Bevezetés az R-be
Betekintés az expressziós chipek kiértekelésébe

Solymosi Norbert

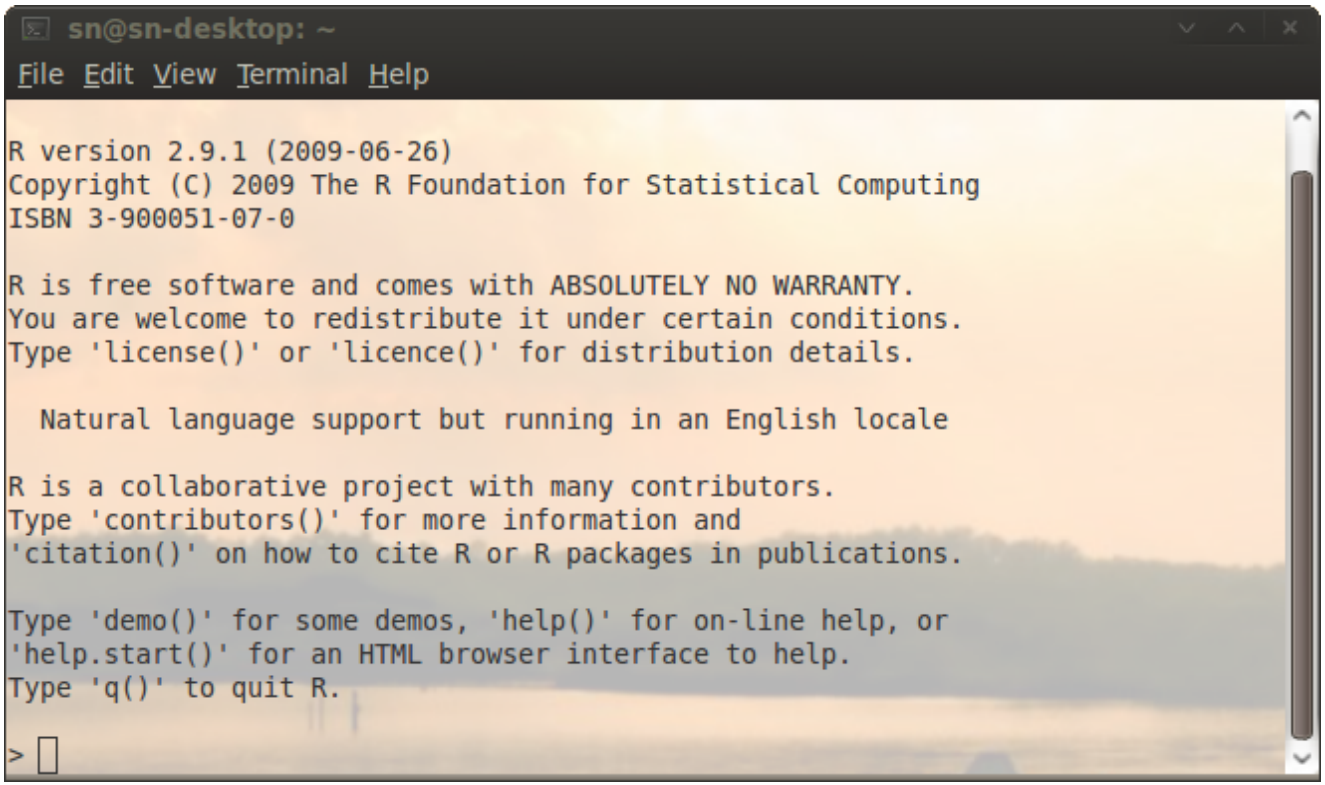
Állatorvostudományi Egyetem

Neuroinformatika
SE Szentágothai Doktori Iskola

2016. szeptember 28.

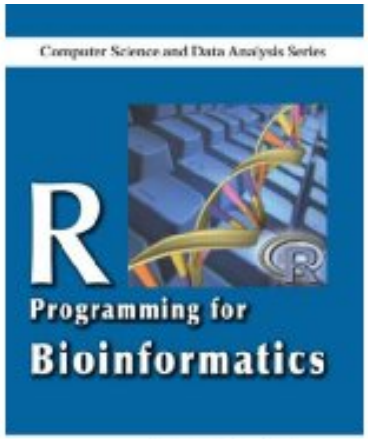
http://www.r-project.org/

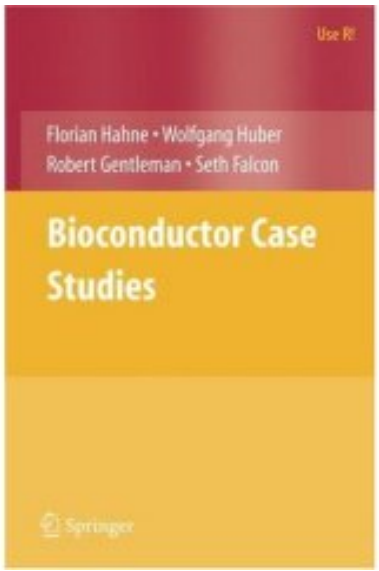
- S, S-Plus
- Robert Gentleman, Ross Ihaka
- Szkript-nyelv
- Függvények (csomagok, könyvtárak)

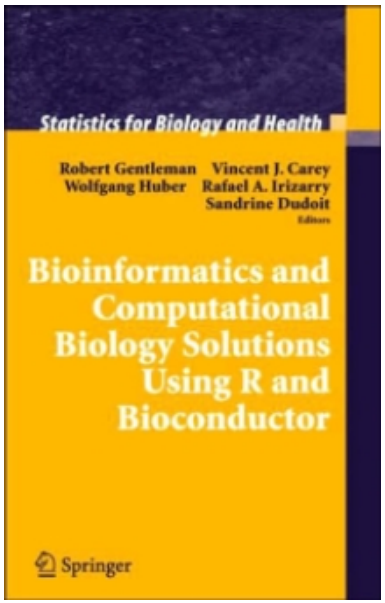


http://www.bioconductor.org/

- Robert Gentleman
- Csomagok
  - Software
  - Metadata (Annotation, CDF and Probe)
  - Custom CDF
  - Experiment Data
  - Complete Taxonomy







Telepítés

R

- http://cran.r-project.org/
- Binárisok – forrásból
- Alaptelepítés csomagjai
- Csomagok telepítése

```
> install.packages('vcd')
```

Bioconductor

- Csomagok telepítése

```
> setRepositories()
> install.packages('affy')
```

- Csomagcsoportok telepítése

```
> source('http://bioconductor.org/biocLite.R')
> biocLite('RBioinf')
```

```
>
> 1 + 2

[1] 3

objektum <- kifejezés.
> a <- 1 + 2
> a

[1] 3

> (a <- 1 + 2)

[1] 3

> (a <- 5)

[1] 5

> fuggveny.neve(arg1,arg2,...)
> length(a)

[1] 1
```

Könyvtárak

- A függvények könyvtárakban érhetők el
- Könyvtárak telepítése

```
> setRepositories()

--- Please select repositories for use in this session ---

1: + CRAN
2:   Omegahat
3:   BioC software
4:   BioC annotation
5:   BioC experiment
6:   BioC extra
7:   R-Forge

Enter one or more numbers separated by spaces
1:

> install.packages('vcd')
```

- Könyvtárak betöltése

```
> library(lattice)
```

Súgó

```
> help(t.test)
> ?t.test

t.test                  package:stats                  R Documentation

Student's t-Test

Description:
     Performs one and two sample t-tests on vectors of data.

Usage:
     t.test(x, ...)

     ## Default S3 method:
     t.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, var.equal = FALSE,
            conf.level = 0.95, ...)

     ## S3 method for class 'formula':
     t.test(formula, data, subset, na.action, ...)

Arguments:
       x: a (non-empty) numeric vector of data values.

       y: an optional (non-empty) numeric vector of data values.

alternative: a character string specifying the alternative hypothesis,
          must be one of '"two.sided"' (default), '"greater"' or
```

Fájlok, adatok

```
> setwd('/home/user/chip')
> getwd()
[1] "/home/user/chip"
```

```
read.table(file, header = FALSE, sep = "", quote = "\"'", dec = ".",
row.names, col.names, as.is = !stringsAsFactors, na.strings = "NA",
colClasses = NA, nrows = -1, skip = 0, check.names = TRUE,
fill = !blank.lines.skip, strip.white = FALSE, blank.lines.skip = TRUE,
comment.char = "#", allowEscapes = FALSE, flush = FALSE,
stringsAsFactors = default.stringsAsFactors(), fileEncoding = "",
encoding = "unknown")
```

függvény	sep	dec	quote	fill
read.line	""	.	\"'	!blank.lines.skip
read.csv	,	.	\"	TRUE
read.csv2	;	,	\"	TRUE
read.delim	\t	.	\"	TRUE
read.delim2	\t	,	\"	TRUE

Fájlok, adatok

```
write()
write.table()

save()
save(list = ls(all=TRUE), file = "minden_objektum.RData")
save.image()

dput()
dget()

dump()
source()

savehistory()
loadhistory()
```

Vektor

```
> (a <- 1:5)

[1] 1 2 3 4 5

> (a <- c(9,4,6,7,1,2,5))

[1] 9 4 6 7 1 2 5

> a[3]

[1] 6

> (a <- vector(mode = "numeric", length = 5))
> (a <- numeric(length = 5))

[1] 0 0 0 0 0

> (a <- vector(mode = "logical", length = 5))
> (a <- logical(length = 5))

[1] FALSE FALSE FALSE FALSE FALSE

> (a <- vector(mode = "character", length = 5))
> (a <- character(length = 5))

[1] "" "" "" "" ""
```

Mátrix

```
> a <- 1:6
> (m <- matrix(a, nr = 3))

     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6

> (m <- matrix(a, nr = 3, byrow = T))

     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6

> dim(a) <- c(3, 2)
> a

     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

Mátrix

```
> (x <- matrix(1:9, nc = 3))

     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> x[2, 2]

[1] 5

> x[2, ]

[1] 2 5 8

> x[, 2]

[1] 4 5 6
```

```
> x[-1, ]

     [,1] [,2] [,3]
[1,]    2    5    8
[2,]    3    6    9

> x[, -1]

     [,1] [,2]
[1,]    4    7
[2,]    5    8
[3,]    6    9

> x[-1, -1]

     [,1] [,2]
[1,]    5    8
[2,]    6    9

> x[-c(1, 3), ]

[1] 2 5 8
```

Data frame

```
> x <- 1:4
> n <- 10
> (r <- data.frame(x, n))

  x  n
1 1 10
2 2 10
3 3 10
4 4 10

> (r <- data.frame(oszlop1 = x, oszlop2 = n))

  oszlop1 oszlop2
1       1      10
2       2      10
3       3      10
4       4      10

> r$oszlop1

[1] 1 2 3 4

> r[,'oszlop1']

[1] 1 2 3 4
```

Lista

```
> x <- matrix(1:9, nc = 3)
> y <- 1:5
> allista <- list(c("a", "b", "c"),
+ c(8, 5, 2, 4, 1, 3))
> lista <- list(x, y, allista)
> names(lista) <- c("r", "t", "z")
> lista
$r
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

$t
[1] 1 2 3 4 5

$z
$z[[1]]
[1] "a" "b" "c"

$z[[2]]
[1] 8 5 2 4 1 3
```

```
> lista[[1]]
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> lista$r
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Affymetrix expressziós chip

Gén

↓

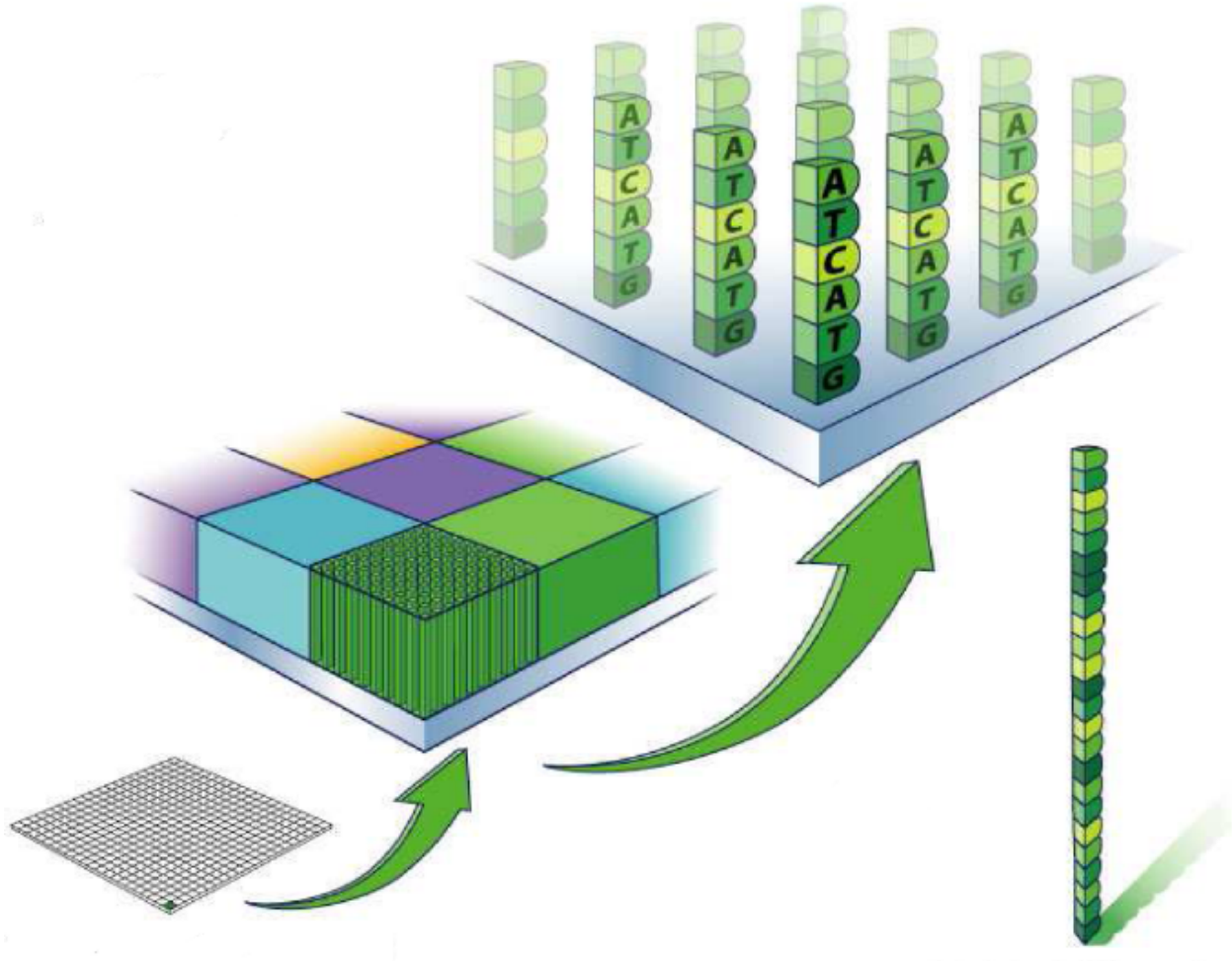
Egy vagy több probeset

↓

Több probe (PM, MM)

↓

25 mer



Affymetrix expressziós chip

Gén

↓

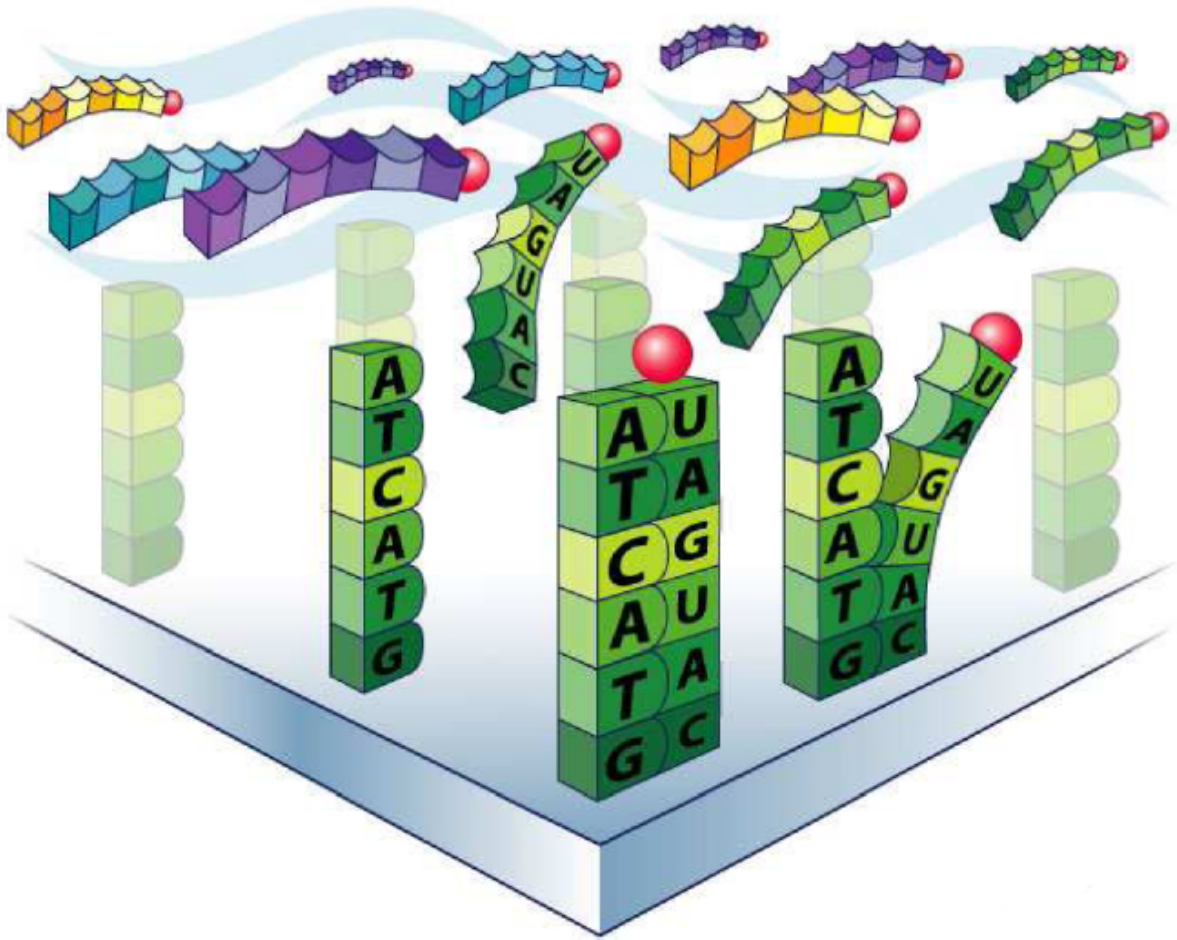
Egy vagy több probeset

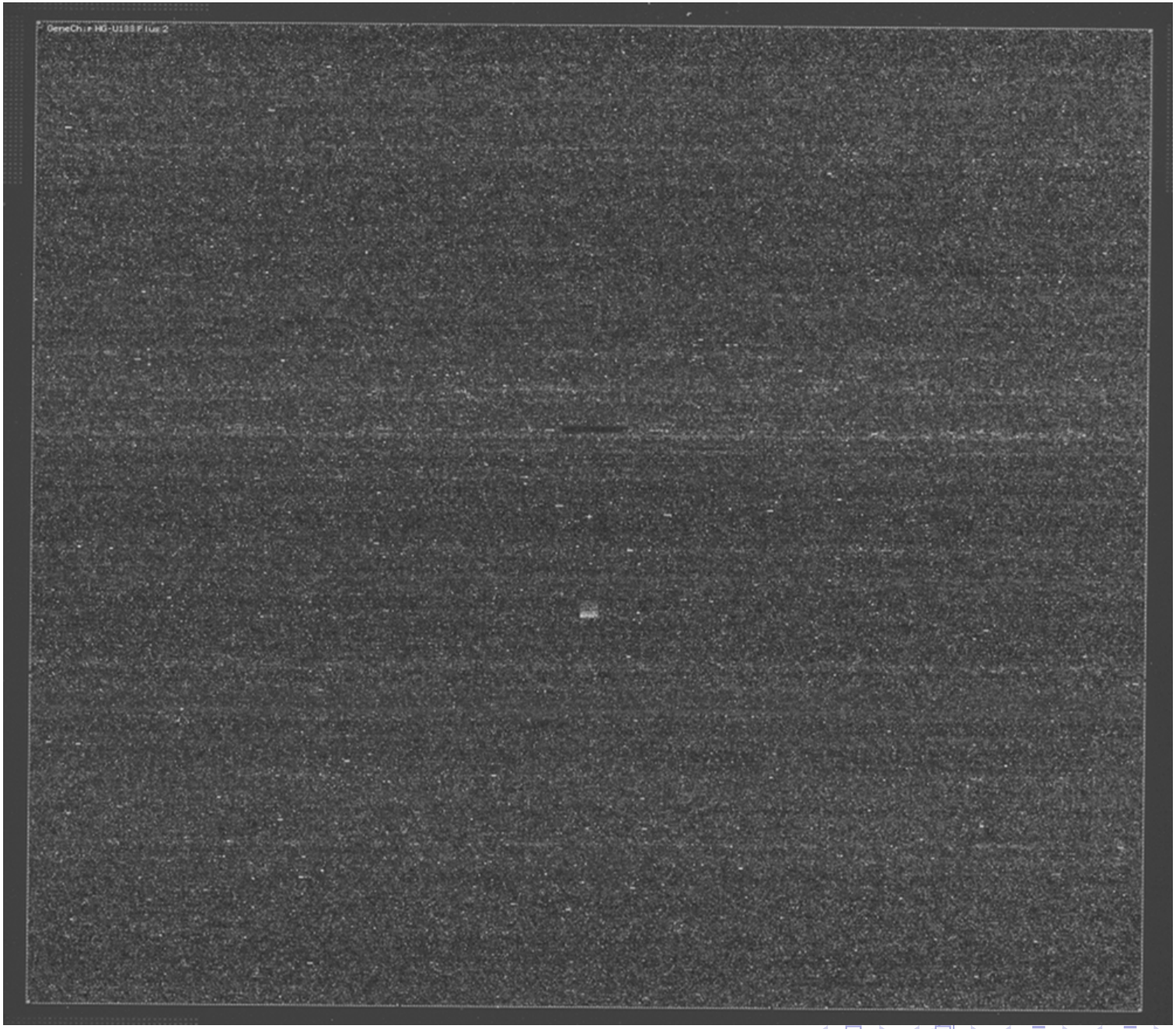
↓

Több probe (PM, MM)

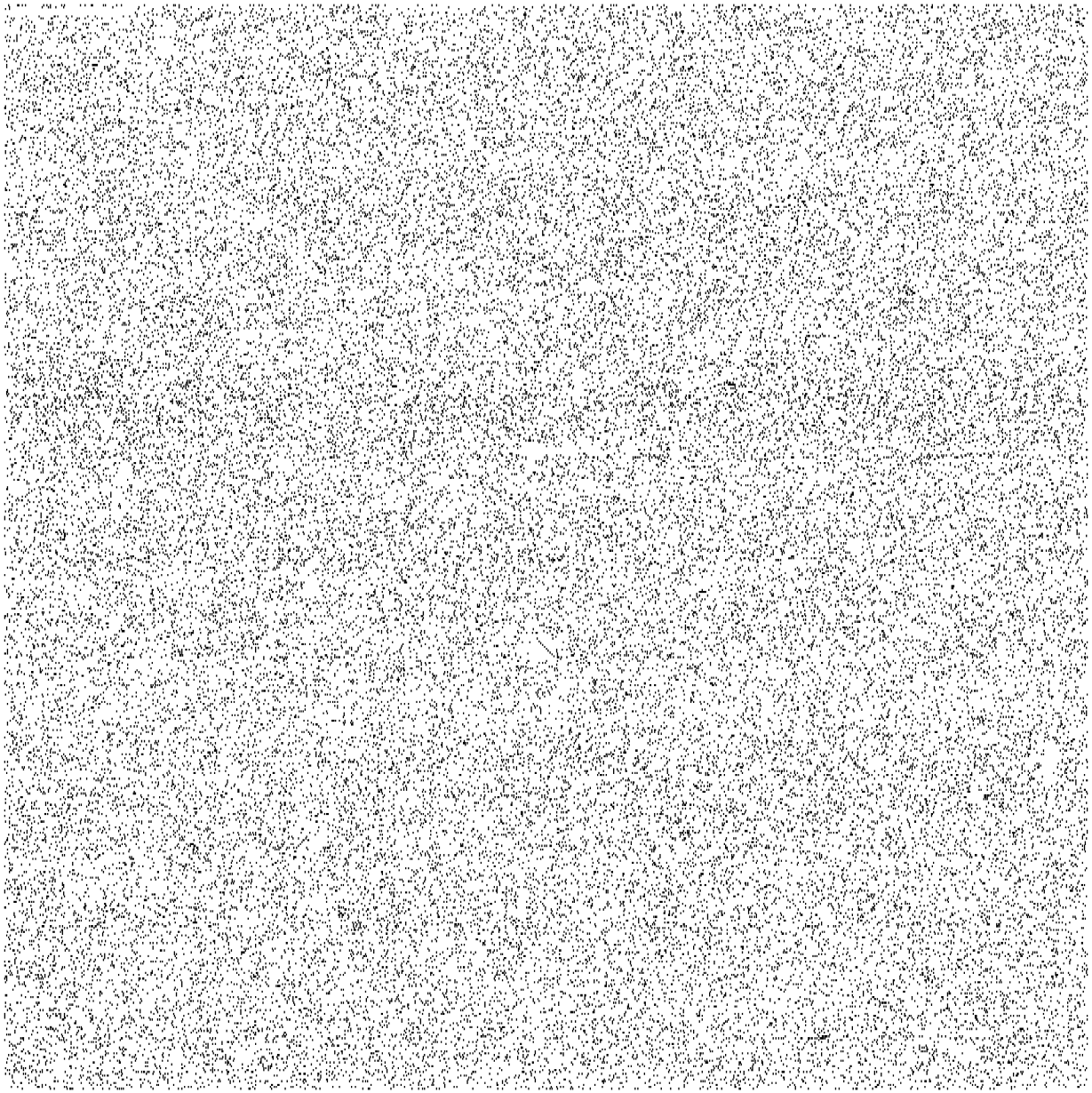
↓

25 mer

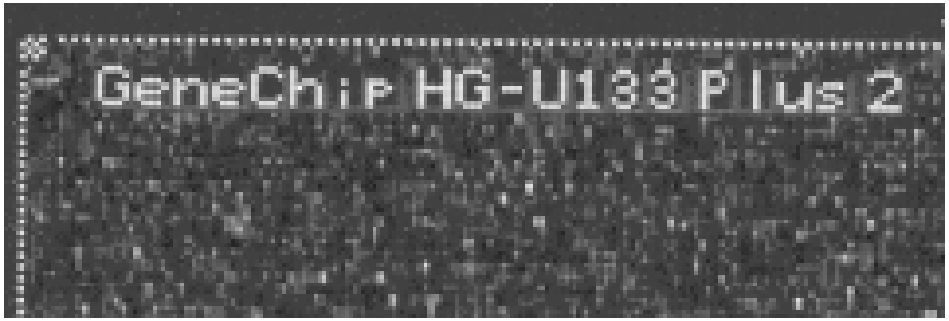




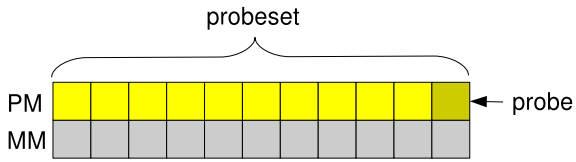
Minden probeset
1. probe-ja



Intenzitás – Probe



- Probe: 8×8 pixel
- Rács illesztése
- Keret elhagyása
- 75. percentilis $\rightarrow$ intenzitás
- CEL-állomány:
  - PM- és MM-intenzitás



AffyBatch

```
> library('affy')
> setwd('munkakönyvtár')
> beolvasott.chipek = ReadAffy()

> library("CLL")

Loading required package: affy
Loading required package: Biobase

Welcome to Bioconductor

  Vignettes contain introductory material. To view, type
  'openVignette()'. To cite Bioconductor, see
  'citation("Biobase")' and for packages 'citation(pkgname)'.

> data("CLLbatch")
```

AffyBatch

```
> CLLbatch

AffyBatch object
size of arrays=640x640 features (91212 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=24
number of genes=12625
annotation=hgu95av2
notes=
```

The AffyBatch object has 24 samples that were affixed to Affymetrix hgu95av2 arrays. These 24 samples came from 24 CLL patients that were either classified as stable or progressive in regards to disease progression.

The CLL package contains the chronic lymphocytic leukemia (CLL) gene expression data. The CLL data had 24 samples that were either classified as progressive or stable in regards to disease progression. The CLL microarray data came from Dr. Sabina Chiaretti at Division of Hematology, Department of Cellular Biotechnologies and Hematology, University La Sapienza, Rome, Italy and Dr. Jerome Ritz at Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

AffyBatch

- Minták

```
> sampleNames(CLLbatch)
```

```
 [1] "CLL10.CEL" "CLL11.CEL" "CLL12.CEL" "CLL13.CEL" "CLL14.CEL" "CLL15.CEL"
 [7] "CLL16.CEL" "CLL17.CEL" "CLL18.CEL" "CLL19.CEL" "CLL1.CEL"  "CLL20.CEL"
[13] "CLL21.CEL" "CLL22.CEL" "CLL23.CEL" "CLL24.CEL" "CLL2.CEL"  "CLL3.CEL"
[19] "CLL4.CEL"  "CLL5.CEL"  "CLL6.CEL"  "CLL7.CEL"  "CLL8.CEL"  "CLL9.CEL"
```

```
> length(sampleNames(CLLbatch))
```

```
[1] 24
```

- Probesetek

```
> featureNames(CLLbatch)[1:10]
```

```
 [1] "1000_at"   "1001_at"   "1002_f_at" "1003_s_at" "1004_at"   "1005_at"
 [7] "1006_at"   "1007_s_at" "1008_f_at" "1009_at"
```

```
> length(featureNames(CLLbatch))
```

```
[1] 12625
```

AffyBatch

PM

```
> pm(CLLbatch, "1000_at")[,1:2]
```

	CLL10.CEL	CLL11.CEL
1000_at1	476.5	668.0
1000_at2	280.0	431.0
1000_at3	107.0	130.0
1000_at4	240.0	629.0
1000_at5	124.0	391.5
1000_at6	577.5	368.5
1000_at7	96.0	143.0
1000_at8	91.0	152.3
1000_at9	155.0	405.3
1000_at10	143.0	232.3
1000_at11	116.5	198.0
1000_at12	159.8	486.0
1000_at13	490.0	1587.0
1000_at14	940.0	3615.0
1000_at15	386.8	1794.0
1000_at16	174.5	615.5

MM

```
> mm(CLLbatch, "1000_at")[,1:2]
```

	CLL10.CEL	CLL11.CEL
1000_at1	853.5	1256.0
1000_at2	185.0	233.0
1000_at3	89.0	92.0
1000_at4	95.0	137.0
1000_at5	80.0	87.3
1000_at6	633.0	550.8
1000_at7	79.0	91.0
1000_at8	79.0	65.0
1000_at9	106.3	158.0
1000_at10	106.0	118.0
1000_at11	104.0	126.0
1000_at12	88.3	101.5
1000_at13	362.0	812.5
1000_at14	601.0	1916.8
1000_at15	279.0	745.0
1000_at16	151.8	395.0

AffyBatch – leíró adatok

AffyBatch-ben eddigi info

```
> head(pData(CLLbatch))
```

```
          sample
CLL10.CEL      1
CLL11.CEL      2
CLL12.CEL      3
CLL13.CEL      4
CLL14.CEL      5
CLL15.CEL      6
```

A minták leírása

```
> data(disease)
> head(disease)
```

```
  SampleID  Disease
1    CLL10     <NA>
2    CLL11 progres.
3    CLL12   stable
4    CLL13 progres.
5    CLL14 progres.
6    CLL15 progres.
```

Egyesítve az AffyBatch-ben

```
> rownames(disease) = disease$SampleID
> uj.nev = sub('\\.CEL$', '',
+ sampleNames(CLLbatch))
> uj.nev[1:5]

[1] "CLL10" "CLL11" "CLL12" "CLL13" "CLL14"

> sampleNames(CLLbatch) = uj.nev
> e.id = match(rownames(disease), sampleNames(CLLbatch))
> vmd = data.frame(labelDescription = c('Sample ID',
+ 'Disease status: progressive or stable disease'))
> phenoData(CLLbatch) = new('AnnotatedDataFrame',
+ data = disease[e.id, ], varMetadata = vmd)

> head(pData(CLLbatch))
```

```
      SampleID  Disease
CLL10    CLL10     <NA>
CLL11    CLL11 progres.
CLL12    CLL12   stable
CLL13    CLL13 progres.
CLL14    CLL14 progres.
CLL15    CLL15 progres.
```

Mértékek

- Átlagos háttér
- Átskálázási factor
- Jelenlét %
- 3'/5' arány
- Hibridizációs kontrollok

- 4×4 grid
- mindegyik cellán belül az alsó 2% lesz a háttér
- ezek átlaga az átlagos háttér
- $\frac{\text{legnagyobb}}{\text{legkisebb}} < 3$

Mértékek

- Átlagos háttér
- Átskálázási factor
- Jelenlét %
- 3'/5' arány
- Hibridizációs kontrollok

- MAS 5.0 normalizáció
- alapgondolat, hogy a transzkriptumok kis hányada különbözik csak
- minták trimmelt átlagos intenzitása megegyezik
- alsó, felső 2% elhagyása
- az $\frac{\text{legnagyobb}}{\text{átlag}} < 3$

Mértékek

- Átlagos háttér
- Átskálázási factor
- Jelenlét %
- 3'/5' arány
- Hibridizációs kontrollok

- Probesetek hány % expresszálódik?
- $PM > MM$: hiányzik, határeset, jelen van
- $\frac{\text{legnagyobb}}{\text{legkisebb}} < 3$

Mértékek

- Átlagos háttér
- Átskálázási factor
- Jelenlét %
- 3'/5' arány
- Hibridizációs kontrollok

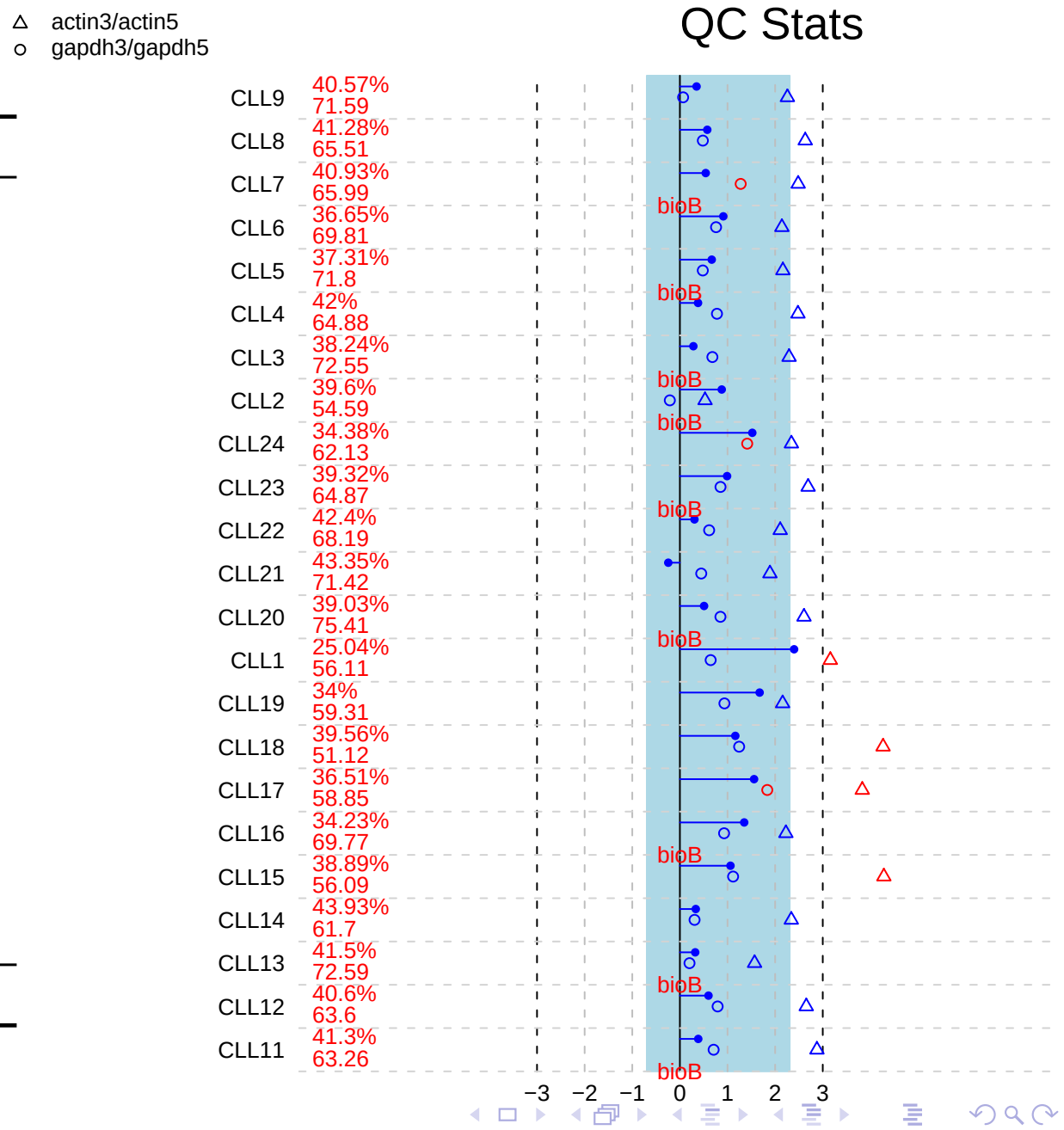
- RNS-minősége, bomlottsága
- β-aktin, GAPDH
- a legtöbb sejt azonos szinten expresszálja
- hosszúak, probesetek az 5'- és a 3'-végről, ill. közepéről
- a 3'- és az 5'-jel aránya
- ha magas $\leftarrow$ pl. bomlott
- β-aktin: < 3
- GAPDH: ≈ 1

Mértékek

- Átlagos háttér
- Átskálázási factor
- Jelenlét %
- 3'/5' arány
- Hibridizációs kontrollok

- BioB, BioC, BioD, CreX
- *Bacillus subtillis*
- intenzitás $\sim$ hibridizáció, szkennelés
- BioB: mindegyik chipen jelen kell lennie, de legalább 70%-ukban

Minta	H. átlag	S. faktor	Jelenlét %
CLL1	56.11	5.27	25.04
CLL2	54.59	1.84	39.60
CLL3	72.55	1.22	38.24
CLL4	64.88	1.30	42.00
CLL5	71.80	1.59	37.31
CLL6	69.81	1.88	36.65
CLL7	65.99	1.46	40.93
CLL8	65.51	1.49	41.28
CLL9	71.59	1.27	40.57
CLL11	63.26	1.31	41.30
CLL12	63.60	1.52	40.60
CLL13	72.59	1.25	41.50
CLL14	61.70	1.26	43.93
CLL15	56.09	2.09	38.89
CLL16	69.77	2.55	34.23
CLL17	58.85	2.94	36.51
CLL18	51.12	2.24	39.56
CLL19	59.31	3.19	34.00
CLL20	75.41	1.42	39.03
CLL21	71.42	0.84	43.35
CLL22	68.19	1.24	42.40
CLL23	64.87	1.99	39.32
CLL24	62.13	2.87	34.38
Arány	1.47	2.75	1.75



Probe-adatok

```
> library("annotate")
Loading required package: AnnotationDbi
> annotation(CLLbatch)

[1] "hgu95av2"

> library(hgu95av2probe)
> data(hgu95av2probe)
> hgu95av2probe

Object of class probetable data.frame with 201800 rows and 6 columns.

> as.data.frame(hgu95av2probe[1:10,-5])

                    sequence   x   y Probe.Set.Name Target.Strandedness
1   TGGCTCCTGCTGAGGTCCCCTTTCC 395 301        1138_at           Antisense
2   GGCTGTGAATTCCTGTACATATTTC 322 441        1138_at           Antisense
3   GCTTCAATTCCATTATGTTTTAATG 213 419        1138_at           Antisense
4   GCCGTTTGACAGAGCATGCTCTGCG 279 435        1138_at           Antisense
5   TGACAGAGCATGCTCTGCGTTGTTG 473 299        1138_at           Antisense
6   CTCTGCGTTGTTGGTTTCACCAGCT 587 205        1138_at           Antisense
7   GGTTTCACCAGCTTCTGCCCTCACA 423 491        1138_at           Antisense
8   TTCTGCCCTCACATGCACAGGGATT 196 519        1138_at           Antisense
9   CCTCACATGCACAGGGATTTAACAA 240 469        1138_at           Antisense
10 TCCTTGGTACTCTGCCCTCCTGTCA 425 593        1138_at           Antisense
```

Probeset-kapcsolódások (hgu95av2.db)

```
> hgu95av2()

Quality control information for hgu95av2:
This package has the following mappings:
hgu95av2ACCNUM has 12625 mapped keys (of 12625 keys)
hgu95av2ALIAS2PROBE has 37934 mapped keys (of 37934 keys)
hgu95av2CHR has 11957 mapped keys (of 12625 keys)
hgu95av2CHRLENGTHS has 25 mapped keys (of 25 keys)
hgu95av2CHRLOC has 11789 mapped keys (of 12625 keys)
hgu95av2CHRLOCEND has 11789 mapped keys (of 12625 keys)
hgu95av2ENSEMBL has 11639 mapped keys (of 12625 keys)
hgu95av2ENSEMBL2PROBE has 9021 mapped keys (of 9021 keys)
hgu95av2ENTREZID has 11960 mapped keys (of 12625 keys)
hgu95av2ENZYME has 1978 mapped keys (of 12625 keys)
hgu95av2ENZYME2PROBE has 725 mapped keys (of 725 keys)
hgu95av2GENENAME has 11960 mapped keys (of 12625 keys)
hgu95av2GO has 11363 mapped keys (of 12625 keys)
hgu95av2GO2ALLPROBES has 9581 mapped keys (of 9581 keys)
hgu95av2GO2PROBE has 6774 mapped keys (of 6774 keys)
hgu95av2MAP has 11919 mapped keys (of 12625 keys)
hgu95av2OMIM has 10350 mapped keys (of 12625 keys)
hgu95av2PATH has 4585 mapped keys (of 12625 keys)
hgu95av2PATH2PROBE has 203 mapped keys (of 203 keys)
hgu95av2PFAM has 11878 mapped keys (of 12625 keys)
hgu95av2PMID has 11898 mapped keys (of 12625 keys)
hgu95av2PMID2PROBE has 206993 mapped keys (of 206993 keys)
hgu95av2PROSITE has 11878 mapped keys (of 12625 keys)
hgu95av2REFSEQ has 11883 mapped keys (of 12625 keys)
hgu95av2SYMBOL has 11960 mapped keys (of 12625 keys)
hgu95av2UNIGENE has 11905 mapped keys (of 12625 keys)
hgu95av2UNIPROT has 11764 mapped keys (of 12625 keys)
```

```
> (chrom = buildChromLocation('hgu95av2.db'))

Instance of a chromLocation class with the following fields:
  Organism:  Homo sapiens
  Data source:  hgu95av2.db
  Number of chromosomes for this organism:  25
  Chromosomes of this organism and their lengths in base pairs:
    1 : 247249719
    10 : 135374737
    11 : 134452384
    12 : 132349534
    13 : 114142980
    14 : 106368585
    15 : 100338915
    16 : 88827254
    17 : 78774742
    18 : 76117153
    19 : 63811651
    2 : 242951149
    20 : 62435964
    21 : 46944323
    22 : 49691432
    3 : 199501827
    4 : 191273063
    5 : 180857866
    6 : 170899992
    7 : 158821424
    8 : 146274826
    9 : 140273252
    M : 16571
    X : 154913754
    Y : 57772954
```

```
> chromLocs(chrom)[['Y']][1:10]
  266_s_at   31911_at 32930_f_at 32991_f_at   35885_at 35929_s_at   35930_at
-19611913   14324840   15145847   -6793958   13322553    9914563    9914563
  36321_at   37583_at   38182_at
  13283691  -20326688   20217829

> get('35885_at', probesToChrom(chrom))

[1] "Y"

> probesets = featureNames(CLLbatch)
> getSYMBOL(probesets[1:3], 'hgu95av2.db')

  1000_at   1001_at 1002_f_at
  "MAPK3"    "TIE1" "CYP2C19"

> mget(probesets[1:3], hgu95av2SYMBOL)

$`1000_at`
[1] "MAPK3"

$`1001_at`
[1] "TIE1"

$`1002_f_at`
[1] "CYP2C19"

> get(probesets[1:3], hgu95av2SYMBOL)

[1] "MAPK3"
```

```
> hgu95av2SYMBOL$'1000_at'

[1] "MAPK3"

> hgu95av2SYMBOL[['1000_at']]

[1] "MAPK3"

> hgu95av2GENENAME$'1000_at'

[1] "mitogen-activated protein kinase 3"

> hgu95av2ENSEMBL$'1000_at'

[1] "ENSG00000102882"

> hgu95av2ACCNUM$'1000_at'

[1] "X60188"

> sym.2.hgu95av2 = revmap(hgu95av2SYMBOL)
> sym.2.hgu95av2$'MAPK3'

[1] "1000_at"
```

GO

```
> toTable(hgu95av2GO['1000_at'])
```

	probe_id	go_id	Evidence	Ontology
1	1000_at	GO:0006468	IDA	BP
2	1000_at	GO:0007049	IEA	BP
3	1000_at	GO:0007265	EXP	BP
4	1000_at	GO:0044419	IEA	BP
5	1000_at	GO:0005829	EXP	CC
6	1000_at	GO:0005634	IDA	CC
7	1000_at	GO:0005654	EXP	CC
8	1000_at	GO:0005730	IDA	CC
9	1000_at	GO:0005856	IDA	CC
10	1000_at	GO:0000166	IEA	MF
11	1000_at	GO:0005515	IPI	MF
12	1000_at	GO:0004674	EXP	MF
13	1000_at	GO:0004674	IEA	MF
14	1000_at	GO:0004707	EXP	MF
15	1000_at	GO:0004707	NAS	MF
16	1000_at	GO:0016740	IEA	MF

KEGG

```
> (ps.paths = as.list(hgu95av2PATH)[['1000_at']])
 [1] "04010" "04012" "04150" "04350" "04360" "04370" "04510" "04520" "04540"
[10] "04620" "04650" "04664" "04720" "04730" "04810" "04910" "04912" "04916"
[19] "04930" "05010" "05210" "05211" "05212" "05213" "05214" "05215" "05216"
[28] "05218" "05219" "05220" "05221" "05223"
> library(KEGG.db)
> pathways.by.ids = as.list(KEGGPATHID2NAME)
> pathways.by.names = as.list(KEGGPATHNAME2ID)
> length(pathways.by.names)
[1] 336
for (ps in ps.paths) print(pathways.by.ids[[ps]])
[1] "MAPK signaling pathway"
[1] "ErbB signaling pathway"
[1] "mTOR signaling pathway"
[1] "TGF-beta signaling pathway"
[1] "Axon guidance"
[1] "VEGF signaling pathway"
[1] "Focal adhesion"
[1] "Adherens junction"
[1] "Gap junction"
[1] "Toll-like receptor signaling pathway"
[1] "Natural killer cell mediated cytotoxicity"
[1] "Fc epsilon RI signaling pathway"
[1] "Long-term potentiation"
[1] "Long-term depression"
[1] "Regulation of actin cytoskeleton"
[1] "Insulin signaling pathway"
[1] "GnRH signaling pathway"
[1] "Melanogenesis"
[1] "Type II diabetes mellitus"
[1] "Alzheimer's disease"
[1] "Colorectal cancer"
[1] "Renal cell carcinoma"
[1] "Pancreatic cancer"
[1] "Endometrial cancer"
[1] "Glioma"
[1] "Prostate cancer"
[1] "Thyroid cancer"
[1] "Melanoma"
[1] "Bladder cancer"
[1] "Chronic myeloid leukemia"
[1] "Acute myeloid leukemia"
[1] "Non-small cell lung cancer"
```

KEGG

```
> pathways.by.names[['Colorectal cancer']]

[1] "05210"

> pathways.by.ids[['05210']]

[1] "Colorectal cancer"

> (probesets.in.path = as.list(hgu95av2PATH2PROBE)[['05210']][1:10])

 [1] "34055_at"   "34056_g_at" "34415_at"   "36451_at"   "39199_at"
 [6] "1564_at"    "2022_at"    "2023_g_at"  "40972_at"   "1912_s_at"

> as.character(getSYMBOL(probesets.in.path, 'hgu95av2.db'))

 [1] "ACVR1B" "ACVR1B" "ACVR1B" "ACVR1B" "ACVR1B" "AKT1"   "AKT2"   "AKT2"
 [9] "AKT2"   "APC"

> unique(as.character(getSYMBOL(probesets.in.path, 'hgu95av2.db')))

[1] "ACVR1B" "AKT1"   "AKT2"   "APC"
```

PubMed

```
> absts = pm.getabst('37809_at', "hgu95av2")
> absts['37809_at'][[58]]
```

```
An object of class 'pubMedAbst':
Title: HOX expression patterns identify a common signature for
     favorable AML.
PMID: 18668134
Authors: M Andreeff, V Ruvolo, S Gadgil, C Zeng, K Coombes, W Chen, S
     Kornblau, AE Barón, HA Drabkin
Journal: Leukemia
Date: Nov 2008
```

```
> abstText(absts['37809_at'][[58]])
```

```
[1] "Deregulated HOX expression, by chromosomal translocations and myeloid-lymphoid leukemia (MLL)
rearrangements, is causal in some types of leukemia. Using real-time reverse transcription-PCR,
we examined the expression of 43 clustered HOX, polycomb, MLL and FLT3 genes in 119 newly
diagnosed adult acute myeloid leukemias (AMLs) selected from all major cytogenetic groups.
Downregulated HOX expression was a consistent feature of favorable AMLs and, among these cases,
inv(16) cases had a distinct expression profile. Using a 17-gene predictor in 44 additional
samples, we observed a 94.7% specificity for classifying favorable vs intermediate/unfavorable
cytogenetic groups. Among other AMLs, HOX overexpression was associated with nucleophosmin (NPM)
mutations and we also identified a phenotypically similar subset with wt-NPM. In many unfavorable
and other intermediate cytogenetic AMLs, HOX levels resembled those in normal CD34+ cells, except
that the homogeneity characteristic of normal samples was not present. We also observed that HOXA9
levels were significantly inversely correlated with survival and that BMI-1 was overexpressed in
cases with 11q23 rearrangements, suggesting that p19(ARF) suppression may be involved in
MLL-associated leukemia. These results underscore the close relationship between HOX expression
patterns and certain forms of AML and emphasize the need to determine whether these differences
play a role in the disease process."
```

Mérési adatok előkészítése elemzésre

- **Probe**-intenzitás $\rightarrow$ összehasonlítható **probeset**-expressziós érték
- Lépései:
  1. Háttér-korrekció
     Chipen belüli normálás
  2. Normalizáció
     Chipek összehasonlíthatósága
  3. Expressziós érték számítása
     Intenzitásból expressziós érték
- Számos módszer mindegyik lépésre
- N.B. a függvény neve $\neq$ eljárás
  - Pl. `rma()`

    háttér-korrekció RMA-korrekció
    normalizáció kvantilis
    expresszió-számítás medián

ExpressionSet

```
> eset = rma(CLLbatch)

Background correcting
Normalizing
Calculating Expression

> eset

ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 23 samples
  element names: exprs
phenoData
  sampleNames: CLL11, CLL12, ..., CLL9  (23 total)
  varLabels and varMetadata description:
    SampleID: Sample ID
    Disease: Disease status: progressive or stable disease
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation: hgu95av2
```

ExpressionSet

```
> eset$Disease

 [1] progres. stable   progres. progres. progres. progres. stable   stable
 [9] progres. stable   stable   progres. stable   progres. stable   stable
[17] progres. progres. progres. progres. progres. progres. stable
Levels: progres. stable

> eset[,eset$Disease=='stable']

ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 9 samples
  element names: exprs
phenoData
  sampleNames: CLL12, CLL17, ..., CLL9  (9 total)
  varLabels and varMetadata description:
    SampleID: Sample ID
    Disease: Disease status: progressive or stable disease
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation: hgu95av2
```

Expressziós mátrix

```
> e.tab = exprs(eset)
> dim(e.tab)
[1] 12625    23
> colnames(e.tab)[1:5]
[1] "CLL11" "CLL12" "CLL13" "CLL14" "CLL15"
> rownames(e.tab)[1:5]
[1] "1000_at"   "1001_at"   "1002_f_at" "1003_s_at" "1004_at"
> e.tab[1:5,1:5]
            CLL11    CLL12    CLL13    CLL14    CLL15
1000_at   8.314906 8.508307 8.170304 8.095857 7.915544
1001_at   4.563419 4.419578 4.664504 4.497282 4.777937
1002_f_at 4.004539 4.064893 4.157137 4.074372 3.843771
1003_s_at 6.197371 6.449624 6.326980 6.132313 6.090583
1004_at   8.086685 8.267743 8.137425 8.018026 7.591456
```

Expressziós különbségek

- Expressziós értékek összehasonlítása, fenotípus szerint
- Probeseteket külön-külön hasonlítja össze
- Gyakoribb módszerek:
  - Fold-change
    - átlagok, mediánok hányadosa
    - általában log2-skálán
    - nem kezeli a csoportokon belüli variabilitást
    - kisebb mintaszámok
  - Paraméteres próbák
    - variabilitást bevonja
    - a feltételek ritkán teljesülnek
    - nagyobb az ereje
  - Nemparaméteres próbák
    - variabilitást bevonja
    - enyhébb feltételek
    - kisebb az ereje
  - Egyebek: ROC, permutációs próbák, stb.

ALL

```
> library(ALL)
> data(ALL)
> ALL

ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 128 samples
  element names: exprs
phenoData
  sampleNames: 01005, 01010, ..., LAL4  (128 total)
  varLabels and varMetadata description:
    cod:  Patient ID
    diagnosis:  Date of diagnosis
    ...: ...
    date last seen:  date patient was last seen
    (21 total)
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at  (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

ALL

```
> head(pData(ALL))
```

	cod	diagnosis	sex	age	BT	remission	CR	date.cr	t(4;11)	t(9;22)
01005	1005	5/21/1997	M	53	B2		CR	8/6/1997	FALSE	TRUE
01010	1010	3/29/2000	M	19	B2		CR	6/27/2000	FALSE	FALSE
03002	3002	6/24/1998	F	52	B4		CR	8/17/1998	NA	NA
04006	4006	7/17/1997	M	38	B1		CR	9/8/1997	TRUE	FALSE
04007	4007	7/22/1997	M	57	B2		CR	9/17/1997	FALSE	FALSE
04008	4008	7/30/1997	M	17	B1		CR	9/27/1997	FALSE	FALSE

	cyto.normal	citog	mol.biol	fusion protein	mdr	kinet	ccr
01005	FALSE	t(9;22)	BCR/ABL	p210	NEG	dyploid	FALSE
01010	FALSE	simple alt.	NEG	<NA>	POS	dyploid	FALSE
03002	NA	<NA>	BCR/ABL	p190	NEG	dyploid	FALSE
04006	FALSE	t(4;11)	ALL1/AF4	<NA>	NEG	dyploid	FALSE
04007	FALSE	del(6q)	NEG	<NA>	NEG	dyploid	FALSE
04008	FALSE	complex alt.	NEG	<NA>	NEG	hyperd.	FALSE

	relapse	transplant	f.u	date last seen
01005	FALSE	TRUE	BMT / DEATH IN CR	<NA>
01010	TRUE	FALSE	REL	8/28/2000
03002	TRUE	FALSE	REL	10/15/1999
04006	TRUE	FALSE	REL	1/23/1998
04007	TRUE	FALSE	REL	11/4/1997
04008	TRUE	FALSE	REL	12/15/1997

ALL

B-sejtes tumor minták indexeinek kigyűjtése

```
> b.sejt.idxs = grep("^B", as.character(ALL$BT))
```

Molekuláris biológiai tulajdonság alapján 2 csoport indexeinek lekérdezése (BCR/ABL-translokáció, negatív a vizsgált eltérésekre)

```
> mol.tipus.idxs = which(as.character(ALL$mol.biol) %in% c('NEG', 'BCR/ABL'))
```

Közös metszetük alapján új eSet létrehozása

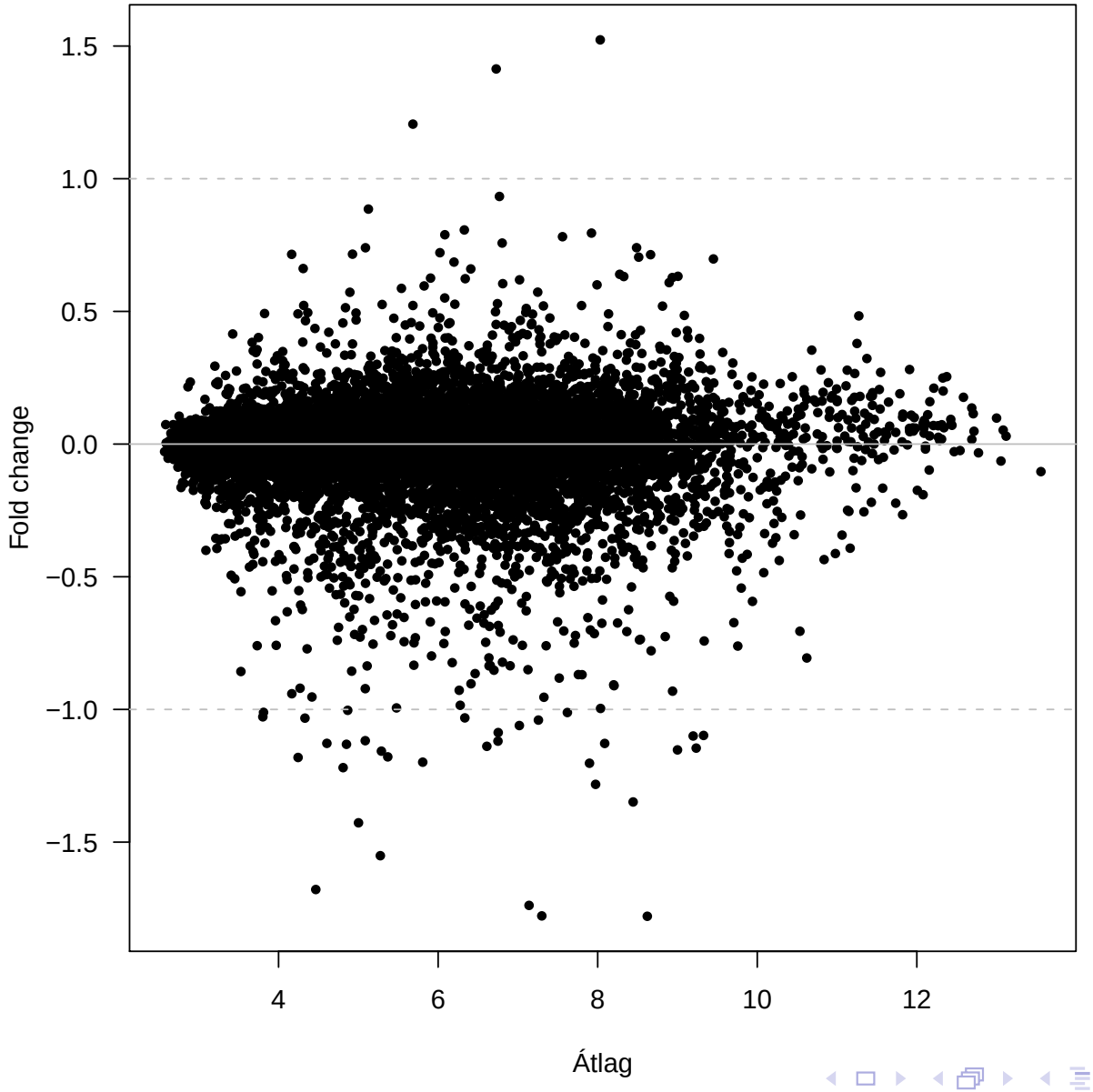
```
> ALL.bcr.neg = ALL[,intersect(b.sejt.idxs, mol.tipus.idxs)]
> ALL.bcr.neg$mol.biol = factor(ALL.bcr.neg$mol.biol)
> ALL.bcr.neg
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 79 samples
  element names: exprs
phenoData
  sampleNames: 01005, 01010, ..., 84004  (79 total)
  varLabels and varMetadata description:
    cod:  Patient ID
    diagnosis:  Date of diagnosis
    ...: ...
    date last seen:  date patient was last seen
    (21 total)
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at  (12625 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

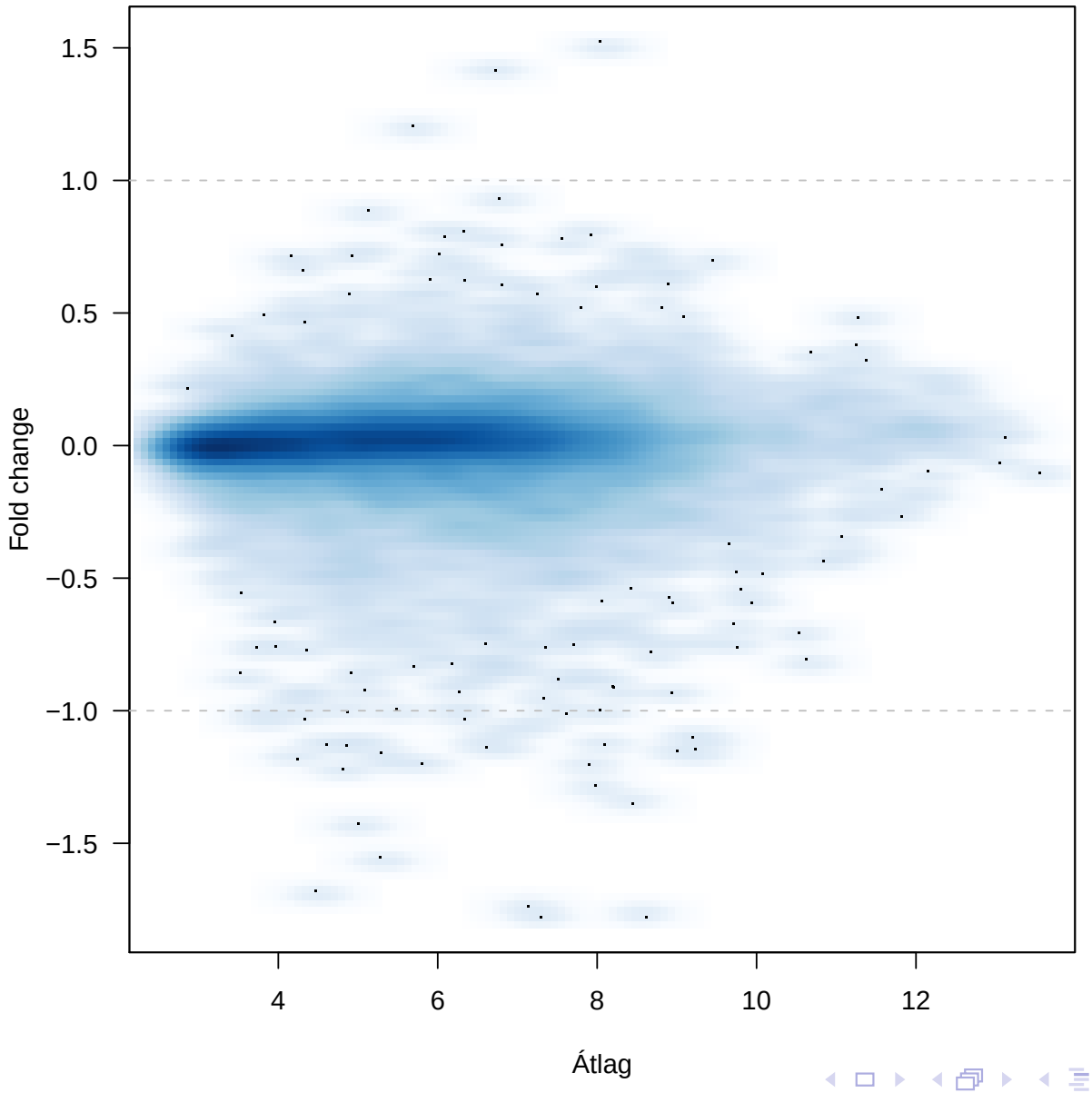
Fold change

```
> ALL.bcr.neg.exprs.m = exprs(ALL.bcr.neg)
> csoportok = as.numeric(ALL.bcr.neg$mol.biol)
> table(csoportok)

csoportok
 1  2
37 42

> idx1 = which(csoportok==1)
> idx2 = which(csoportok==2)

> Fc = rowMeans(ALL.bcr.neg.exprs.m[,idx2])-rowMeans(ALL.bcr.neg.exprs.m[,idx1])
> atlag = rowMeans(ALL.bcr.neg.exprs.m)

> plot(atlag, Fc, xlab = 'átlag', ylab = 'Fold change', las=1, pch=20)
> abline(h=0, col='grey')
> abline(h=1, col='grey', lty=2)
> abline(h=-1, col='grey', lty=2)

> smoothScatter(atlag, Fc, xlab = 'Átlag', ylab = 'Fold change', las=1)
> abline(h=1, col='grey', lty=2)
> abline(h=-1, col='grey', lty=2)
```

Fold change



Fold change



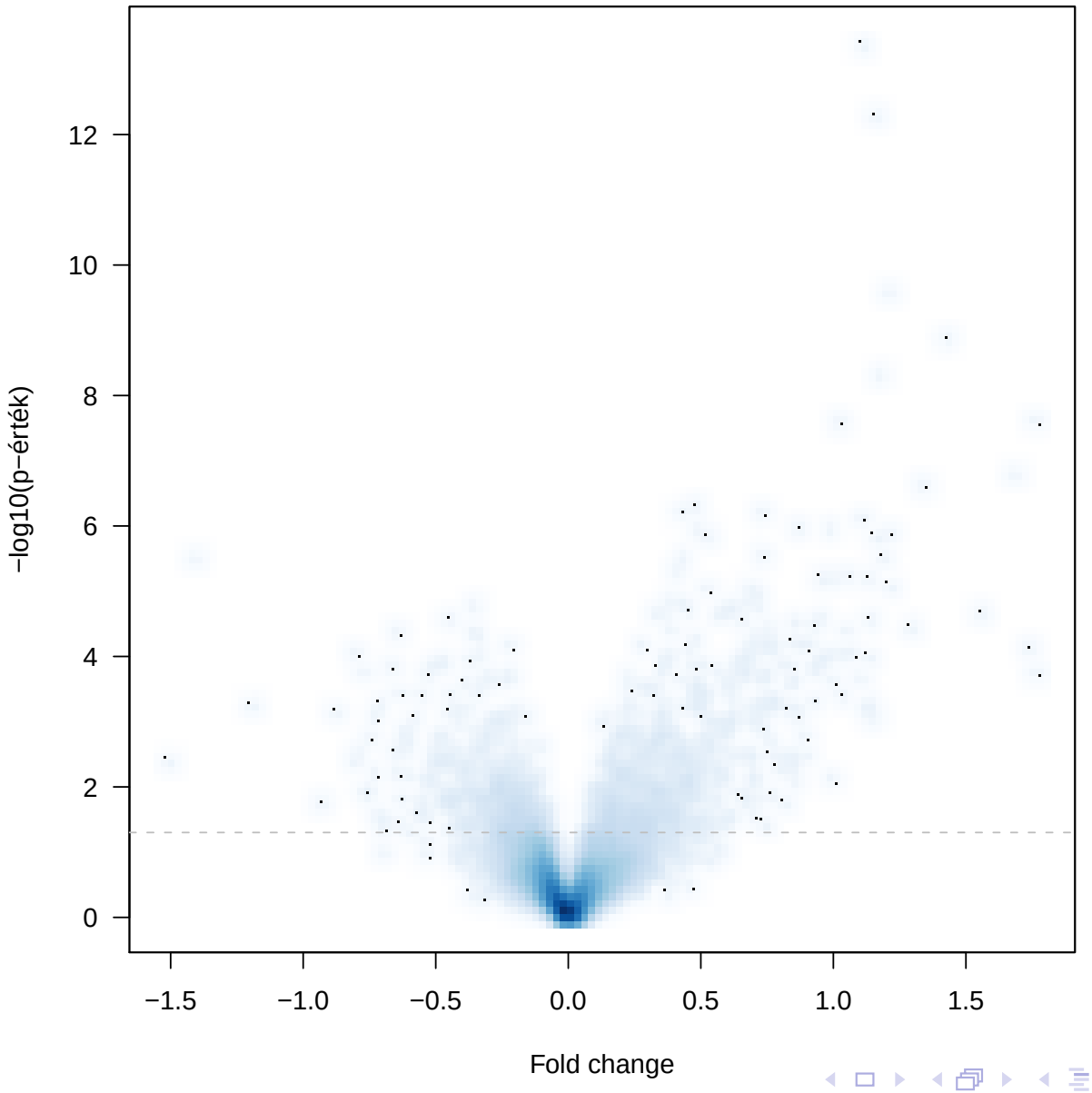
Fold change

```
> sum(abs(Fc)>=1)

[1] 36

> (dif.ps = rownames(ALL.bcr.neg.exprs.m[abs(Fc)>=1, ]))

 [1] "1211_s_at"  "1635_at"    "1636_g_at"  "1674_at"    "32434_at"
 [6] "32542_at"   "33232_at"   "33440_at"   "33462_at"   "34472_at"
[11] "35926_s_at" "36275_at"   "36536_at"   "36543_at"   "36591_at"
[16] "36617_at"   "36638_at"   "36927_at"   "37006_at"   "37014_at"
[21] "37015_at"   "37027_at"   "37043_at"   "37363_at"   "37403_at"
[26] "38052_at"   "38111_at"   "38514_at"   "38578_at"   "39329_at"
[31] "39730_at"   "40202_at"   "40504_at"   "40516_at"   "40953_at"
[36] "41123_s_at"

> sort(unique(as.character(getSYMBOL(dif.ps, 'hgu95av2.db'))))

 [1] "ABL1"    "ACTN1"   "AHNAK"   "AHR"     "ALDH1A1" "ANXA1"   "CD27"
 [8] "CNN3"    "CRADD"   "CRIP1"   "CTGF"    "ENPP2"   "F13A1"   "F3"
[15] "FHL1"    "FZD6"    "ID1"     "ID3"     "IFI44L"  "IGJ"     "IGLL1"
[22] "KLF9"    "LILRB1"  "MARCKS"  "MTSS1"   "MX1"     "P2RY14"  "PON2"
[29] "SCHIP1"  "SEMA6A"  "TUBA4A"  "VCAN"    "YES1"    "ZEB1"
```

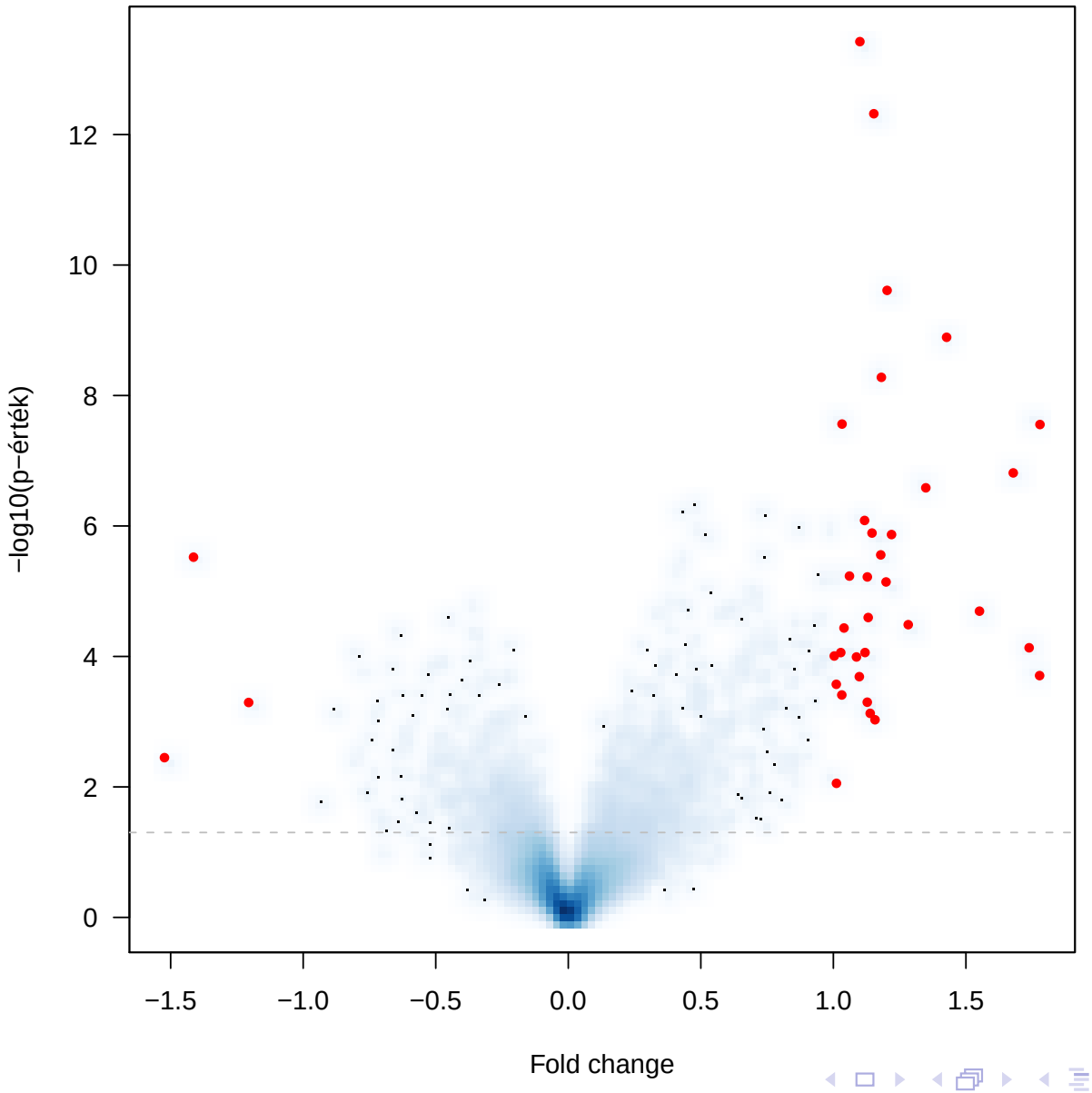
t-próba

```
> library(genefilter)
> t.eredmeny = rowttests(ALL.bcr.neg, 'mol.biol')
> names(t.eredmeny)

[1] "statistic" "dm"        "p.value"

> log.p = -log10(t.eredmeny$p.value)
> Fc = t.eredmeny$dm

> smoothScatter(Fc, log.p,
+ xlab = 'Fold change', ylab = '-log10(p-érték)', las=1)
> abline(h = -log10(0.05), col='grey', lty=2)

> szign.Fc1.ps = which(t.eredmeny$p.value<=0.05 & abs(t.eredmeny$dm)>=1)

> points(Fc[szign.Fc1.ps], log.p[szign.Fc1.ps], pch=20, col='red')
```

t-próba



t-próba



t-próba

```
> length(szign.Fc1.ps)

[1] 36

> (dif.ps = rownames(ALL.bcr.neg.exprs.m[szign.Fc1.ps, ]))

 [1] "1211_s_at"  "1635_at"    "1636_g_at"  "1674_at"    "32434_at"
 [6] "32542_at"   "33232_at"   "33440_at"   "33462_at"   "34472_at"
[11] "35926_s_at" "36275_at"   "36536_at"   "36543_at"   "36591_at"
[16] "36617_at"   "36638_at"   "36927_at"   "37006_at"   "37014_at"
[21] "37015_at"   "37027_at"   "37043_at"   "37363_at"   "37403_at"
[26] "38052_at"   "38111_at"   "38514_at"   "38578_at"   "39329_at"
[31] "39730_at"   "40202_at"   "40504_at"   "40516_at"   "40953_at"
[36] "41123_s_at"

> sort(unique(as.character(getSYMBOL(dif.ps, 'hgu95av2.db'))))

 [1] "ABL1"    "ACTN1"   "AHNAK"   "AHR"     "ALDH1A1" "ANXA1"   "CD27"
 [8] "CNN3"    "CRADD"   "CRIP1"   "CTGF"    "ENPP2"   "F13A1"   "F3"
[15] "FHL1"    "FZD6"    "ID1"     "ID3"     "IFI44L"  "IGJ"     "IGLL1"
[22] "KLF9"    "LILRB1"  "MARCKS"  "MTSS1"   "MX1"     "P2RY14"  "PON2"
[29] "SCHIP1"  "SEMA6A"  "TUBA4A"  "VCAN"    "YES1"    "ZEB1"
```

Többszörös összehasonlítás

- első fajú hiba (α) valószínűsége ↑
- korrigálni szokták, másodfajú hiba (β) ↑, erő ↓

```
> library(multtest)
> permut.t = mt.maxT(ALL.bcr.neg.exprs.m, classlabel=csoportok-1, B=1000)

b=10 b=20 b=30 b=40 b=50 b=60 b=70 b=80 b=90 b=100
b=110 b=120 b=130 b=140 b=150 b=160 b=170 b=180 b=190 b=200
b=210 b=220 b=230 b=240 b=250 b=260 b=270 b=280 b=290 b=300
b=310 b=320 b=330 b=340 b=350 b=360 b=370 b=380 b=390 b=400
b=410 b=420 b=430 b=440 b=450 b=460 b=470 b=480 b=490 b=500
b=510 b=520 b=530 b=540 b=550 b=560 b=570 b=580 b=590 b=600
b=610 b=620 b=630 b=640 b=650 b=660 b=670 b=680 b=690 b=700
b=710 b=720 b=730 b=740 b=750 b=760 b=770 b=780 b=790 b=800
b=810 b=820 b=830 b=840 b=850 b=860 b=870 b=880 b=890 b=900
b=910 b=920 b=930 b=940 b=950 b=960 b=970 b=980 b=990 b=1000

> names(permut.t)

[1] "index"    "teststat" "rawp"     "adjp"

> sum(permut.t$rawp<=0.05)

[1] 1262
```

Többszörös összehasonlítás

Westfall-Young módszere alapján:

```
> sum(permut.t$adjp<=0.05)

[1] 27

> korrigalt.szig.ps = permut.t[permut.t$adjp <= 0.05, ]
> korrigalt.szig.ps[1:12,]

           index  teststat  rawp  adjp
1636_g_at    714 -9.130386 0.001 0.001
39730_at    9823 -8.604144 0.001 0.001
1635_at      713 -7.167919 0.001 0.001
1674_at      756 -6.737666 0.001 0.001
40504_at   10604 -6.413755 0.001 0.002
40202_at   10299 -6.330859 0.001 0.002
37015_at    7082 -5.919240 0.001 0.003
37027_at    7094 -5.709362 0.001 0.003
32434_at    2456 -5.645085 0.001 0.003
40167_s_at 10263 -5.508733 0.001 0.004
40480_s_at 10579 -5.468690 0.001 0.005
39837_s_at  9930 -5.450287 0.001 0.006
```

Többszörös összehasonlítás

```
> (dif.ps = rownames(permut.t[permut.t$adjp <= 0.05, ]))

 [1] "1636_g_at"  "39730_at"   "1635_at"    "1674_at"    "40504_at"
 [6] "40202_at"   "37015_at"   "37027_at"   "32434_at"   "40167_s_at"
[11] "40480_s_at" "39837_s_at" "36591_at"   "33774_at"   "41274_at"
[16] "37403_at"   "37014_at"   "41815_at"   "37363_at"   "39631_at"
[21] "34472_at"   "32542_at"   "39329_at"   "35162_s_at" "32148_at"
[26] "33362_at"   "40855_at"

> sort(unique(as.character(getSYMBOL(dif.ps, 'hgu95av2.db'))))

 [1] "ABL1"     "ACTN1"    "ACVR2A"   "AHNAK"    "ALDH1A1"  "ANXA1"
 [7] "CASP8"    "CDC42EP3" "EMP2"     "FARP1"    "FHL1"     "FYN"
[13] "FZD6"     "KLF9"     "MARCKS"   "MINA"     "MTSS1"    "MX1"
[19] "PON2"     "SAMD4A"   "SYNE2"    "TUBA4A"   "WSB2"     "YES1"
[25] "ZNF467"
```

Nem specifikus szűrés

- Cél az összehasonlítások számának csökkentése
- Annotáció alapján
- Variabilitás alapján
  - Kis varianciájú probesetek kihagyása

```
> szorasok = esApply(ALL.bcr.neg, 1, sd)
> nagy.var.idxs = (szorasok > quantile(szorasok, 0.2))
> (nagy.var.eset = ALL.bcr.neg[nagy.var.idxs, ])
ExpressionSet (storageMode: lockedEnvironment)
assayData: 10100 features, 79 samples
  element names: exprs
phenoData
  sampleNames: 01005, 01010, ..., 84004  (79 total)
  varLabels and varMetadata description:
    cod:  Patient ID
    diagnosis:  Date of diagnosis
    ...: ...
    date last seen:  date patient was last seen
    (21 total)
featureData
  featureNames: 1000_at, 1001_at, ..., AFFX-YEL024w/RIP1_at  (10100 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

Lineáris modell

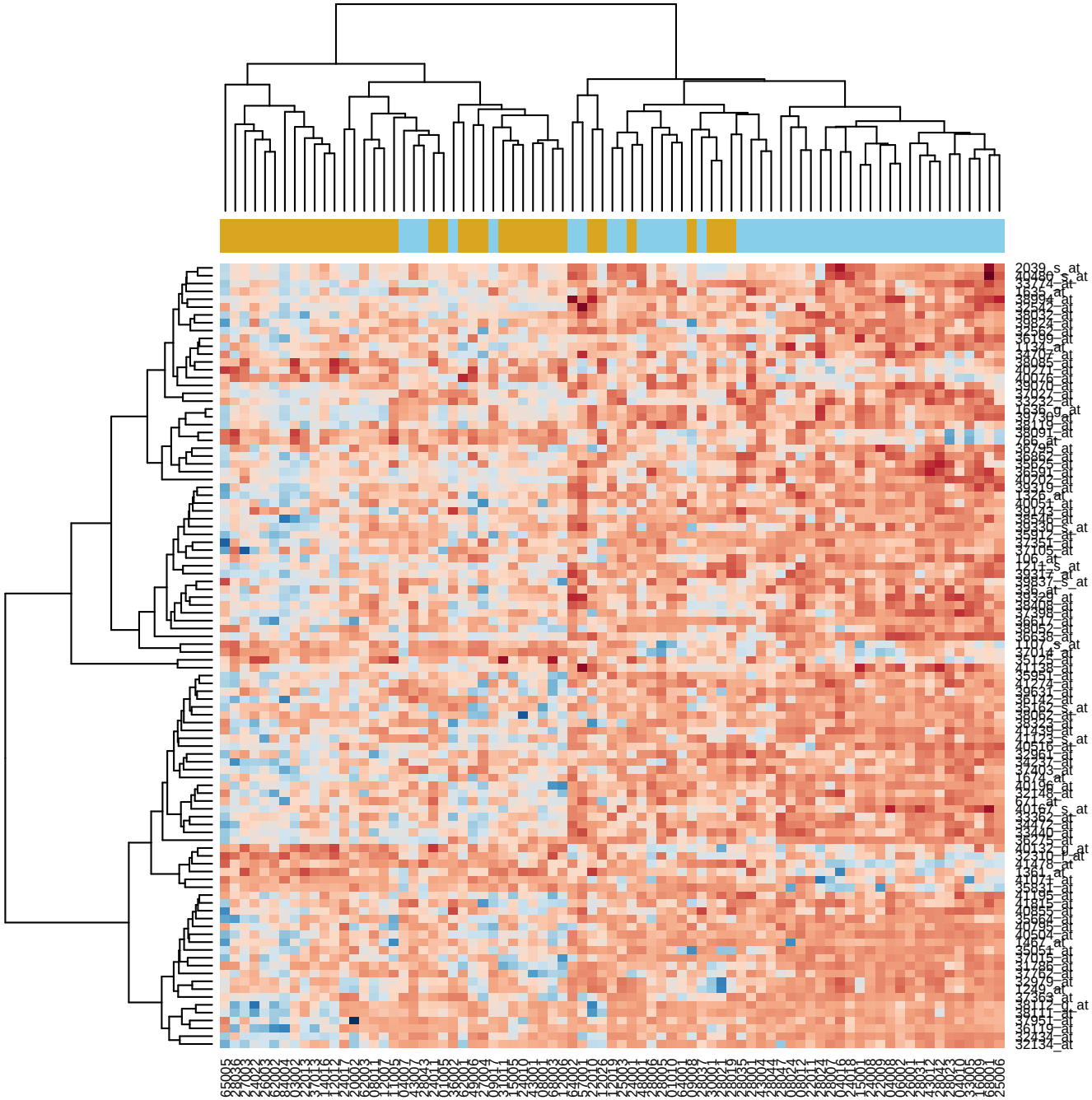
Kettő vagy több csoport esetén is használható

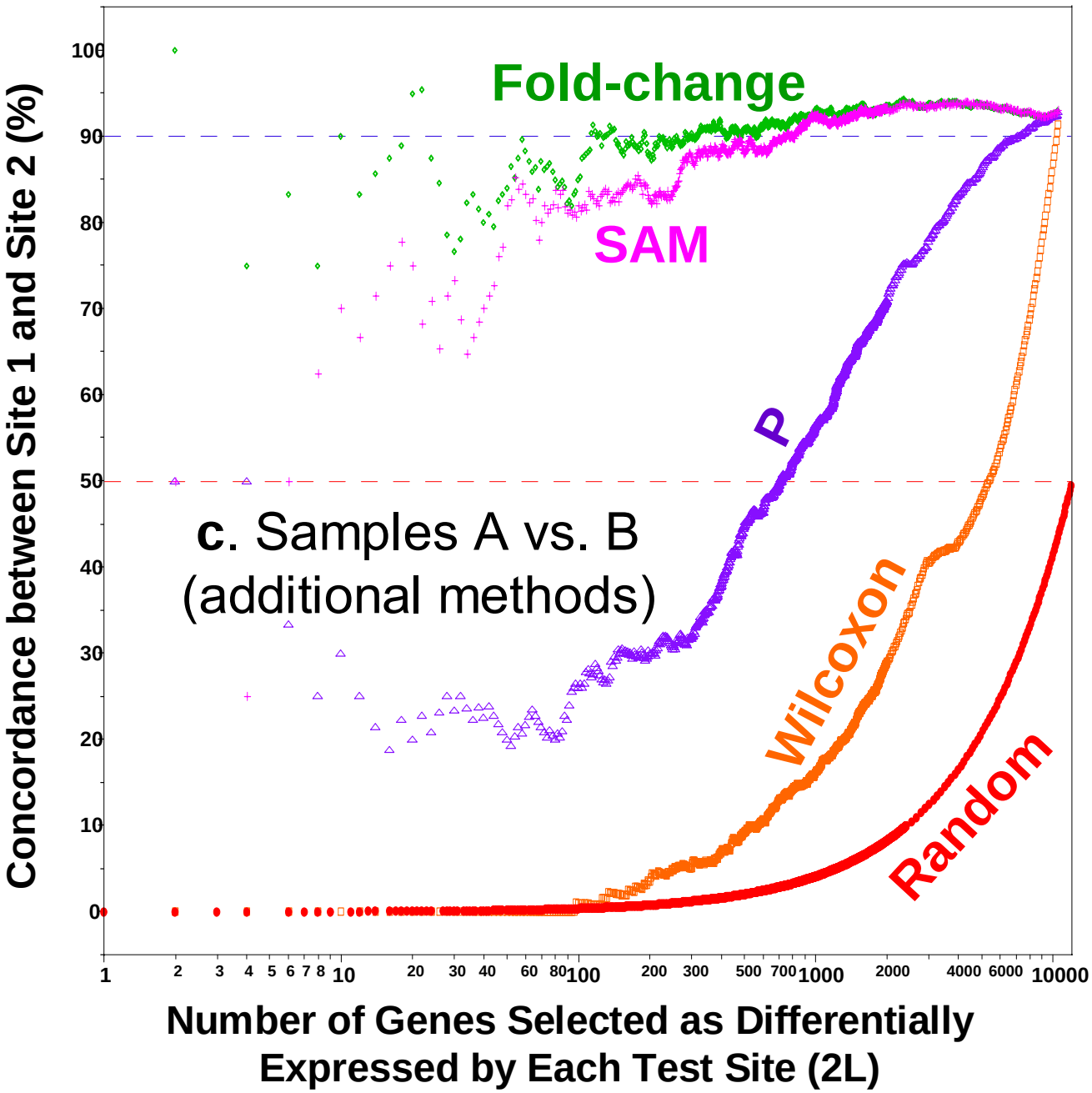
```
> library(limma)
> design = model.matrix(~factor(csoportok))
> illesztett = lmFit(ALL.bcr.neg, design)
> se.korrigalt = eBayes(illesztett)
> top100 = topTable(se.korrigalt,coef=2,adjust.method='fdr',number=100)
> top100[1:10, ]
```

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
714	1636_g_at	-1.1000116	9.196420	-9.386530	1.531812e-14	1.933913e-10	21.773880
9823	39730_at	-1.1525269	9.000049	-8.815214	2.028724e-13	1.280632e-09	19.443353
713	1635_at	-1.2026753	7.897095	-7.398075	1.208549e-10	5.085978e-07	13.636715
756	1674_at	-1.4272115	5.001771	-7.020362	6.486736e-10	2.047376e-06	12.102060
10604	40504_at	-1.1810295	4.244478	-6.683873	2.854764e-09	7.208279e-06	10.746992
10299	40202_at	-1.7793784	8.621443	-6.296601	1.536039e-08	2.868208e-05	9.206850
7082	37015_at	-1.0327017	4.330511	-6.288545	1.590294e-08	2.868208e-05	9.175072
2456	32434_at	-1.6785501	4.466311	-5.881601	9.015004e-08	1.422680e-04	7.586693
7094	37027_at	-1.3487023	8.444161	-5.749020	1.573117e-07	2.206734e-04	7.077075
9930	39837_s_at	-0.4757069	7.144313	-5.548352	3.621192e-07	4.571755e-04	6.314169

Heatmap

```
> library(RColorBrewer)

> cella.szin = colorRampPalette(brewer.pal(10, 'RdBu'))(256)
> oszlop.szin = ifelse(csoportok ==1, 'goldenrod', 'skyblue')

> e.top.m = exprs(ALL.bcr.neg[top100$ID,])

> heatmap(e.top.m, col=cella.szin, ColSideColors = oszlop.szin)
```





MAQC Consortium (2006)

GSEA

- Egyedi gének $\leftrightarrow$ géncsoportok
- Géncsoportok:
  - BioCarta, BioCyc
  - GO, GOA
  - KEGG
  - PFAM
  - Kromoszóma sáv (band)
  - Korábbi elemzések eredményei
  - Egyéb
- Egyszerű:

$$z_K = \frac{1}{\sqrt{n}} \sum_{k \in K} t_k$$

  - permutációs teszt (fenotípus)

Subramanian et al. (2005); Tian et al. (2005)

Hipergeometrikus teszt

Géncsoport	*Differenciáltan expresszált*	
	Igen	Nem
Benn	n_{11}	n_{12}
Kinn	n_{21}	n_{22}

2×2 tábla esetén:

$$OR = \frac{n_{11}}{n_{21}} / \frac{n_{12}}{n_{22}}$$

Alexa et al. (2006)

Tian

Q1: A géncsoporton belül a fenotípussal ugyanolyan összefüggést mutatnak a gének, mint azon kívül.

$$T_k = \frac{1}{m_k} \sum_{i=1}^{B} G_{ki} t_i$$

- $i = 1, \ldots, B$ gén, $j = 1, \ldots, n$ minta
- t_i az i-dik gén fenotípussal való kapcsolatát kifejező mérőszám
- $k = 1, \ldots, K$ géncsoport
- $G_{ki} = 1$ ha az i-dik gén tagja a k-dik géncsoportnak
- $m_k = \sum_{i=1}^{B} G_{ki}$ a k-dik géncsoport génjeinek száma
- Permutációs teszt (gének)
- NT_k standardizált

Tian et al. (2005)

Tian

Q2: A géncsoportban nincsen olyan gén, aminek expressziója kapcsolatban lenne a fenotípussal.

$$E_k = \frac{1}{m_k} \sum_{i=1}^{B} G_{ki} t_i$$

- $i = 1, \ldots, B$ gén, $j = 1, \ldots, n$ minta
- t_i az i-dik gén fenotípussal való kapcsolatát kifejező mérőszám
- $k = 1, \ldots, K$ géncsoport
- $G_{ki} = 1$ ha az i-dik gén tagja a k-dik géncsoportnak
- $m_k = \sum_{i=1}^{B} G_{ki}$ a k-dik géncsoport génjeinek száma
- Permutációs teszt (fenotípus)
- NE_k standardizált

Tian et al. (2005)

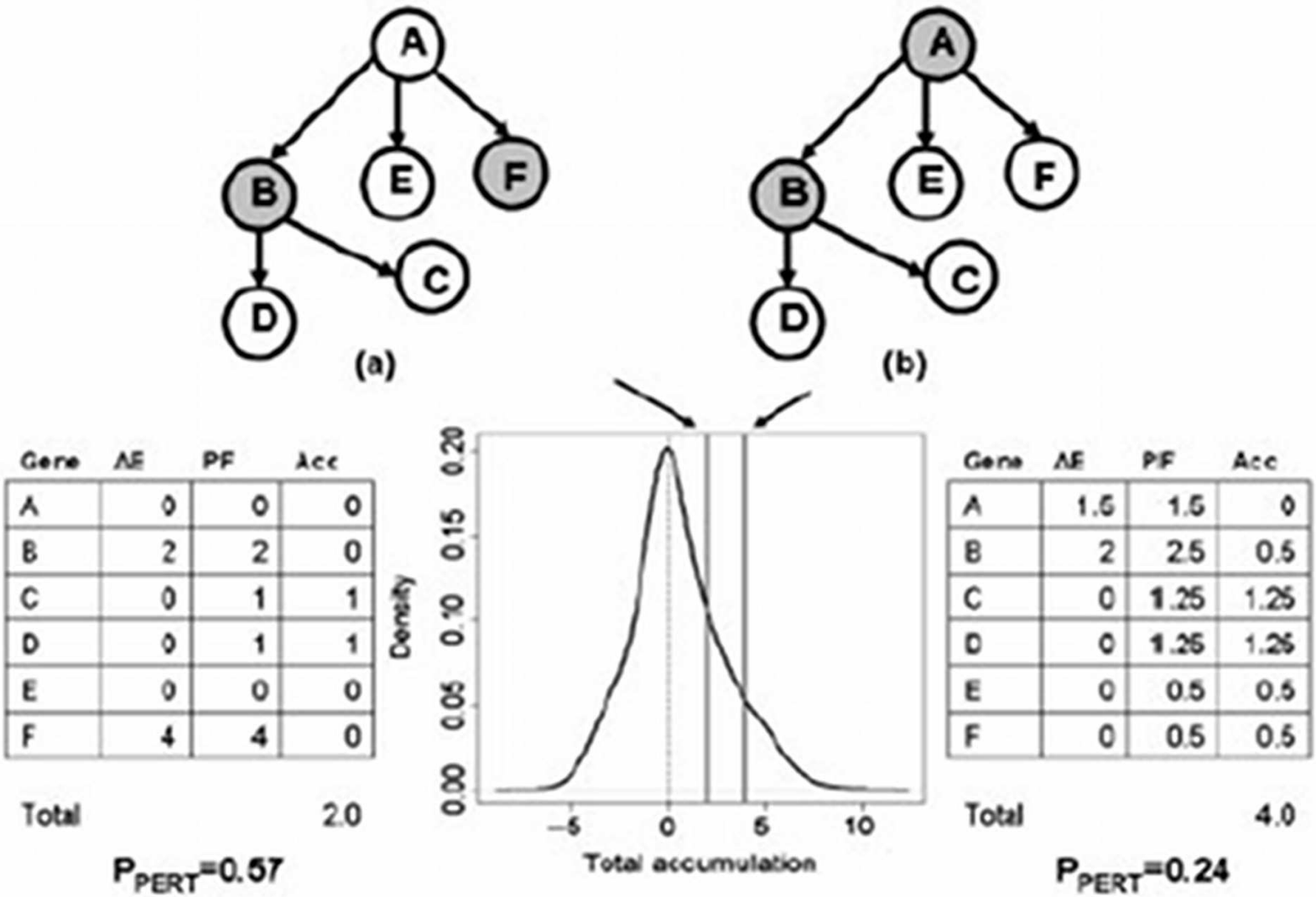
Tarca

1. A DE-gének felülreprezentáltak egy pathwayben
   - $P_{NDE} = P(X \geq N_{de}|H_0)$
2. A perturbáció mértéke
   - perturbációs faktor: $PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \times \frac{PF(g_j)}{N_{ds}(g_j)}$
     - g_i fold-change
     - $\Delta E(g_i)$ normalizált expresszióváltozás
     - irányított él $A \rightarrow B$, akkor A forrása B-nek, B célja A-nak
     - g_j forrása g_i-nek, $N_{ds}(gj)$ a g_j összes ilyen célgénjének a száma
     - β_{ij} a két gén interakciójának erőssége, pl. $+1$ aktiválás, -1 gátlás
   - $P_{PERT} = P(T_A \geq t_A|H_0)$
   - total net accumulated perturbation $t_A = \sum_i Acc(g_i)$
   - net perturbation accumulation: $Acc(g_i) = PF(g_i) - \Delta E(g_i)$
3. Global probability $P_G = c_i - c_i \times \ln(c_i)$

$$c_i = P_{NDE}(i) \times P_{PERT}(i)$$

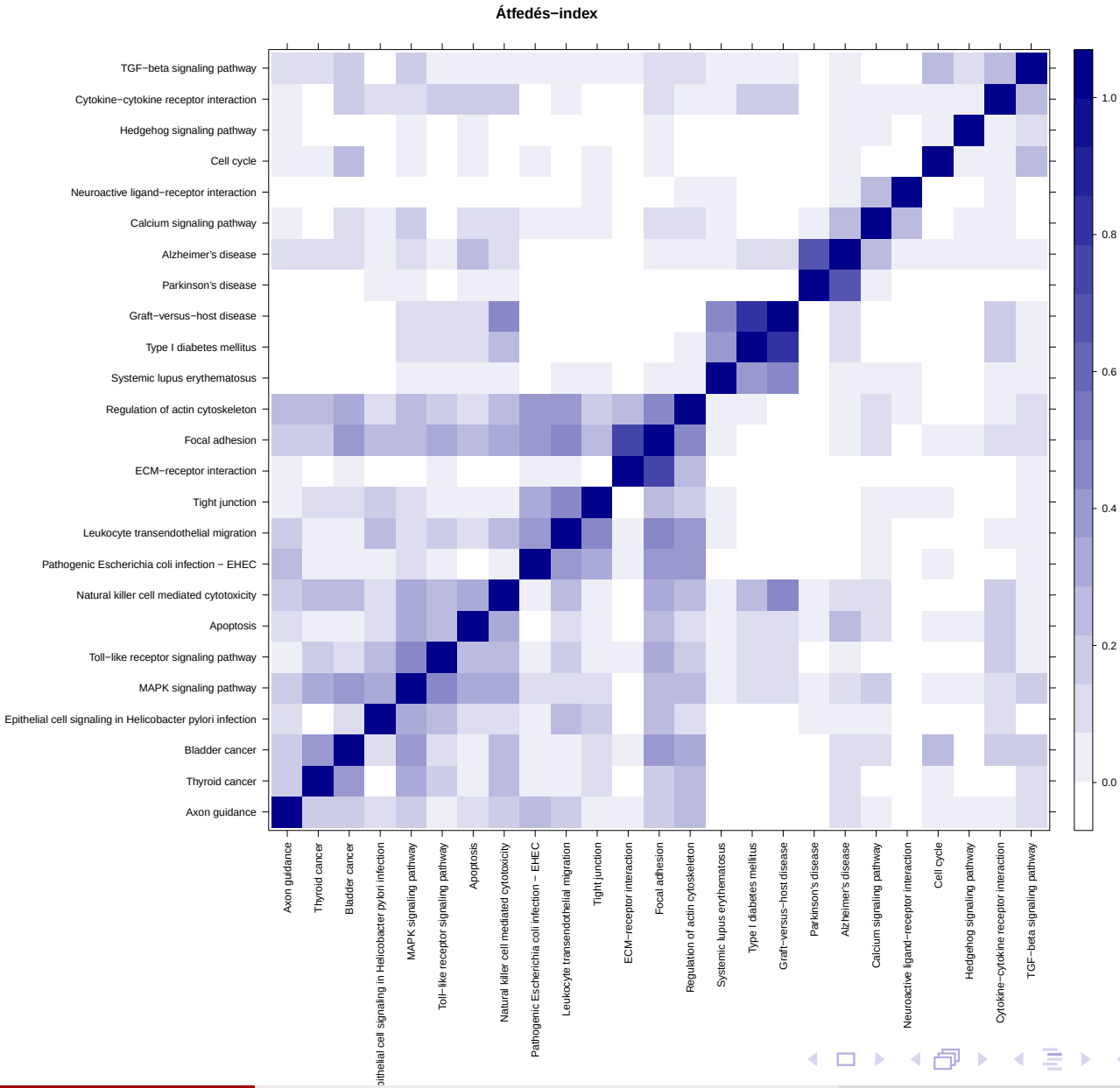
Tarca et al. (2009)

Tarca



Tarca et al. (2009)

Géncsoportok átfedése



Források

Alexa, A., J. Rahnenführer, and T. Lengauer (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics 22*(13), 1600–1607.

Dinya, E. and N. Solymosi (2016). *Biometria a klinikumban 2. Feladatok megoldása R-környezetben*. Budapest: Medicina.

Hahne, F., W. Huber, R. Gentleman, and S. Falcon (2008). *Bioconductor Case Studies*. Springer Publishing Company, Incorporated.

MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology 24*(9), 1151–1161.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS 102*(43), 15545–15550.

Tarca, A. L., S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero (2009). A novel signaling pathway impact analysis. *Bioinformatics 25*(1), 75–82.

Tian, L., S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park (2005). Discovering statistically significant pathways in expression profiling studies. *PNAS 102*(38), 13544–13549.

`http://www.bioconductor.org/pub/biocases/websupp/index.html`