **aditivnega** šuma
  - npr: FR=FG, FR±10, FG±10
    - e = log2(510/490) = 0.057
    - e = log2(30/10) = 1.58



- visok signal **ozadja**
  - prispevek ozadja ni zgolj aditiven
  - izloči, če FR/BR < 1.5 ali FG/BG < 1.5

- **korelacija** med slikovnimi točkami FR in FG
  - nizka -> izloči

Wang X, Ghosh S, Guo SW: Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Res 2001; 29(15): E75-75.

2. Nivo genov

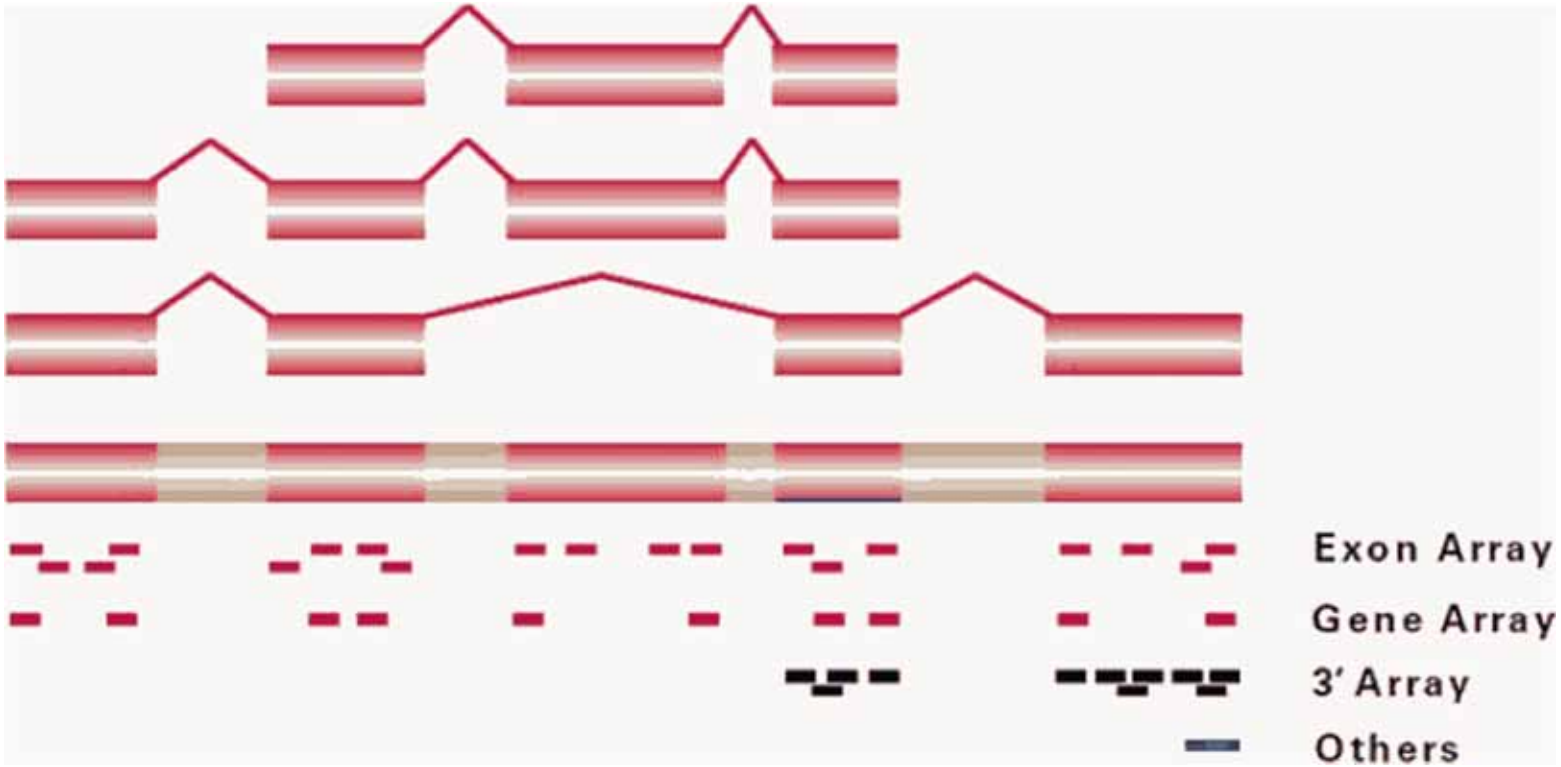
- običajno **več sond** za en gen
  - podvojene sonde
  - različne sonde za isti gen
- **razlike** med sondami so **posledica**
  - alternativnih spojitvenih variant
  - napačnih anotacij sond
  - razlik pri hibridizaciji
  - napak pri izdelavi mikromrež

- **testiramo** razlike in **izločimo** sonde **pri vseh čipih hkrati**
  - razlike med podvojenimi sondami (tehnične narave)
  - razlike med različnimi sondami za isti gen (biološke in tehnične narave)

- izločimo lahko tudi **nizko variabilne gene** pod različnimi eksperimentalnimi pogoji
  - njihovo izražanje je lahko zgolj posledica šuma

Affymetrix

- 3‘ ekspresijski čipi
- gene čipi
- eksonski čipi



3. Nivo čipov

- posledica
  - slabše izdelave **čipov**
  - razgrajene **RNA**
  - napak pri **označevanju, hibridizaciji, skeniranju**

- **odstranimo** podatke celega čipa, če
  1. izločenih veliko sond
  2. slabo razmerje med ospredjem in ozadjem
  3. nizka varianca pri posameznem kanalu (Cy3, Cy5)
  4. veliko saturiranih sond (ponovimo skeniranje)
  5. neustrezne interne kontrole (spikes)

Imputacija manjkajočih vrednosti sond

- oceni manjkajoče/izločene vrednosti na podlagi preostalih vrednosti

- možnosti
  - mediana izražanja gena preko vseh čipov
  - z upoštevanjem korelacije med geni

Obravnava manjkajočih vrednosti

- metode **prilagojene** za obravnavo manjkajočih vrednosti

- npr. test diferenčne izraženosti genov
  - **linearna regresija** namesto klasične analize variance (ANOVA)

Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Normalizacija podatkov

- odstranimo razlike, ki so **tehnične** narave
  - nastavitev PMT pri laserskem odčitavanju
  - količina hibridizirane RNA
  - učinkovitost vgradnje barvil
- na **nivoju**
  - blokov
  - čipov
  - poskusa
- **pristop**
  - izbira normalizacijskih **sond**
  - izbira norm. **algoritma**

Izbira normalizacijskih sond

- **hišni geni**
  - njihovo izražanje se ne spreminja
    - vključeni v osnovne celične funkcije
    - preko pogojev poskusa (ocenimo iz podatkov – pristranost !!)
  - ne obstajajo ?!
- **zunanja RNA (spikes)**
  - vsem vzorcem dodamo zunanjo RNA
  - problem količine zunanje RNA
    - uskladiti s količino RNA pri obravnavanih vzorcih
    - napaka pipete
  - uporabno za **po meri narejene / diagnostične** čipe
- **vsi geni**
  - predpostavka: skupna količina RNA je pri vseh vzorcih konstantna
    - velja le za čipe, ki pokrivajo **celoten genom**
    - ne velja za po meri narejene / diagnostične čipe
  - najpogosteje uporabljeno

Normalizacija dvobarvnih čipov

- za vsak čip ***j*** in gen ***k*** izračunamo normalizacijski faktor

$$C_{jk}$$

- prilagodimo izraženost genov

$$x_{jk} = log_2\left(\frac{R_{jk}}{G_{jk}}\right) - C_{jk}$$

- pristopi
  - globalen pristop (skupen C_j za vse gene)

$$C_j = \underset{k \in G}{\text{median}}\left(log_2\left(\frac{R_{jk}}{G_{jk}}\right)\right)$$

  - C_{jk} odvisen od povprečne intenzitete

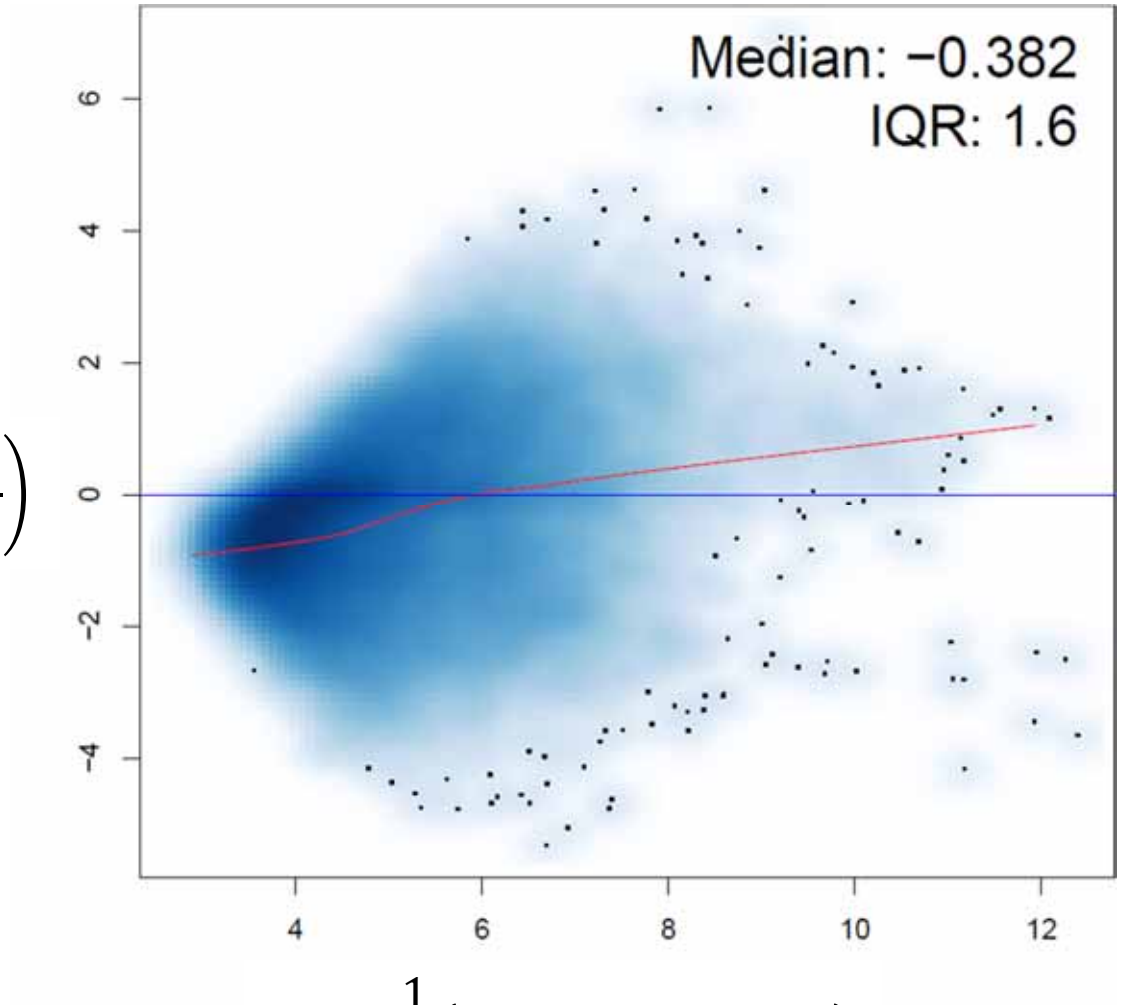
$$A_{jk} = \frac{1}{2}\left(log_2 R_{jk} + log_2 G_{jk}\right)$$

  - C_{jk} odvisen od lokacije na čipu (po blokih)
  - C_{jk} odvisen od povprečne intenzitete in lokacije na čipu

C_{jk} odvisen od povprečne intenzitete vseh genov

- normalizacija mikromrež z geni iz celotnega genoma
- **MA graf**

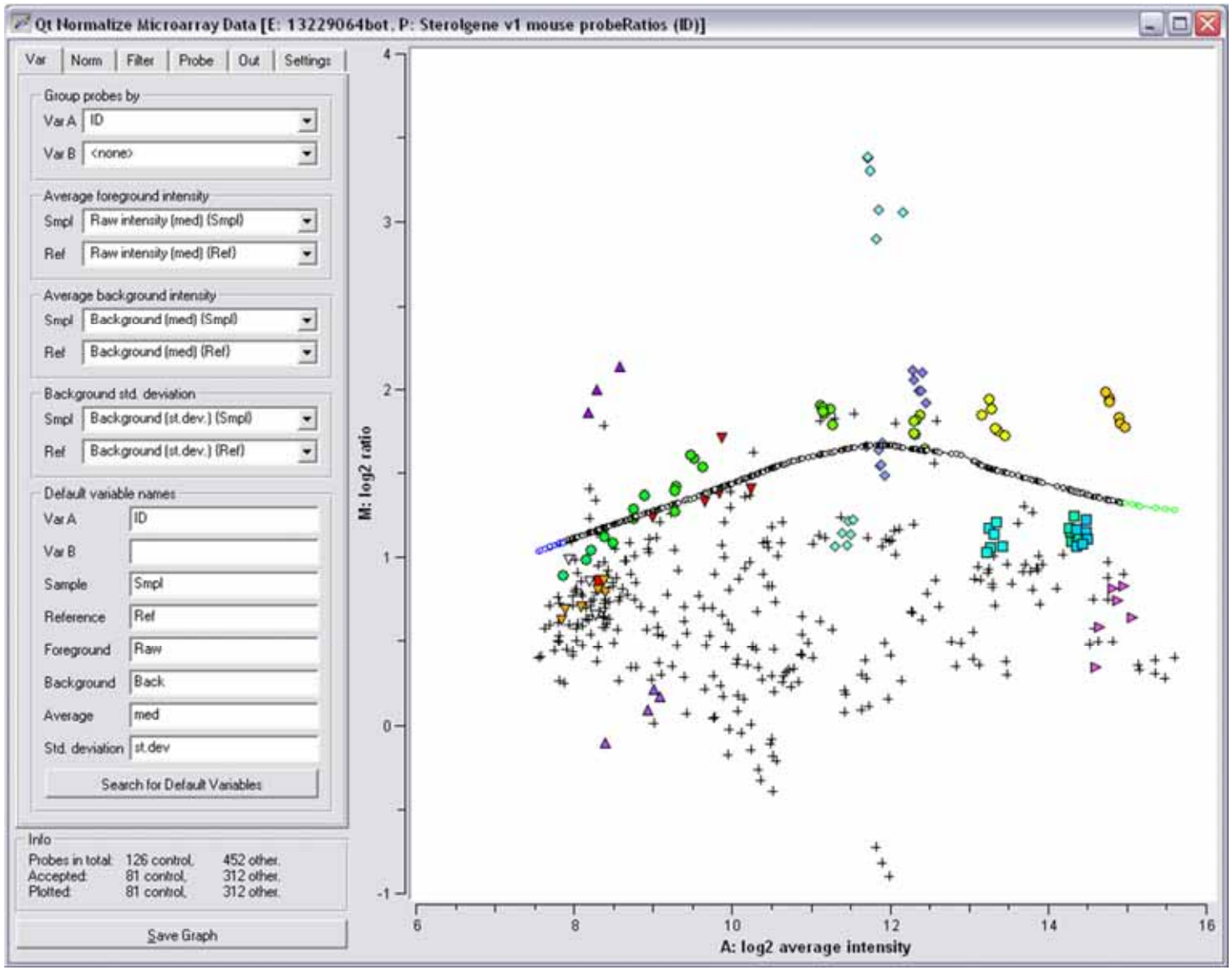
$$M_{jk} = log_2\left(\frac{R_{jk}}{G_{jk}}\right)$$



$$A_{jk} = \frac{1}{2}\left(log_2 R_{jk} + log_2 G_{jk}\right)$$

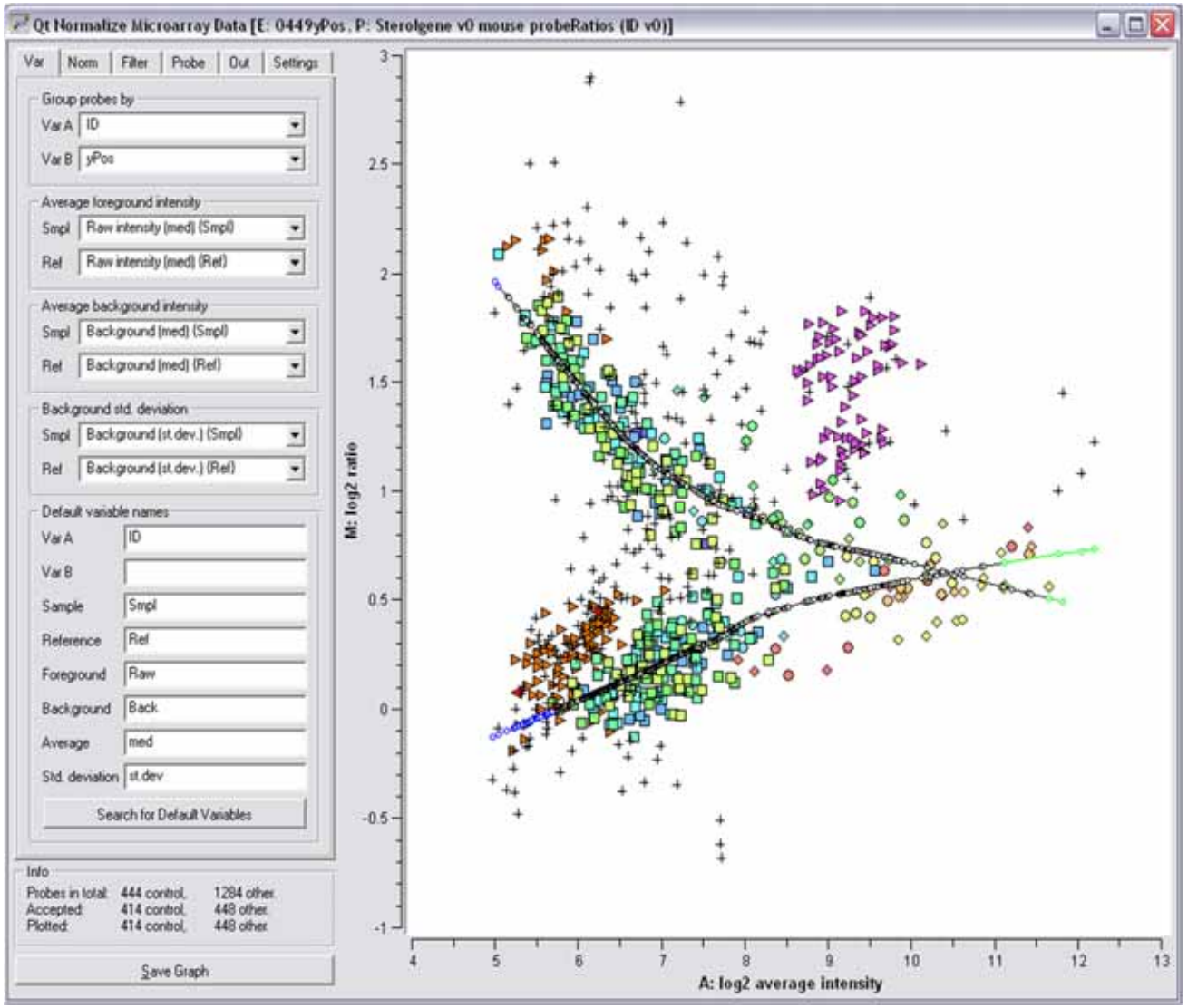
C_{jk} odvisen od povprečne intenzitete „spikes“ sond

– normalizacija po meri narejene mikromreže Steroltalk v1



C_{jk} odvisen od povprečne intenzitete in lokacije „spikes“ sond

– normalizacija po meri narejene mikromreže Sterolgene v0



Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Primerjava razredov

- cilja
  - izbira **diferenčno izraženih** **genov** med razredi
  - analiza **obogatenosti** **genskih skupin/poti**

- razrede določimo **v naprej** (neodvisno od rezultatov meritev)
  - rakaste ~ normalne celice
- sočasne primerjave **več razredov**
  - rakaste ~ normalne celice
  - pred ~ po administraciji zdravila
- **večnivojski** razredi
  - rak tipa1 ~ rak tipa 2 ~ normalno tkivo

Primerjava razredov

1. **Test diferenčne izraženosti posameznih genov**
    1. Studentov t-test
    2. permutacijski t-test
    3. Wilcoxon rang-sum test
    4. več razredov
    5. parni vzorci

2. **Identifikacija diferenčno izraženih genov**
    1. Bonferronijev popravek
    2. permutacijska metoda
    3. nadzor števila lažno pozitivnih (FP)
    4. nadzor FDR (False Discovery Rate)

3. **Regresijski modeli**

Primerjava razredov

1. **Test diferenčne izraženosti posameznih genov**
   1. Studentov t-test
   2. permutacijski t-test
   3. Wilcoxon rang-sum test
   4. več razredov
   5. parni vzorci

2. **Identifikacija diferenčno izraženih genov**
   1. Bonferronijev popravek
   2. permutacijska metoda
   3. nadzor števila lažno pozitivnih (FP)
   4. nadzor FDR (False Discovery Rate)

3. **Regresijski modeli**

1. Test diferenčne izraženosti posameznih genov

- geni z znanimi funkcijami -> razumevanje bioloških mehanizmov
- geni z neznanimi funkcijami -> napovedovanje novih funkcij

- podatki
  - 2 razreda z J_1 in J_2 vzorci (biološkimi replikami)
  - izražanje gena *x*

$$\overline{x_1} = x_{11}, x_{12}, \ldots x_{1J_1}$$
$$\overline{x_2} = x_{21}, x_{22}, \ldots x_{2J_2}$$

- **razlika srednjih vrednosti izražanja**
  - ni zanesljiva ocena diferenčne izraženosti !!

$$|\overline{x_1} - \overline{x_2}| \geq 1$$

- **Studentov t-test**
  - H_0: porazdelitev izražanja pri razredih J_1 in J_2 je v populaciji enaka
  - ocenimo verjetnost, da je razlika v izražanju taka, kot smo jo izmerili

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{{s_p}^2 \left(\frac{1}{J_1} + \frac{1}{J_2}\right)}}$$

  - kjer je ${s_p}^2$ **združena** (angl. *pooled*) varianca izražanja, pri čemer predpostavimo, da je znotraj razredov enaka

$${s_p}^2 = \frac{(J_1 - 1){s_1}^2 + (J_2 - 1){s_2}^2}{J_1 + J_2 - 2}$$

$${s_i}^2 = \frac{1}{J_i - 1} \sum_{j=1}^{J_i} \left(x_{ij} - \overline{x_i}\right)^2 \; za \; i = 1{,}2$$

  - iz *t*-statistike po tabelah izračunamo ***p-vrednost***
  - *p-vrednost* primerjamo z **izbrano stopnjo značilnosti** (običajno $\boldsymbol{\alpha = 0.05}$)
  - če je $\boldsymbol{p<\alpha}$, zavrnemo ničelno hipotezo H_0 in sklepamo, da je **gen diferenčno izražen**

- **permutacijski t-test**
  - ocenimo, koliko lahko zaupamo *t-statistiki*
  - ne predpostavimo normalnosti porazdelitve izražanja

algoritem

1. izračunamo *t*-statistiko (t)
2. ponavljamo
   1. naključno označimo meritve z razredoma J_1 in J_2
   2. izračunamo *t*-statistiko (t^*)
3. izračunamo *p-vrednost* in primerjamo z izbrano stopnjo značilnosti α

$$p = \frac{1 + \#(|t^*| \geq |t|)}{1 + \#_P}$$

- pogosto lahko pregledamo vseh $\binom{J_1 + J_2}{J_1}$ možnosti; v tem primeru velja

$$p = \frac{\#(|t^*| \geq |t|)}{\binom{J_1 + J_2}{J_1}}$$

	Class 1 data values					Class 2 data values				Parametric *t*-statistic
original data:	−0.18	−0.10	−0.13	0.30	−0.14	0.15	0.84	0.66	0.52	$t = 3.64$
data permutation 1:	−0.18	−0.10	−0.13	0.30	−0.14	0.15	0.84	0.66	0.52	$t^* = 3.64$
data permutation 2:	−0.18	−0.10	−0.13	0.30	0.15	−0.14	0.84	0.66	0.52	$t^* = 2.15$
data permutation 3:	−0.18	−0.10	−0.13	0.15	0.84	0.30	−0.14	0.66	0.52	$t^* = 0.83$
data permutation 4:	−0.18	−0.10	−0.13	−0.14	0.15	0.30	0.84	0.66	0.52	$t^* = 5.48$
⋮			⋮				⋮			⋮
data permutation 124:	0.30	−0.14	0.84	0.66	0.52	−0.18	−0.10	−0.13	0.15	$t^* = -2.48$
data permutation 125:	0.30	0.15	0.84	0.66	0.52	−0.18	−0.10	−0.13	−0.14	$t^* = -4.49$
data permutation 126:	−0.14	0.15	0.84	0.66	0.52	−0.18	−0.10	−0.13	0.30	$t^* = -2.48$

$$\text{permutation } p\text{-value} = \frac{\#\text{ of the 126 permuations where } |t^*| \geq |t|}{126} = \frac{3}{126}$$

Simon RM *et al*. Design and Analysis of DNA Microarray Investigations. Springer, 2004, p. 70.

- **Wilcoxon rang-sum test**
  - neparametrična alternativa t-testa

  algoritem

  1. zamenjamo $J_1 + J_2$ vrednost izražanja z rangi
  2. izračunamo vsoto rangov za razreda 1 in 2
  3. izračunamo *p-vrednost* po tabelah in primerjamo z izbrano stopnjo značilnosti α

  - manjša občutljivost

    $$J_1 = 2, J_2 = 3: \#_p = 10, p_{min} = 0.1$$

  + hiter izračun

- **več razredov**
  - $J_1, J_2 \ldots J_K$ vzorcev za ***K*** razredov
  - H_0: porazdelitev izražanja je enaka za vse razrede
  - alternativna hipoteza: **vsaj pri enem** izmed razredov se porazdelitev izražanja razlikuje od ostalih razredov

  - *F*-statistika (posplošitev t-testa)
  - permutacijski F-test (posplošitev)
  - neparametričen: Kruskal-Wallis test
  - neparametričen, urejeni razredi: Jonckhreere test

  - **post hoc** primerjave
  - primerjave razredov po parih: $\left(\frac{(K-1)*K}{2}\right)$ parov

- **parni vzorci**
  - izražanje gena x izmerjeno v parih
    npr: pri isti osebi pred/po administraciji zdravila
    $$(x_{11}, x_{21}), (x_{12}, , x_{22}), \dots (x_{1J}, x_{2J})$$
  - parni t-test
    $$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{{s_s}^2/J}}$$
    $${s_d}^2 = \frac{1}{J-1}\sum_{j=1}^{J}\left[(x_{1j} - x_{2j}) - (\overline{x_1} - \overline{x_2})\right]^2$$
  - permutacijski parni t-test
    - permutacije izvedemo z upoštevanjem parnosti: oznake razredov permutiramo za vsak par posebej
    - možnih permutacij: 2^J
  - Wilcoxon signed-rank test
    - rangiramo razlike po parih: $d_1 = x_{11} - x_{21}$

Primerjava razredov

1. **Test diferenčne izraženosti posameznih genov**
   1. Studentov t-test
   2. permutacijski t-test
   3. Wilcoxon rang-sum test
   4. več razredov
   5. parni vzorci

2. **Identifikacija diferenčno izraženih genov**
   1. Bonferronijev popravek
   2. permutacijska metoda
   3. nadzor števila lažno pozitivnih (FP)
   4. nadzor FDR (False Discovery Rate)

3. **Regresijski modeli**

2. Identifikacija diferenčno izraženih genov

- testiramo veliko število hipotez (genov)
- velika možnost **lažno pozitivnih** (FP) zaključkov
  - če pri 30.000 genih dovolimo 5% napako
  - v povprečju dobimo 1.500 lažno pozitivnih rezultatov !!
- **problem multiplih primerjav:** kako nadzorovati št. FP

- **Bonferronijev popravek**
  - *p-vrednosti* posameznih genov pomnožimo s št. genov (K)
  - popravljene *p-vrednost* primerjamo z izbrano stopnjo značilnosti α
  - preveč strikten
    - npr. pri K=10.000 je gen diferenčno izražen pri $\mathrm{p} \leq 0.000005$
  - za tako nizke p-vrednosti
    - Studentov t-test ni dovolj natančen
    - permutacijski test ni dovolj občutljiv

- **permutacijska metoda**

  algoritem

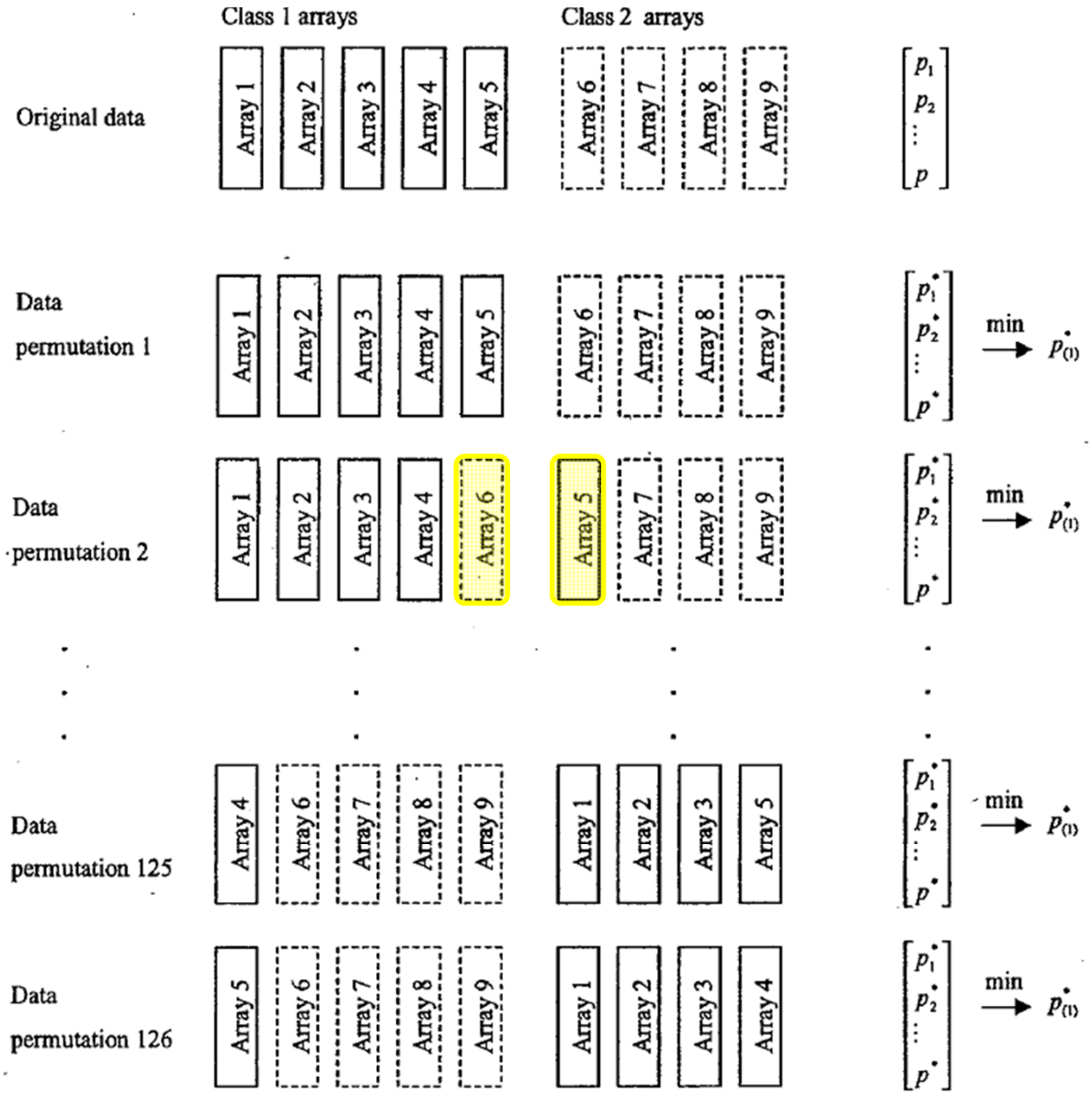
  1. izračunamo *p-vrednosti* genov (p_k)
  2. ponavljamo
     1. permutiramo oznake razredov
     2. izračunamo *p-vrednosti* genov ($p_k{}^*$) in poiščemo minimum ($p_{(1)}{}^*$)
  3. izračunamo popravljene *p-vrednosti* (adj_p_k)

$$adj\_p_k = \frac{1 + \#(p_{(1)}{}^* \leq p_k)}{1 + \#_P}$$

  4. popravljene *p-vrednost* primerjamo z izbrano stopnjo značilnosti α

  \+ manj strikten od Bonferronijevega popravka

  – (pre)pogosto lahko pregledamo vseh $\binom{J_1 + J_2}{J_1}$ možnosti



Simon R.M. et al. Design and Analysis of DNA Microarray Investigations. Springer, 2004.

- **nadzor števila lažno pozitivnih (FP, angl. *False Positive*)**
  - Bonferronijev popravek in permutacijska metoda zagotavljata, da med odkritimi geni ni FP genov
  - posplošimo **permutacijski test**, da dovolimo v povprečju največ ***U*** FP genov

$$adj\_p_k = \frac{1 + \#(p_{(U+1)}{}^* \leq p_k)}{1 + \#_P}$$

- **nadzor stopnje lažnih odkritij (FDR, angl. *False Discovery Rate*) – postopek Benjamini-Hochberg (BH)**
  1. uredimo nepopravljene *p-vrednosti*
  $$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$$
  2. poiščemo najvišji indeks ***D***, pri katerem velja
  $$p_{(D)} * \frac{K}{D} < \alpha$$
  3. izračunamo popravljene *p-vrednosti* (***D*** predstavlja št. odkritih genov)
  $$adj\_p_{(k)} = p_{(k)} * \frac{K}{D}$$

- postopek Tusher *et al.* (**SAM**)
- **permutacijski FDR**

Primerjava razredov

1. **Test diferenčne izraženosti posameznih genov**
   1. Studentov t-test
   2. permutacijski t-test
   3. Wilcoxon rang-sum test
   4. več razredov
   5. parni vzorci

2. **Identifikacija diferenčno izraženih genov**
   1. Bonferronijev popravek
   2. permutacijska metoda
   3. nadzor števila lažno pozitivnih (FP)
   4. nadzor FDR (False Discovery Rate)

3. **Regresijski modeli**

3. Regresijski modeli

$$y_j = \alpha + \beta z_j + e_j$$

- $j = 1,2 \ldots J$ vzorcev
- y: odvisna spremenljivka: izražanje gena
- z: neodvisna spremenljivka (faktor vpliva)
- α: srednja vrednost izražanja gena
- e: napaka
- β: regresijski koeficient

Hipoteza $H_0: \beta = 0$: ni povezave med izražanjem gena in faktorjem z

- **večje število neodvisnih** spremenljivk (faktorjev)

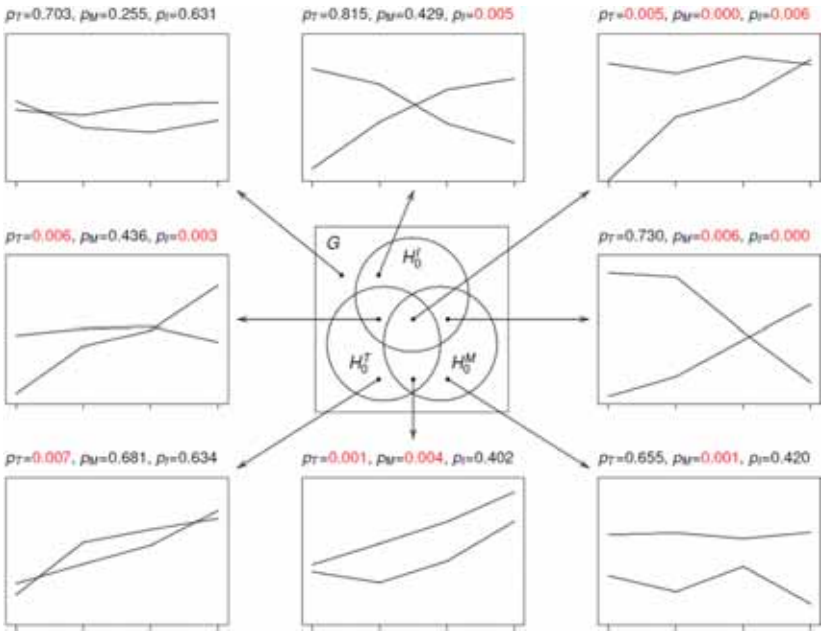
$$y_j = \alpha + \beta_1 z_j + \beta_2 w_j + e_j$$

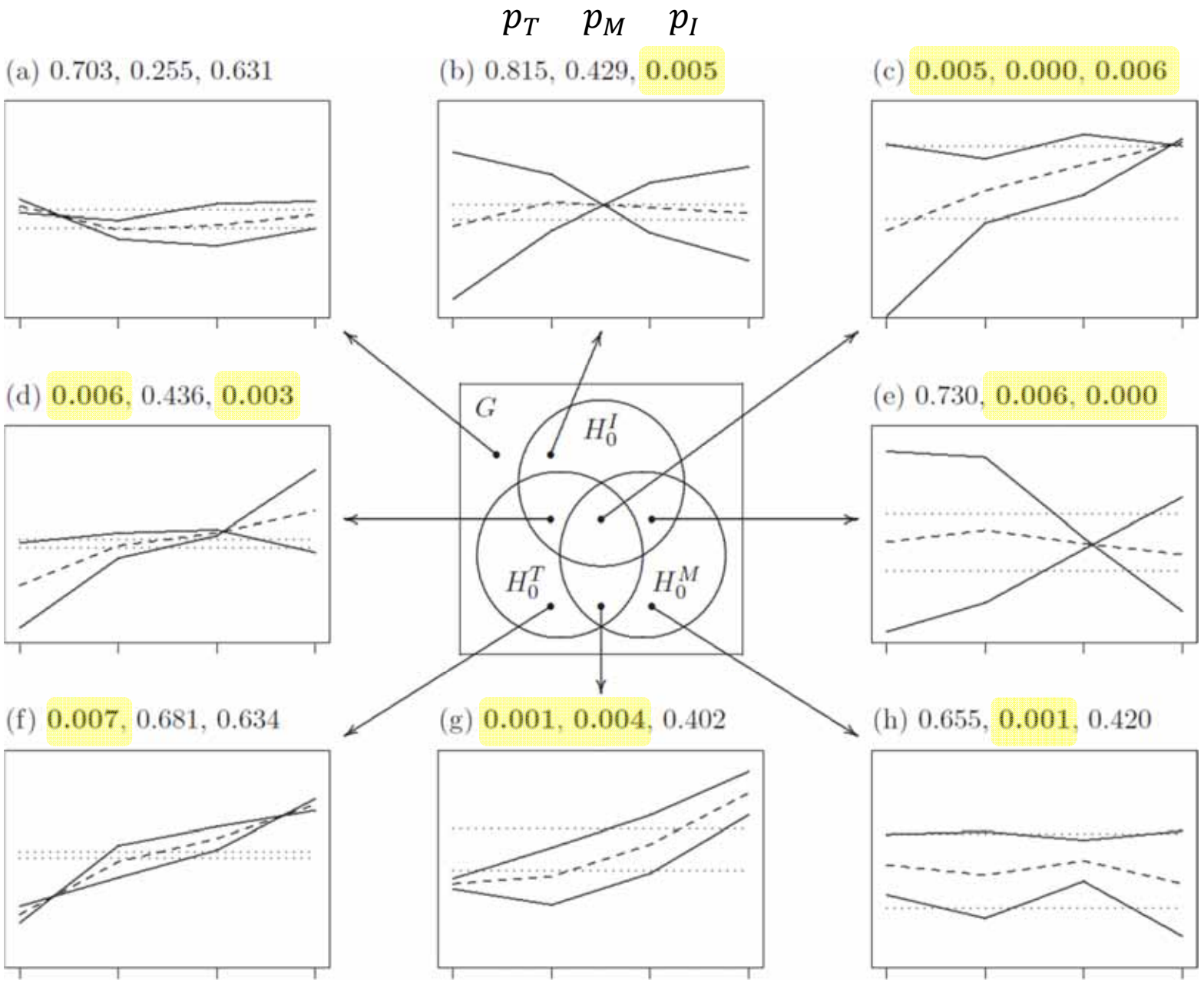
- obravnava motečih dejavnikov (diskretne/zvezne vrednosti)

Regresijski model z dvema faktorjema

$$y_j = \alpha + \beta_T T_j + \beta_M M_j + \beta_I (T_j * M_j) + e_j$$

- ${H_0}^T: \beta_T = 0$: ni povezave med izražanjem gena in časom T
- ${H_0}^M: \beta_M = 0$: ni povezave med izražanjem gena in vplivom mutacije M
- ${H_0}^I: \beta_I = 0$: ni **interakcije** med vplivom časa T in mutacije M





Juvan P. Artificial intelligence methods for discovery of relationships in genetic data. Doctoral Dissertation, University of Ljubljana, 2005.

OPEN ACCESS Freely available online

PLoS one

Novel Insights into the Downstream Pathways and Targets Controlled by Transcription Factors CREM in the Testis

Rok Kosir[1,7], Peter Juvan[1], Martina Perse[2], Tomaz Budefeld[3], Gregor Majdic[3,5], Martina Fink[4], Paolo Sassone-Corsi[6], Damjana Rozman[1]*

1 Center for Functional Genomics and Bio-Chips, Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, **2** Medical Experimental Centre, Institute of Pathology, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, **3** Center for Animal Genomics, Veterinary Faculty, University of Ljubljana, Ljubljana, Slovenia, **4** Department of Haematology, University Medical Center Ljubljana, Ljubljana, Slovenia, **5** Institute of Physiology, Faculty of Medicine, University of Maribor, Maribor, Slovenia, **6** Department of Pharmacology, University of California Irvine, Irvine, California, United States of America, **7** Diagenomi Ltd, Ljubljana, Slovenia

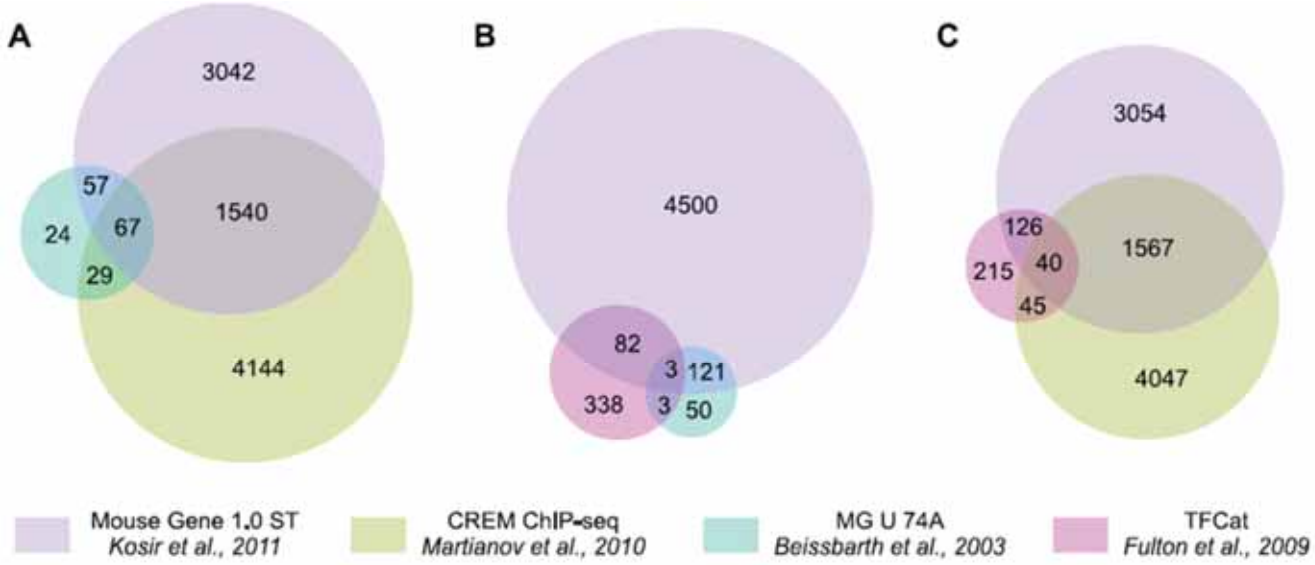


Figure 3. Euler diagrams representing genes from different datasets. Euler diagrams were drawn to visualize the comparisons of genes from different datasets. The size of the circles corresponds to the number of genes present in each dataset. Three comparisons were made in order to retrieve the data for further functional analysis of DE genes.

doi:10.1371/journal.pone.0031798.g003

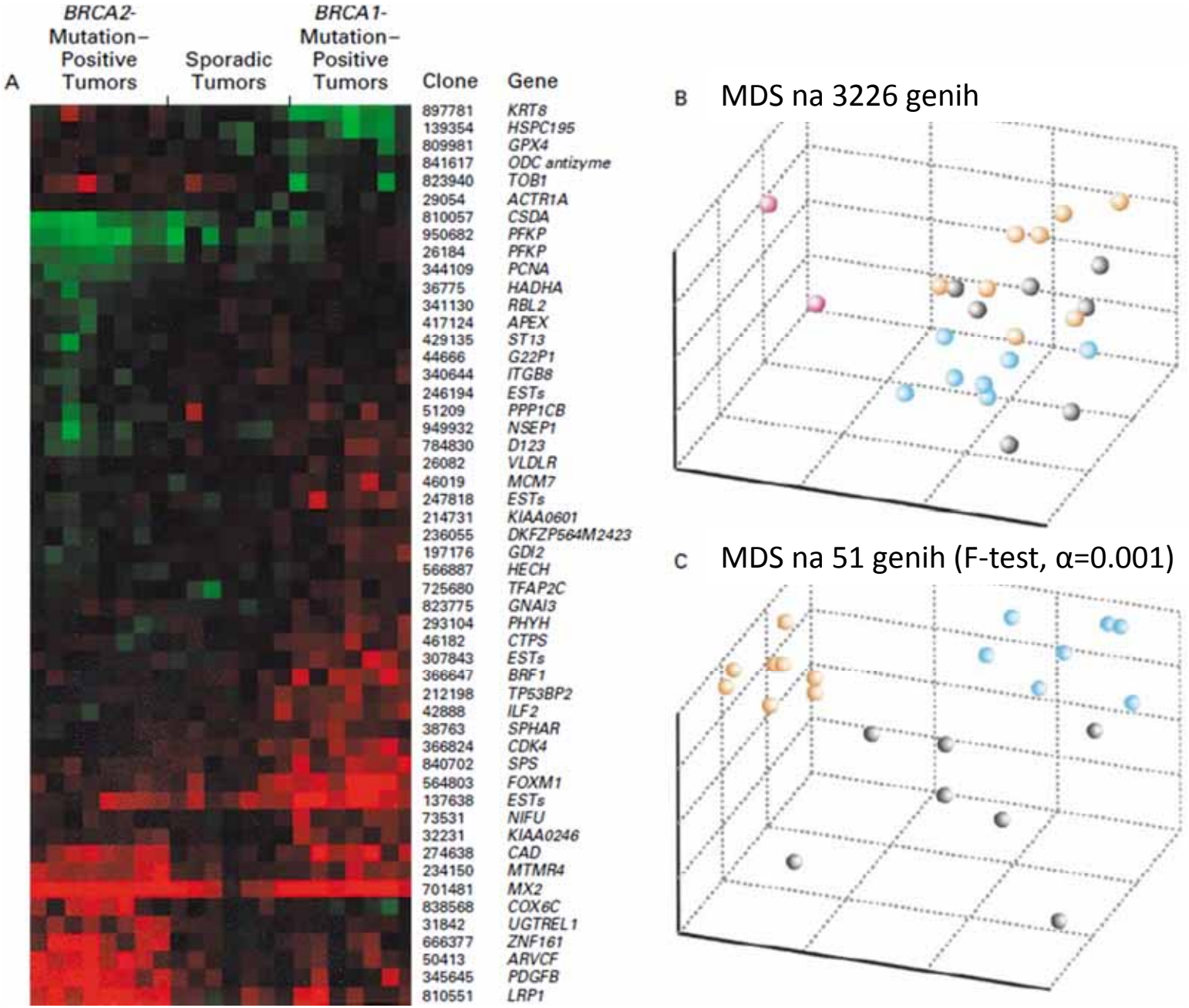
Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Napovedovanje razreda

- cilj: na podlagi izražanja genov **določi razred neznanega** vzorca
  - uporaba: diagnostični čipi

1. priprava **učnih primerov**
   - meritve izražanja genov večjega števila vzorcev s poznanim razredom
2. izdelava **napovednega modela**
   - izbira spremenljivk
   - izbira metode **nadzorovanega** strojnega učenja
   - validacija modela
3. uporaba modela za **napovedovanje** razreda neznanih vzorcev
   - ločevanje tumorskega in normalnega tkiva
   - izbira metode zdravljenja
   - ...



Hedenfalk I *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine* 334(8).

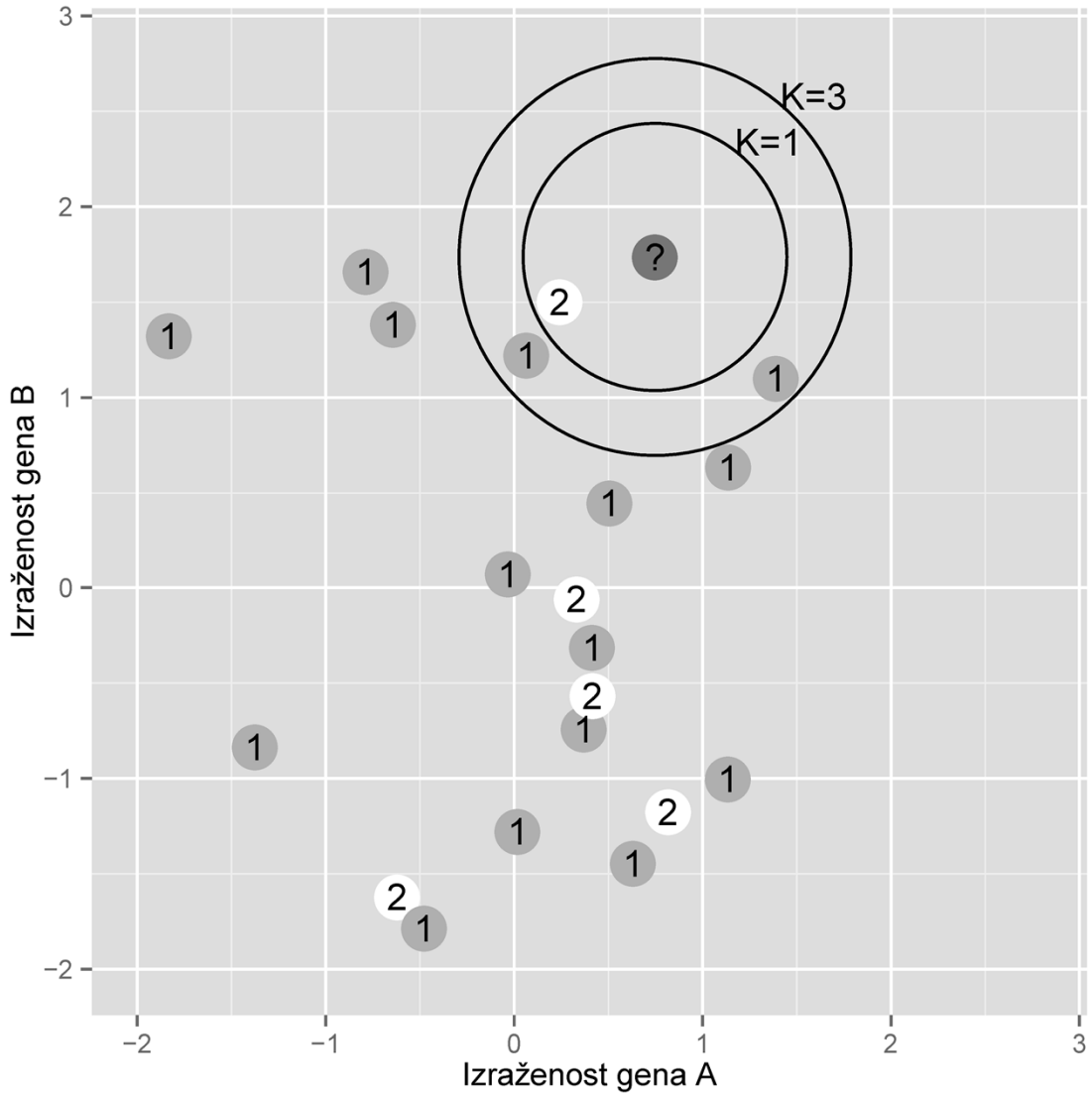
Izbira spremenljivk (genov)

- št. spremenljivk (genov) **<<** števila primerov (vzorcev)
  - problem pretiranega prileganja (angl. *overfitting*)
- diferenčno izraženi geni
  - t-test, F-test, permutacijski in neparametrični testi
  - izbira ustrezne α (0.01, 0.001, ...)
- analiza glavnih komponent (PCA)
  - redukcija dimenzij
  - komponenta: linearna kombinacija genov
  - vključenih veliko genov
  - problem interpretacije modela
- gruče genov

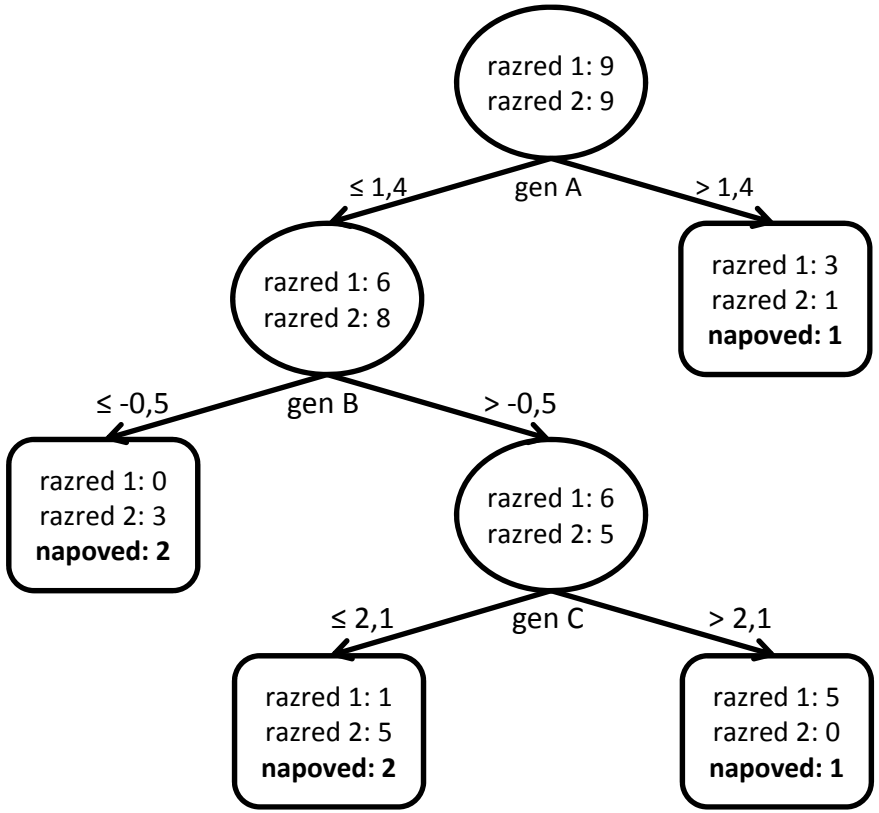
Metode nadzorovanega strojnega učenja

- diskriminantna analiza
  - poišči tako linearno kombinacijo izbranih genov,
  - da si bodo razredi glede na vrednosti linearne kombinacije med seboj čim bolj različni
- metoda K najbližjih sosedov
  - napovej razred na podlagi K primerov,
  - ki so najbolj podobni z ozirom na izraženost izbranih genov
  - običajno K<7
- odločitvena drevesa
- metoda podpornih vektorjev
- ...

Metoda najbližjih sosedov



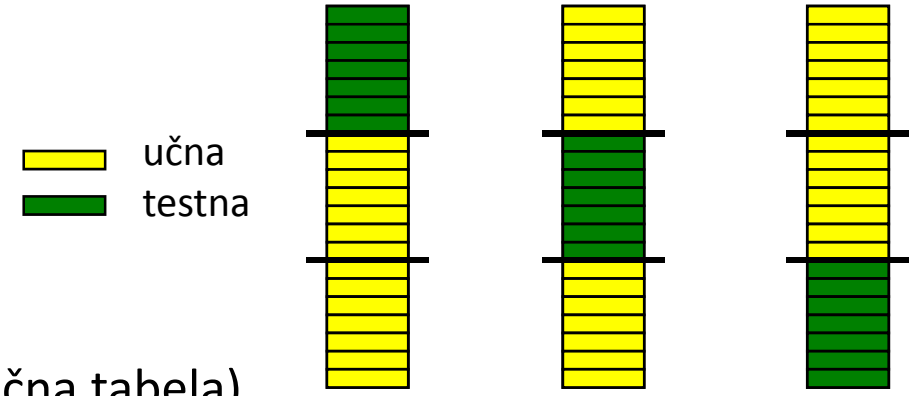
Odločitvena drevesa



Simon RM *et al.* Design and analysis of DNA microarray investigations. Springer, 2003.

Validacija modela: prečno preverjanje

- izbira učne / testne množice



- izračunamo statistiko (kontingenčna tabela)

		Resnični razred +	Resnični razred -	Skupaj
Napovedan razred	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
	Skupaj	TP+FN	FP+TN	N

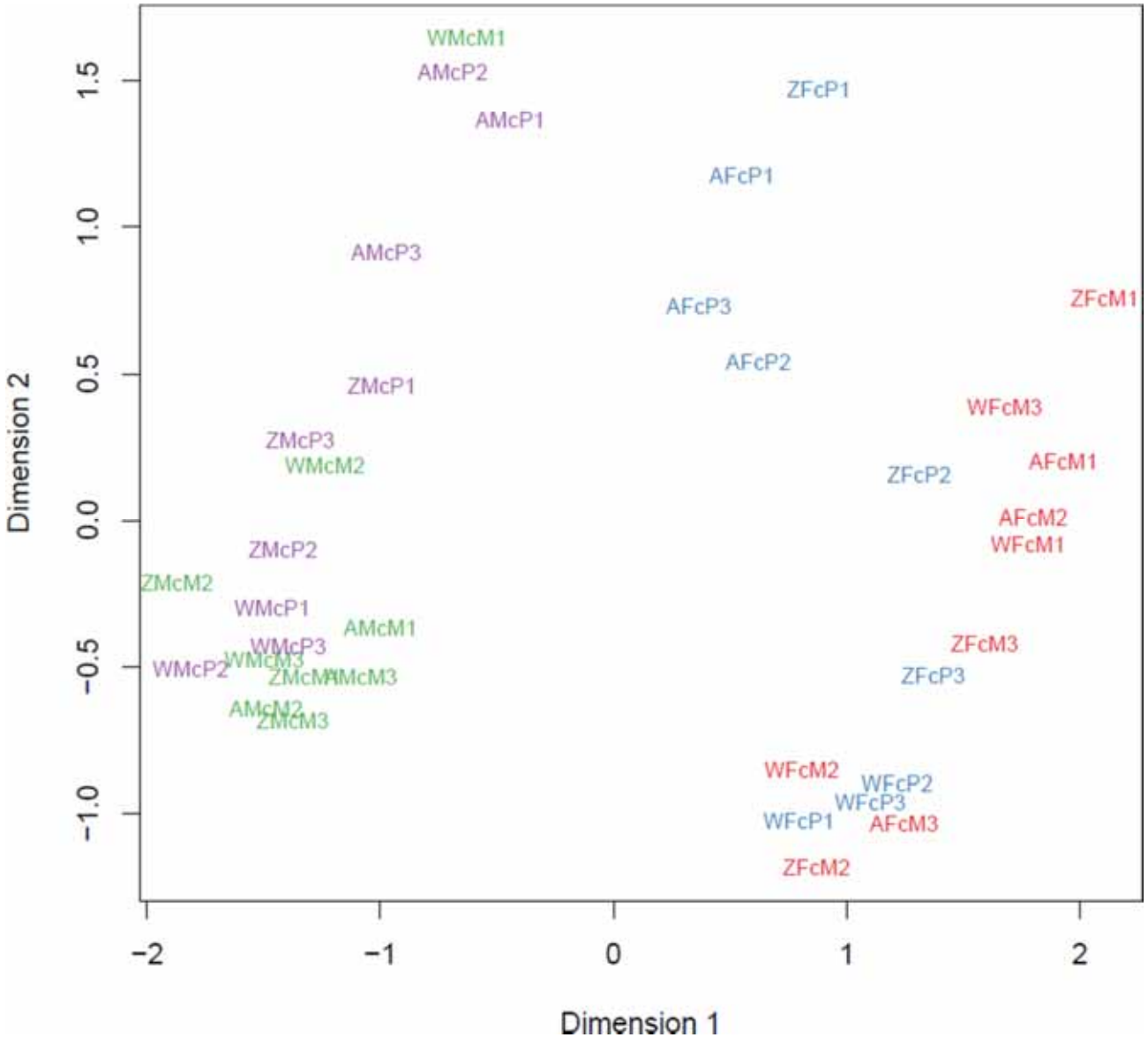
- izračunamo statistike, npr.
  - napovedno točnost: (TP+TN)/N
  - občutljivost: TP/(TP+FN)
  - specifičnost: TN/(FP+TN)

Poskus z DNA mikromrežami

1. Načrtovanje poskusa
2. Analiza slike
3. Kontrola kvalitete
4. Normalizacija podatkov
5. Primerjava razredov
6. Napovedovanje razreda
7. Oblikovanje razredov

Oblikovanje razredov

- cilji
  - na podlagi izražanja genov **določiti tipične razrede** v populaciji
  - odkriti **biološke značilnosti** posameznih razredov (npr. različnih vrst tumorjev)
  - odkriti tipične **profile izražanja genov**
- metode **nenadzorovanega** strojnega učenja
  - večrazsežnostno lestvičenje (MDS)
  - gručenje
  - vizualizacijske metode
    - razsevni diagram
    - radviz diagram
- **validacija**
  - razrede je moč oblikovati tudi na **naključnih** podatkih !!!



MDS

Dieta
- **A**ltromin
- **W**estern
- **Z**ero cholesterol

Spol
- Male
- Female

Genotip
- **cP** (Cre +)
- **cM** (Cre –)

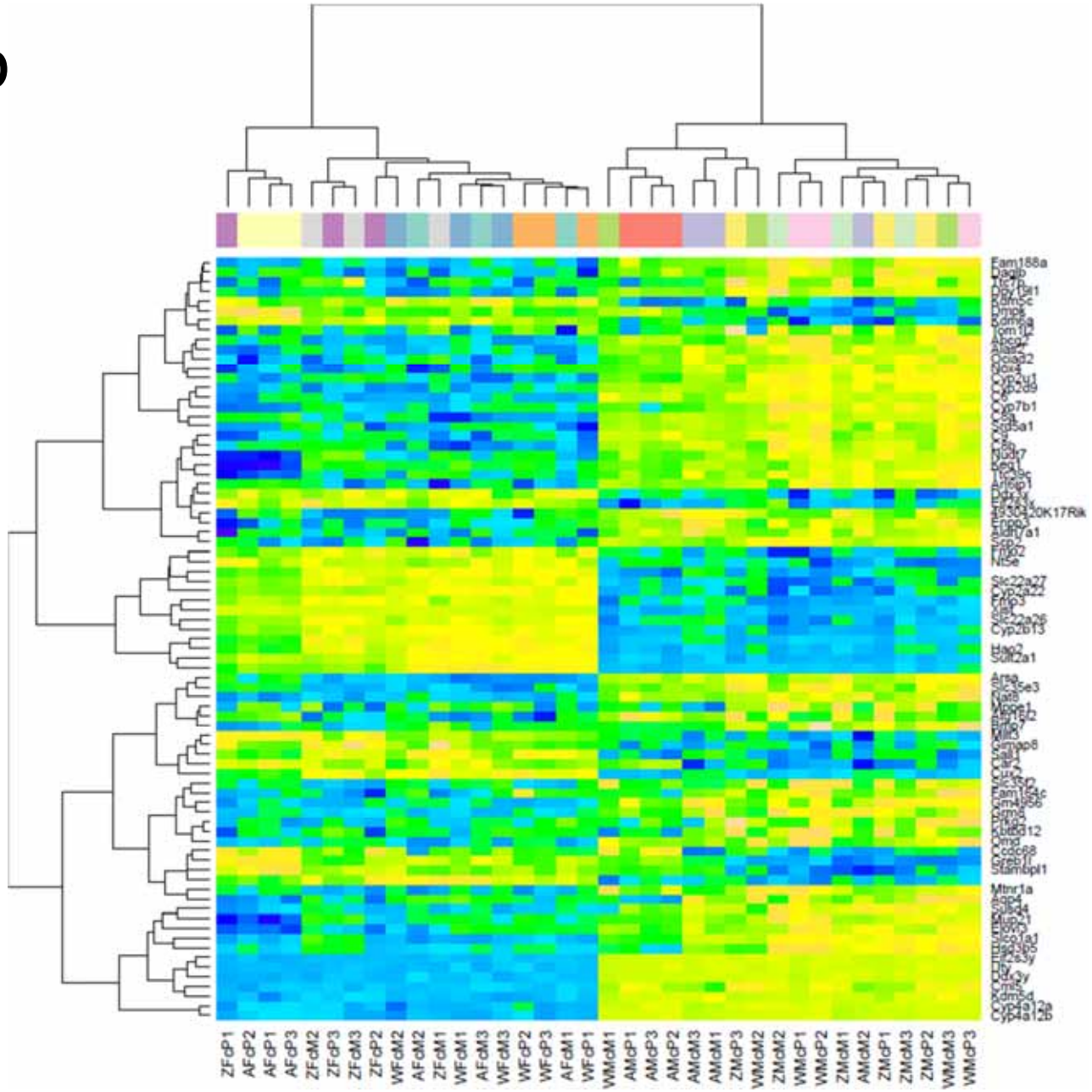
Serija
- **1**
- **2**
- **3**

Lorbek G *et al.* (2012) Unpublished data.

Hierarhično gručenje

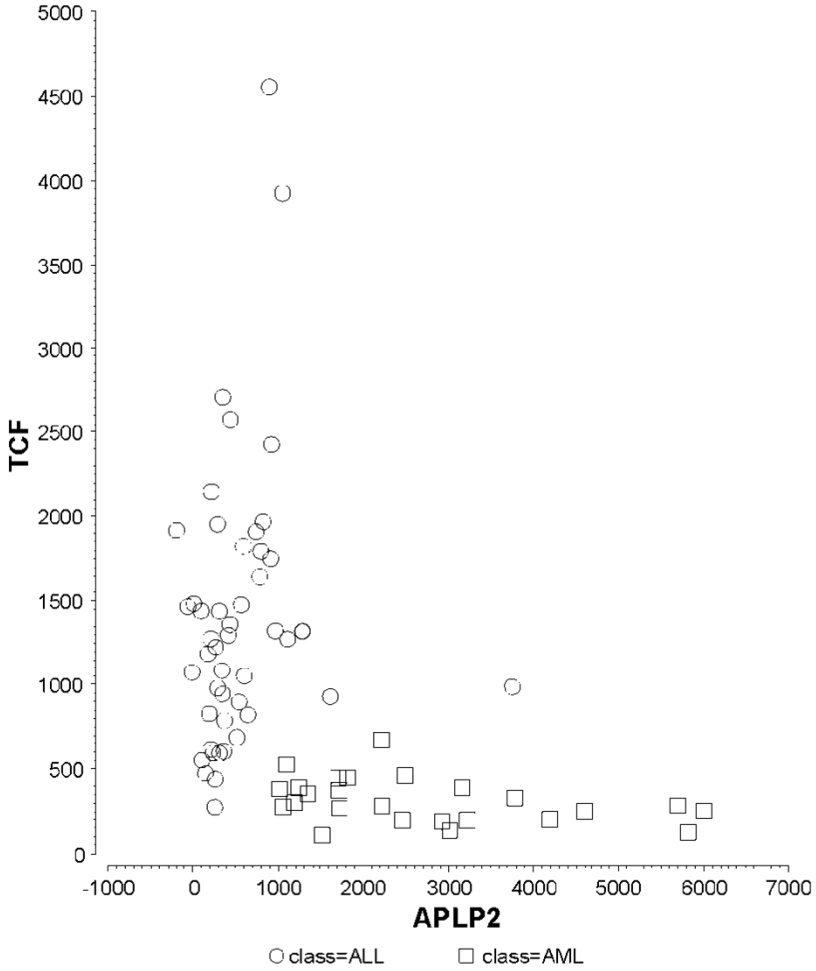
36 vzorcev
- 3 diete
- 2 spola
- 2 genotipa
- 3 biološke ponovitve

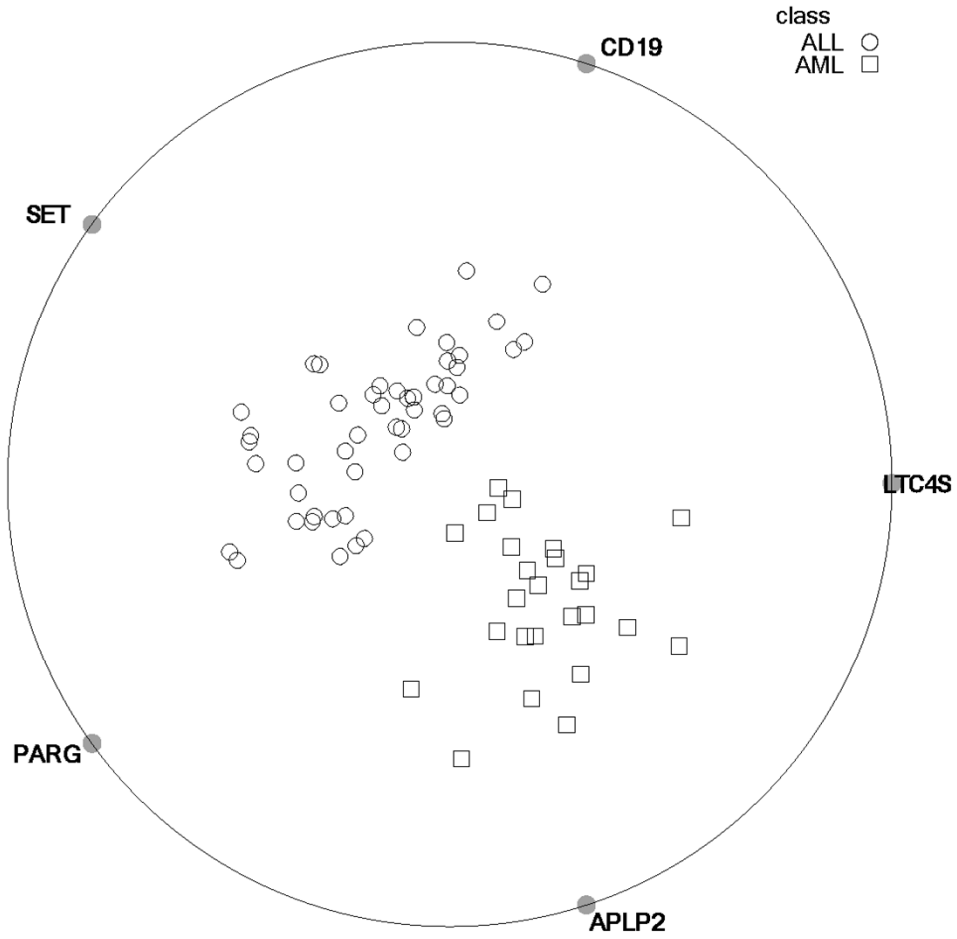
77 genov (α<0.05)



Lorbek G *et al.* (2012) Unpublished data.

Razsevni in radviz diagram





Mramor *et al.* (2006) DNA Microarray Cancer Diagnostics – Simple and Effective Visual Models. *IMS* 11(1).

Orodja

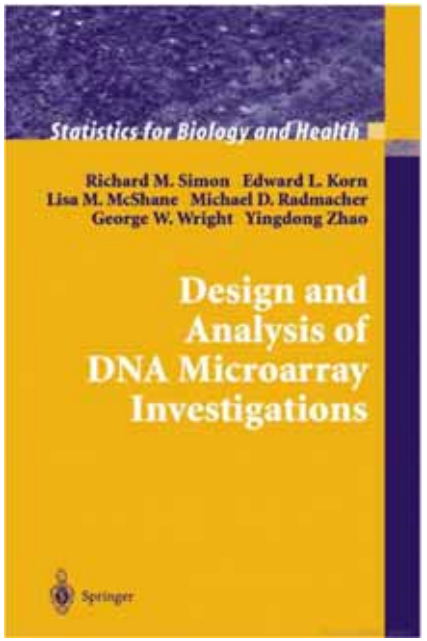
- R Project for Statistical Computing
  http://www.r-project.org/

- Bioconductor
  http://www.bioconductor.org/

- BRB-Array Tools
  http://linus.nci.nih.gov/BRB-ArrayTools.html

Literatura

- Simon RM *et al*. Design and analysis of DNA microarray investigations. Springer, 2003. http://linus.nci.nih.gov/~brb/book.html



- Hovatta I *et al*. DNA microarray data analysis, 2nd edition. CSC, 2005. http://www.csc.fi/english/research/sciences/bioscience/books/microarraybook2nd