The problems encountered during microarray data analysis

Joanna Zyprych

UP Poznań

Październik 4, 2009

A few words about the data...

A few words about the data...

- Acute myeloid leukemia project

A few words about the data...

- Acute myeloid leukemia project
- One microarray consists of: experimental probe - RNA sample from a patient or a healthy person and control probe - RNA isolated from cell line HL60 (a subtype of AML)

A few words about the data...

- Acute myeloid leukemia project
- One microarray consists of: experimental probe - RNA sample from a patient or a healthy person and control probe - RNA isolated from cell line HL60 (a subtype of AML)
- 86 hybridization: 1-2 HL60 versus Control,3-68 HL60 versus Leukemia,69-86 HL60 versus Control

Gpr file from GenePix for AML experiment.

Block	Column	Row	Name	Flags
1	1	1	ERG-Operon	100
1	2	1	ERG-Operon	100
1	3	1	ERG-Operon	100
1	4	1	FLT3-Operon	-50
1	5	1	FLT3-Operon	-50
1	6	1	FLT3-Operon	-50
1	7	1	GAPDHS-Operon	-50
1	8	1	GAPDHS-Operon	-50
1	9	1	GAPDHS-Operon	-50

Gpr file from GenePix for AML experiment.

Block	Column	Row	Name	Flags
1	1	1	ERG-Operon	100
1	2	1	ERG-Operon	100
1	3	1	ERG-Operon	100
1	4	1	FLT3-Operon	-50
1	5	1	FLT3-Operon	-50
1	6	1	FLT3-Operon	-50
1	7	1	GAPDHS-Operon	-50
1	8	1	GAPDHS-Operon	-50
1	9	1	GAPDHS-Operon	-50

- The last column gives us the specified knowledge which weights should be given to spots

Gpr file from GenePix for AML experiment.

Block	Column	Row	Name	Flags
1	1	1	ERG-Operon	100
1	2	1	ERG-Operon	100
1	3	1	ERG-Operon	100
1	4	1	FLT3-Operon	-50
1	5	1	FLT3-Operon	-50
1	6	1	FLT3-Operon	-50
1	7	1	GAPDHS-Operon	-50
1	8	1	GAPDHS-Operon	-50
1	9	1	GAPDHS-Operon	-50

- The last column gives us the specified knowledge which weights should be given to spots
- For flags less than the cutoff value we give weights equal 0 and 1 otherwise

Gpr file from GenePix for AML experiment.

Block	Column	Row	Name	Flags
1	1	1	ERG-Operon	100
1	2	1	ERG-Operon	100
1	3	1	ERG-Operon	100
1	4	1	FLT3-Operon	-50
1	5	1	FLT3-Operon	-50
1	6	1	FLT3-Operon	-50
1	7	1	GAPDHS-Operon	-50
1	8	1	GAPDHS-Operon	-50
1	9	1	GAPDHS-Operon	-50

- The last column gives us the specified knowledge which weights should be given to spots
- For flags less than the cutoff value we give weights equal 0 and 1 otherwise
- We choose cutoff=-50 to downweight bad or absent spots

Problem Number One

Problem

How to calculate the mean intensity for each gene taking into consideration the weight of the spot?

Problem Number One

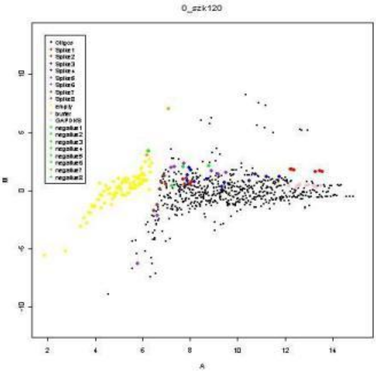
Problem

How to calculate the mean intensity for each gene taking into consideration the weight of the spot?

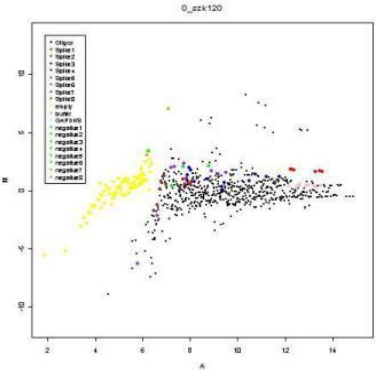
R code

```
> # MA.A - data after normalization
> Mean_intesity <- avedups(MA.A, ndups=3,
weights=MA.A$weights)
```

MA plots: before and after using avedups function

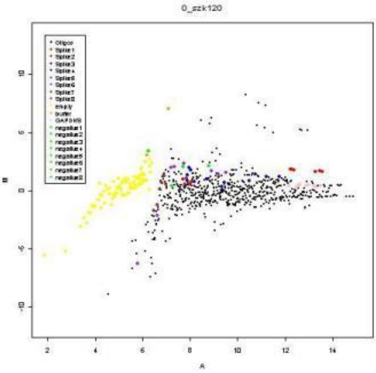


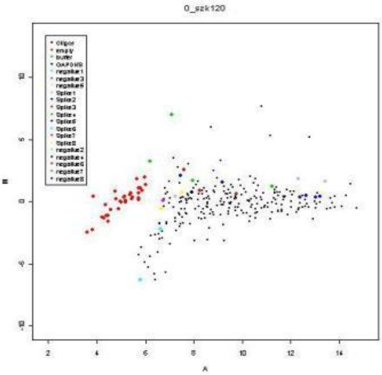
MA plots: before and after using avedups function



before using avedups function

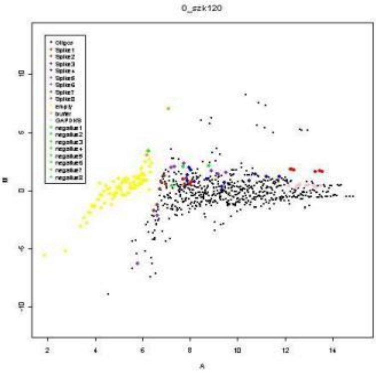
MA plots: before and after using avedups function



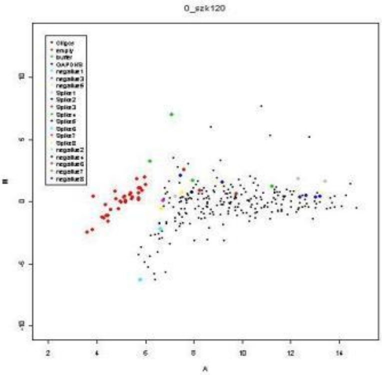


before using avedups function

MA plots: before and after using avedups function



before using avedups function



after using avedups function

Problem Number Two

Question

Which genes are over(under)expressed comparing leukemia and control probe?

Problem Number Two

Question

Which genes are over(under)expressed comparing leukemia and control probe?

The statistics used for these calculations are:

Problem Number Two

Question

Which genes are over(under)expressed comparing leukemia and control probe?

The statistics used for these calculations are:

- two sample t-statistics

Problem Number Two

Question

Which genes are over(under)expressed comparing leukemia and control probe?

The statistics used for these calculations are:

- two sample t-statistics
- sam-statistics

Problem Number Two

Question

Which genes are over(under)expressed comparing leukemia and control probe?

The statistics used for these calculations are:

- two sample t-statistics
- sam-statistics
- fc-statistics

DEDS

DEDS package

Yuanyuan Xiao and Yee Hwa Yang
April 21, 2009
University of California

DEDS

DEDS package

Yuanyuan Xiao and Yee Hwa Yang
April 21, 2009
University of California

deds.stat.linkC(X, L, B, tests = c(”t”, ”fc”, ”sam”,”...”))

DEDS

DEDS package

Yuanyuan Xiao and Yee Hwa Yang
April 21, 2009
University of California

deds.stat.linkC(X, L, B, tests = c(”t”, ”fc”, ”sam”,”...”))

- X: A matrix, in the case of gene expression data, rows correspond to N genes and columns to p mRNA samples

DEDS

DEDS package

Yuanyuan Xiao and Yee Hwa Yang
April 21, 2009
University of California

deds.stat.linkC(X, L, B, tests = c(”t”, ”fc”, ”sam”,”...”))

- X: A matrix, in the case of gene expression data, rows correspond to N genes and columns to p mRNA samples
- L: A vector of integers corresponding to observation (column) class labels

DEDS

DEDS package

Yuanyuan Xiao and Yee Hwa Yang
April 21, 2009
University of California

deds.stat.linkC(X, L, B, tests = c(”t”, ”fc”, ”sam”,”...”))

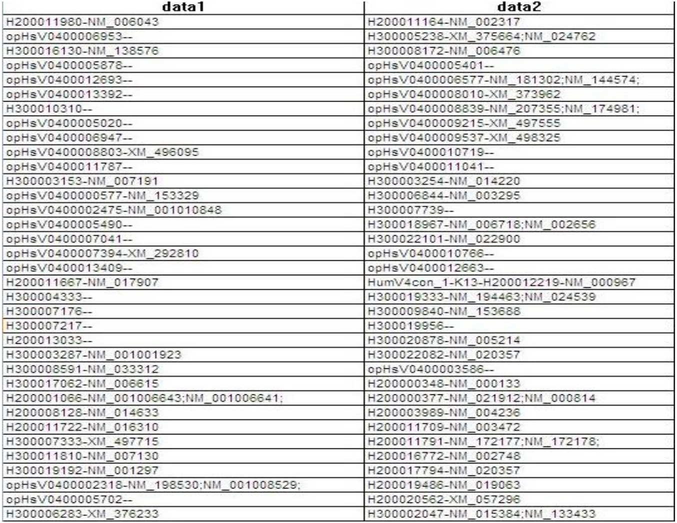
- X: A matrix, in the case of gene expression data, rows correspond to N genes and columns to p mRNA samples
- L: A vector of integers corresponding to observation (column) class labels
- B: The number of permutations

Solution II

R code

```
> library(DEDS)
> # from targets file 0-control, 1-leukemia
> L<-rep(c(0,1,0),c(2,66,18))
> data<-as.matrix(Mean_intesity)
> d <- deds.stat.linkC(data, L, B=200)
> # for the comparisons between the 3 statistics
> t_genes<-topgenes(d,number=50,Mean_intesity$genes$Name,
+ sort.by="t")
> fc_genes<-topgenes(d,number=50,Mean_intesity$genes$Name,
+ sort.by="fc")
> sam_genes<-topgenes(d,number=50,Mean_intesity$genes$Name,
+ sort.by="sam")
```

Data from DEDS...



data1	data2
H200011980-NM_006043	H200011164-NM_002317
opHsV0400006953--	H300005238-XM_375664;NM_024762
H300016130-NM_138576	H300008172-NM_006476
opHsV0400005878--	opHsV0400005401--
opHsV0400012693--	opHsV0400006577-NM_181302;NM_144574;
opHsV0400013392--	opHsV0400008010-XM_373962
H300010310--	opHsV0400008839-NM_207355;NM_174981;
opHsV0400005020--	opHsV0400009215-XM_497555
opHsV0400006947--	opHsV0400009537-XM_498325
opHsV0400008803-XM_496095	opHsV0400010719--
opHsV0400011787--	opHsV0400011041--
H300003153-NM_007191	H300003254-NM_014220
opHsV0400000577-NM_153329	H300006844-NM_003295
opHsV0400002475-NM_001010848	H300007739--
opHsV0400005490--	H300018967-NM_006718;NM_002656
opHsV0400007041--	H300022101-NM_022900
opHsV0400007394-XM_292810	opHsV0400010766--
opHsV0400013409--	opHsV0400012663--
H200011667-NM_017907	HumV4con_1-K13-H200012219-NM_000967
H300004333--	H300019333-NM_194463;NM_024539
H300007176--	H300009840-NM_153688
H300007217--	H300019956--
H200013033--	H300020878-NM_005214
H300003287-NM_001001923	H300022082-NM_020357
H300008591-NM_033312	opHsV0400003586--
H300017062-NM_006615	H200000348-NM_000133
H200001066-NM_001006643;NM_001006641;	H200000377-NM_021912;NM_000814
H200008128-NM_014633	H200003989-NM_004236
H200011722-NM_016310	H200011709-NM_003472
H300007333-XM_497715	H200011791-NM_172177;NM_172178;
H300011810-NM_007130	H200016772-NM_002748
H300019192-NM_001297	H200017794-NM_020357
opHsV0400002318-NM_198530;NM_001008529;	H200019486-NM_019063
opHsV0400005702--	H200020562-XM_057296
H300006283-XM_376233	H300002047-NM_015384;NM_133433

Venn diagram

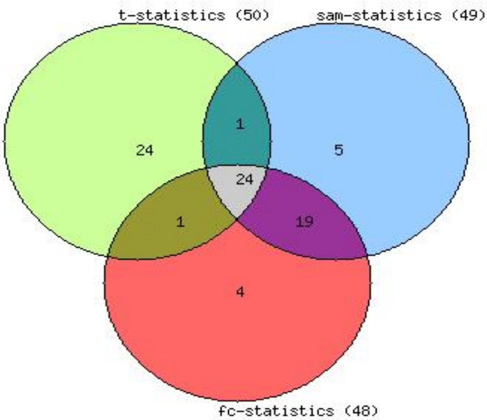
R code

```
> w<-c(data1,data2)
+ hm<-duplicated(w)
```

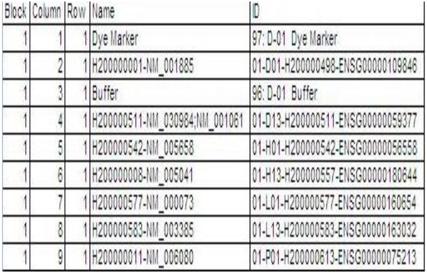
Venn diagram

R code

```
> w<-c(data1,data2)
+ hm<-duplicated(w)
```

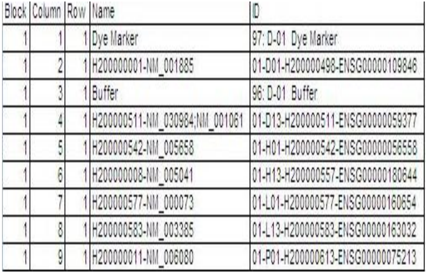


Another example



Block	Column	Row	Name	ID
1	1	1	Dye Marker	97: D-01 Dye Marker
1	2	1	H200000001-NM_001885	01-D01-H200000498-ENSG00000109846
1	3	1	Buffer	96: D-01 Buffer
1	4	1	H200000511-NM_030984;NM_001061	01-D13-H200000511-ENSG00000059377
1	5	1	H200000542-NM_005658	01-H01-H200000542-ENSG00000056558
1	6	1	H200000008-NM_005041	01-H13-H200000557-ENSG00000180644
1	7	1	H200000577-NM_000073	01-L01-H200000577-ENSG00000160654
1	8	1	H200000583-NM_003385	01-L13-H200000583-ENSG00000163032
1	9	1	H200000011-NM_006080	01-P01-H200000613-ENSG00000075213

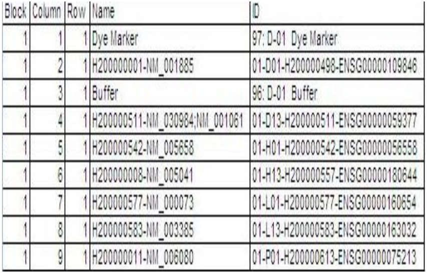
Another example



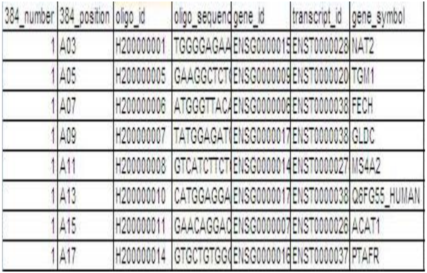
Block	Column	Row	Name	ID
1	1	1	Dye Marker	97: D-01 Dye Marker
1	2	1	H200000001-NM_001885	01-D01-H200000498-ENSG00000109846
1	3	1	Buffer	96: D-01 Buffer
1	4	1	H200000511-NM_030984;NM_001061	01-D13-H200000511-ENSG00000059377
1	5	1	H200000542-NM_005658	01-H01-H200000542-ENSG00000056558
1	6	1	H200000008-NM_005041	01-H13-H200000557-ENSG00000180644
1	7	1	H200000577-NM_000073	01-L01-H200000577-ENSG00000160654
1	8	1	H200000583-NM_003385	01-L13-H200000583-ENSG00000163032
1	9	1	H200000011-NM_006080	01-P01-H200000613-ENSG00000075213

GPR data

Another example



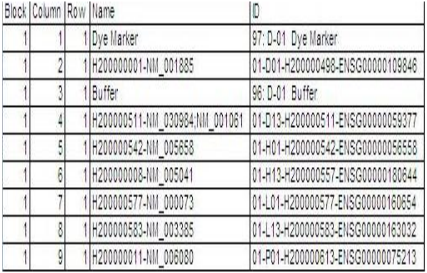
Block	Column	Row	Name	ID
1	1	1	Dye Marker	97: D-01 Dye Marker
1	2	1	H200000001-NM_001885	01-D01-H200000498-ENSG00000109846
1	3	1	Buffer	96: D-01 Buffer
1	4	1	H200000511-NM_030984;NM_001061	01-D13-H200000511-ENSG00000059377
1	5	1	H200000542-NM_005658	01-H01-H200000542-ENSG00000056558
1	6	1	H200000008-NM_005041	01-H13-H200000557-ENSG00000180644
1	7	1	H200000577-NM_000073	01-L01-H200000577-ENSG00000160654
1	8	1	H200000583-NM_003385	01-L13-H200000583-ENSG00000163032
1	9	1	H200000011-NM_006080	01-P01-H200000613-ENSG00000075213



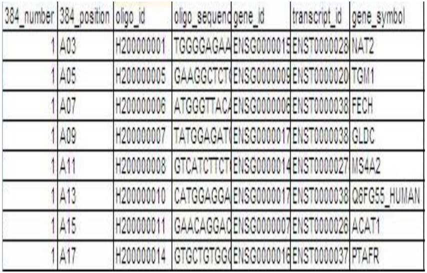
384_number	384_position	oligo_id	oligo_sequenc	gene_id	transcript_id	gene_symbol
1	A03	H200000001	TGGGGAGAA	ENSG0000015	ENST0000028	NAT2
1	A05	H200000005	GAAGGCTCT	ENSG0000009	ENST0000020	TGM1
1	A07	H200000006	ATGGGTTACA	ENSG0000006	ENST0000038	FECH
1	A09	H200000007	TATGGAGAT	ENSG0000017	ENST0000038	GLDC
1	A11	H200000008	GTCATCTTCT	ENSG0000014	ENST0000027	MS4A2
1	A13	H200000010	CATGGAGGA	ENSG0000017	ENST0000038	Q6FG55_HUMAN
1	A15	H200000011	GAACAGGAC	ENSG0000007	ENST0000026	ACAT1
1	A17	H200000014	GTGCTGTGG	ENSG0000016	ENST0000037	PTAFR

GPR data

Another example



Block	Column	Row	Name	ID
1	1	1	Dye Marker	97: D-01 Dye Marker
1	2	1	H200000001-NM_001885	01-D01-H200000498-ENSG00000109846
1	3	1	Buffer	96: D-01 Buffer
1	4	1	H200000511-NM_030984;NM_001061	01-D13-H200000511-ENSG00000059377
1	5	1	H200000542-NM_005658	01-H01-H200000542-ENSG00000056558
1	6	1	H200000008-NM_005041	01-H13-H200000557-ENSG00000180644
1	7	1	H200000577-NM_000073	01-L01-H200000577-ENSG00000160654
1	8	1	H200000583-NM_003385	01-L13-H200000583-ENSG00000163032
1	9	1	H200000011-NM_006080	01-P01-H200000613-ENSG00000075213



384_number	384_position	oligo_id	oligo_sequenc	gene_id	transcript_id	gene_symbol
1	A03	H200000001	TGGGGAGAA	ENSG0000015	ENST0000028	NAT2
1	A05	H200000005	GAAGGCTCT	ENSG0000009	ENST0000020	TGM1
1	A07	H200000006	ATGGGTTAC	ENSG0000006	ENST0000038	FECH
1	A09	H200000007	TATGGAGAT	ENSG0000017	ENST0000038	GLDC
1	A11	H200000008	GTCATCTTCT	ENSG0000014	ENST0000027	MS4A2
1	A13	H200000010	CATGGAGGA	ENSG0000017	ENST0000038	Q6FG55_HUMAN
1	A15	H200000011	GAACAGGA	ENSG0000007	ENST0000026	ACAT1
1	A17	H200000014	GTGCTGTGG	ENSG0000016	ENST0000037	PTAFR

GPR data

GAL data

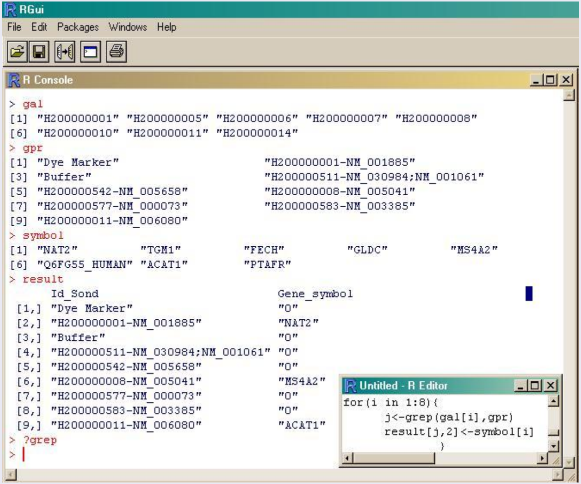
Solution III

R code

```
> gal<-read.table("gal.csv",dec=",", sep=";")
> gpr<-read.table("gpr.csv",dec=",", sep=";")
> gal<-gal[,3]
> gal<-as.character(gal)
> gpr<-gpr[,4]
> gpr<-as.character(gpr)
> symbol<-gal[,9]
> symbol<-as.character(symbol)
> result<-matrix(0,length(gpr),2)
> result[,1]<-gpr
> colnames(result)<-c("Sonda","Gen_symbol")
```

Data after grep function

Finally we obtain id sond in the first column and the gene symbol in the second



```
> gal
[1] "H200000001" "H200000005" "H200000006" "H200000007" "H200000008"
[6] "H200000010" "H200000011" "H200000014"
> gpr
[1] "Dye Marker"                    "H200000001-NM_001885"
[3] "Buffer"                        "H200000511-NM_030984;NM_001061"
[5] "H200000542-NM_005658"          "H200000008-NM_005041"
[7] "H200000577-NM_000073"          "H200000583-NM_003385"
[9] "H200000011-NM_006080"
> symbol
[1] "NAT2"        "TGM1"        "FECH"        "GLDC"        "MS4A2"
[6] "Q6FG55_HUMAN" "ACAT1"       "PTAFR"
> result
      Id_Sond                          Gene_symbol
 [1,] "Dye Marker"                     "0"
 [2,] "H200000001-NM_001885"           "NAT2"
 [3,] "Buffer"                         "0"
 [4,] "H200000511-NM_030984;NM_001061" "0"
 [5,] "H200000542-NM_005658"           "0"
 [6,] "H200000008-NM_005041"           "MS4A2"
 [7,] "H200000577-NM_000073"           "0"
 [8,] "H200000583-NM_003385"           "0"
 [9,] "H200000011-NM_006080"           "ACAT1"
> ?grep
>
```

```
for(i in 1:8){
      j<-grep(gal[i],gpr)
      result[j,2]<-symbol[i]
            }
```

:-)