

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Introduction to the affxparser package

Henrik Bengtsson - hb@stat.berkeley.edu

James Bullard - bullard@berkeley.edu

Kasper Hansen - khansen@stat.berkeley.edu

University of California, Berkeley

May 11, 2006

The affxparser package:

- ▶ R package for reading Affymetrix data files.
- ▶ Provides users and other package a generic low-level interface to data.
- ▶ Currently supported file formats: CDF and CEL (ascii & binary)
- ▶ Fast and memory efficient.
- ▶ Cross platform.
- ▶ Uses Affymetrix open-source Fusion SDK internally (same as in their commercial software)

Installation

To install, in R:

```
source("http://www.bioconductor.org/biocLite.R")  
biocLite("affxparser")
```

For latest patched version:

```
source("http://www.braju.com/R/hbLite.R")  
hbLite("affxparser")
```

For help:

```
library(affxparser)  
?affxparser
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

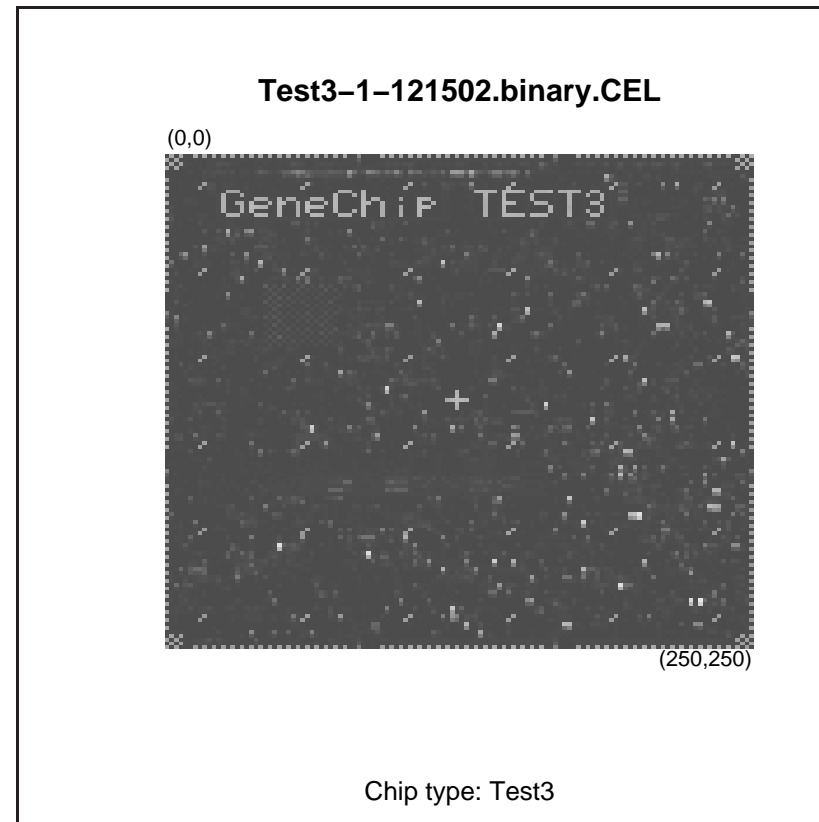
Summary

A first example

1. Copy a CEL file to the working directory of R.
2. Copy the corresponding CDF file to same directory.

3. In R, call:

```
library(affxparser)  
example(readCelRectangle)
```



CEL: **C**ell intensity file. One for each hybridization.

CDF: **C**hip **D**escription **F**ile.

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading PM and MM intensities for a probeset (unit)

First, make sure the corresponding CDF file is the working directory!

```
> cel <- readCelUnits("array1.CEL", units=1)
> cel
$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'$intensities
 [1] 5433.0 3904.8 3161.5 4356.3 1765.0 3417.8 4106.0 7888.3 3906.8
[10] 6054.5 3472.0 4188.5 1475.0 1598.0 1688.5 1435.3 3970.0 1211.0
[19] 2321.0 3718.0  866.8  937.0  848.0  898.0 1156.0  899.0 1370.5
[28] 1305.8 1259.0 1084.0 1401.0 1349.5 1658.0 1055.0 2062.0 1260.0
[37]  990.3 1176.8 1823.3 1778.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

```
> cel <- readCelUnits("array1.CEL", units=1, stratifyBy="pm")
> cel
$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'$intensities
 [1] 5433.0 3161.5 1765.0 4106.0 3906.8 3472.0 1475.0 1688.5 3970.0
[10] 2321.0  866.8  848.0 1156.0 1370.5 1259.0 1401.0 1658.0 2062.0
[19]  990.3 1823.3
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading PM and MM intensities for a probeset (unit)

Make sure the corresponding CDF file is the working directory!

```
> cel <- readCelUnits("array1.CEL", units=1, stratifyBy="pmmm")
> cel
$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at' $'AFFX-MurIL2_at'
$'AFFX-MurIL2_at' $'AFFX-MurIL2_at' $intensities
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 5433.0 3161.5 1765.0 4106.0 3906.8 3472.0 1475 1688.5 3970 2321
[2,] 3904.8 4356.3 3417.8 7888.3 6054.5 4188.5 1598 1435.3 1211 3718
      [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,] 866.8 848 1156 1370.5 1259 1401.0 1658 2062 990.3 1823.3
[2,] 937.0 898 899 1305.8 1084 1349.5 1055 1260 1176.8 1778.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading PM and MM intensities for a probeset (unit)

```
> cel <- readCelUnits("array1.CEL", units=1, stratifyBy="pmmm",
                      addDimnames=TRUE)

> cel
$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`$intensities
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
pm 5433.0 3161.5 1765.0 4106.0 3906.8 3472.0 1475 1688.5 3970 2321
mm 3904.8 4356.3 3417.8 7888.3 6054.5 4188.5 1598 1435.3 1211 3718
      [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
pm 866.8    848    1156 1370.5 1259 1401.0 1658 2062 990.3 1823.3
mm 937.0    898    899 1305.8 1084 1349.5 1055 1260 1176.8 1778.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading multiple arrays

When passing more than one CEL filename, an additional dimension is added:

```
> files <- c("array1.CEL", "array2.CEL", "array3.CEL")
> cel <- readCelUnits(files, units=1)
> cel
$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`$intensities
      [,1] [,2] [,3]
[1,] 5433.0 3786.5 5266.8
[2,] 3904.8 2732.0 2501.0
[3,] 3161.5 1686.0 2628.5
[4,] 4356.3 2438.8 3743.0
...
[38,] 1176.8 1035.3  752.0
[39,] 1823.3 1281.3 1597.0
[40,] 1778.0 1034.0 1275.8
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading multiple arrays

When passing more than one CEL filename, an additional dimension is added:

```
> files <- c("array1.CEL", "array2.CEL", "array3.CEL")
> cel <- readCelUnits(files, units=1)
> cel
$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'
$'AFFX-MurIL2_at'$'AFFX-MurIL2_at'$intensities
      [,1] [,2] [,3]
[1,] 5433.0 3786.5 5266.8
[2,] 3904.8 2732.0 2501.0
[3,] 3161.5 1686.0 2628.5
[4,] 4356.3 2438.8 3743.0
...
[38,] 1176.8 1035.3 752.0
[39,] 1823.3 1281.3 1597.0
[40,] 1778.0 1034.0 1275.8
```

Plot the probe intensities for two of the arrays:

```
> y <- cel[["AFFX-MurIL2_at"]][[1]]$intensities
> plot(y[,1:2], xlim=c(0,6000), ylim=c(0,6000))
> abline(a=0, b=1)
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

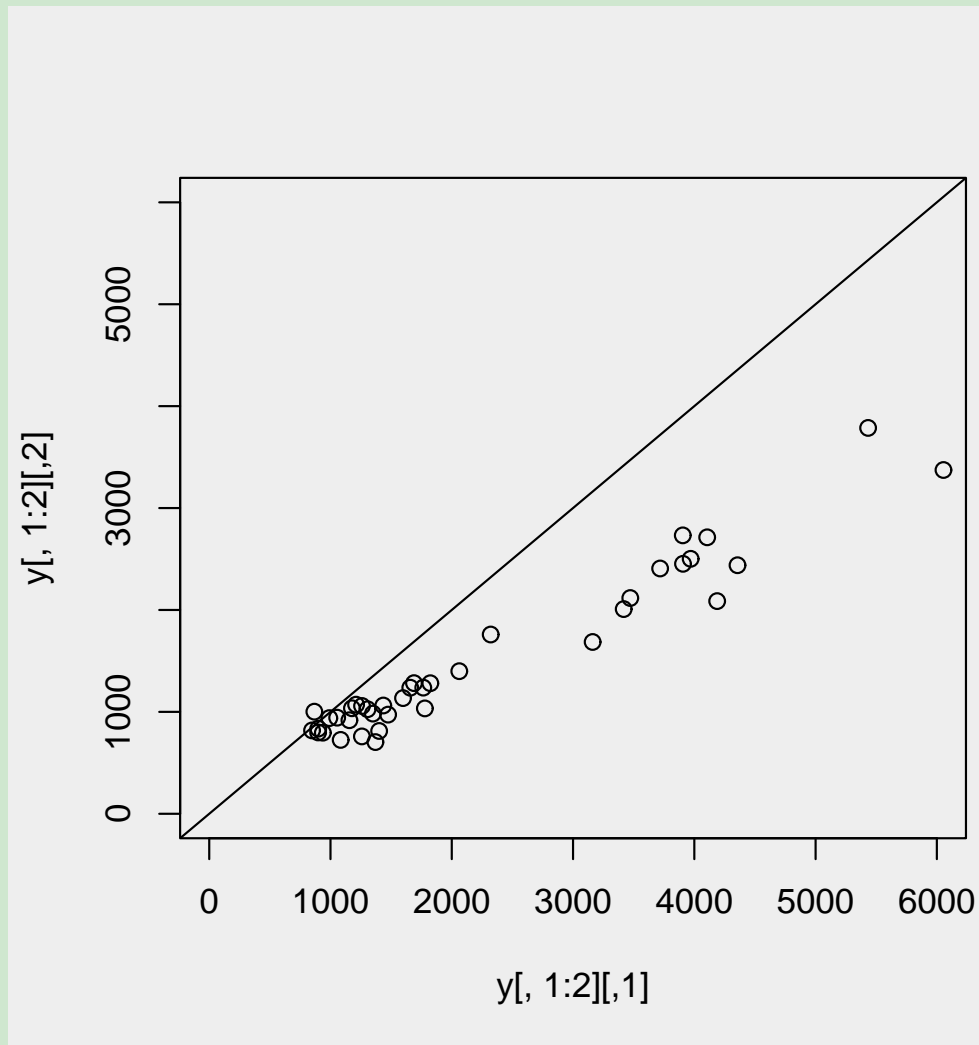
Reading multiple arrays

When passing more than one CEL filename, an additional dimension is added:

```
> files <- c("array1.CEL", "array2.CEL", "array3.CEL")
```

```
> y <- readAffy(files, units=1)
```

Example



Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

**Reading
multiple
arrays**

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading multiple arrays (continued)

The last dimension is always the array dimension:

```
> cel <- readCelUnits(files, units=1, stratifyBy="pmmm")
> cel
$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`
$`AFFX-MurIL2_at`$`AFFX-MurIL2_at`$intensities
, , 1
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 5433.0 3161.5 1765.0 4106.0 3906.8 3472.0 1475 1688.5 3970 2321
[2,] 3904.8 4356.3 3417.8 7888.3 6054.5 4188.5 1598 1435.3 1211 3718
      [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,] 866.8 848 1156 1370.5 1259 1401.0 1658 2062 990.3 1823.3
[2,] 937.0 898 899 1305.8 1084 1349.5 1055 1260 1176.8 1778.0
, , 2
...
, , 3
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 5266.8 2628.5 1526.0 3059.0 3930.3 2189 805.0 1236.5 2405.0 1821
[2,] 2501.0 3743.0 2710.8 5528.5 4791.0 2128 966.8 932.3 881.8 2192
      [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,] 641.5 561 739 911.5 646 674 1228.8 1662.5 647 1597.0
[2,] 596.3 517 686 785.5 558 637 649.8 884.0 752 1275.8
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure
Reading
CEL files
using CDF
Rearranging
CDF
structure

SNP chips

Summary

Reading multiple arrays (continued)

```
> cel <- readCelUnits(files, units=1, stratifyBy="pmmm")  
> y <- cel[["AFFX-MurIL2_at"]][[1]]$intensities
```

All PMs on all arrays:

```
> y[1,,]  
      [,1]  [,2]  [,3]  
[1,] 5433.0 3786.5 5266.8  
[2,] 3161.5 1686.0 2628.5  
[3,] 1765.0 1238.0 1526.0  
[4,] 4106.0 2713.0 3059.0  
...  
[18,] 2062.0 1399.3 1662.5  
[19,]  990.3  938.8  647.0  
[20,] 1823.3 1281.3 1597.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

**Reading
multiple
arrays**

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading multiple arrays (continued)

```
> cel <- readCelUnits(files, units=1, stratifyBy="pmmm")
> y <- cel[["AFFX-MurIL2_at"]][[1]]$intensities
```

All PMs on all arrays:

```
> y[1,,]
      [,1] [,2] [,3]
[1,] 5433.0 3786.5 5266.8
[2,] 3161.5 1686.0 2628.5
[3,] 1765.0 1238.0 1526.0
[4,] 4106.0 2713.0 3059.0
...
[18,] 2062.0 1399.3 1662.5
[19,]  990.3  938.8  647.0
[20,] 1823.3 1281.3 1597.0
```

All PMs and MMs on array #2:

```
> y[, ,2]
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 3786.5 1686.0 1238.0 2713 2453.0 2117.0  973 1282.0 2502.0 1759
[2,] 2732.0 2438.8 2008.5 4079 3373.3 2086.8 1135 1062.3 1070.8 2407
      [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,] 1002.0   818 918.0  703.5   759 812.8  1237 1399.3  938.8 1281.3
[2,]  795.8   798 835.5 1026.0   725 983.0   943 1058.0 1035.3 1034.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading multiple probesets on multiple arrays.

```
> cel <- readCelUnits(files, units=81:99, stratifyBy="pmmm")
> names(cel)
 [1] "31320_at"    "31321_at"    "31322_at"    "31323_r_at"  "31324_at"
 [6] "31325_at"    "31326_at"    "31327_at"    "31328_at"    "31329_at"
[11] "31330_at"    "31331_at"    "31332_at"    "31333_at"    "31334_at"
[16] "31335_at"    "31336_at"    "31337_at"    "31338_at"
> str(cel)
 $ 31320_at  :List of 1
  ..$ 31320_at:List of 1
  .. ..$ intensities: num [1:2, 1:16, 1:3] 10718 22342 6565 7588 5463 ...
 $ 31321_at  :List of 1
  ..$ 31321_at:List of 1
  .. ..$ intensities: num [1:2, 1:16, 1:3] 1955 1759 3718 1938 2864 ...
  ...
 $ 31337_at  :List of 1
  ..$ 31337_at:List of 1
  .. ..$ intensities: num [1:2, 1:16, 1:3] 6264 12014 2364 1700 2898 ...
 $ 31338_at  :List of 1
  ..$ 31338_at:List of 1
  .. ..$ intensities: num [1:2, 1:16, 1:3] 1014 963 1204 1085 2551 ...
```

First dimension is PM/MM, second the probe pairs, and last the arrays.

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

AFFX_CDF_PATH - keep CDF files in one place

Instead of copying the CDF files to the local directory, you can add them to the CDF search path, e.g.

```
options("AFFX_CDF_PATH"="some/path/to/cdfs/;another/path/metadata/")
```

Put this in `~/.Rprofile` so it is set when R starts. See `?Profile`.

Test to see what CDF files `affxparser` finds:

```
> findCdf(firstOnly=FALSE)
[2] "HG_U95A.CDF"
[3] "HG_U95Av2.CDF"
[6] "C:/Documents and Settings/hb/Affymetrix/cdf/Mapping50K_Hind240.CDF"
[7] "C:/Documents and Settings/hb/Affymetrix/cdf/Mapping50K_Xba240.CDF"
```

For more details, see `?findCdf`.

Note: Almost all `readCelNnn()` functions uses `findCdf()` internally.

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Fitting probeset models

Comparison between the multiplicative model of Li & Wong (2001) and the log-additive model of Irizarry et al. (2003):

```
> library(affy)
> files <- list.celfiles(pattern="^CL20010305.*")
> cel <- readCelUnits(files, units=81:89, stratifyBy="pm")
> eMult <- lapply(cel, function(ps) {
+   y <- t(ps[[1]]$intensities)
+   log(li.wong(y)$exprs, base=2)
+ })
> str(eMult)
List of 9
 $ 31320_at : num [1:16] 13.2 12.7 12.5 11.7 11.3 ...
 $ 31321_at : num [1:16] 10.9 12.0 11.4 10.0 10.0 ...
 ...
 $ 31328_at : num [1:16] 11.3 11.2 11.0 12.3 13.6 ...
> eLogAdd <- lapply(cel, function(ps) {
+   y <- ps[[1]]$intensities
+   fit <- medpolish(log(y, base=2), trace=FALSE)
+   fit$overall + fit$row
+ })
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Fitting probeset models

Comparison between the multiplicative model of Li & Wong (2001) and the log-additive model of Irizarry et al. (2003):

```
> library(affy)
> files <- list.celfiles(pattern="^CL20010305.*")
> cel <- readCelUnits(files, units=81:89, stratifyBy="pm")
> eMult <- lapply(cel, function(ps) {
+   y <- t(ps[[1]]$intensities)
+   log(li.wong(y)$exprs, base=2)
+ })
> str(eMult)
List of 9
 $ 31320_at : num [1:16] 13.2 12.7 12.5 11.7 11.3 ...
 $ 31321_at : num [1:16] 10.9 12.0 11.4 10.0 10.0 ...
 ...
 $ 31328_at : num [1:16] 11.3 11.2 11.0 12.3 13.6 ...
> eLogAdd <- lapply(cel, function(ps) {
+   y <- ps[[1]]$intensities
+   fit <- medpolish(log(y, base=2), trace=FALSE)
+   fit$overall + fit$row
+ })
> layout(matrix(1:9, nrow=3))
> for (kk in seq(eMult)) {
+   plot(eMult[[kk]], pch=0, col="blue", ylim=c(8,15), ylab="y")
+   points(eLogAdd[[kk]], pch=2, col="red")
+ }
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

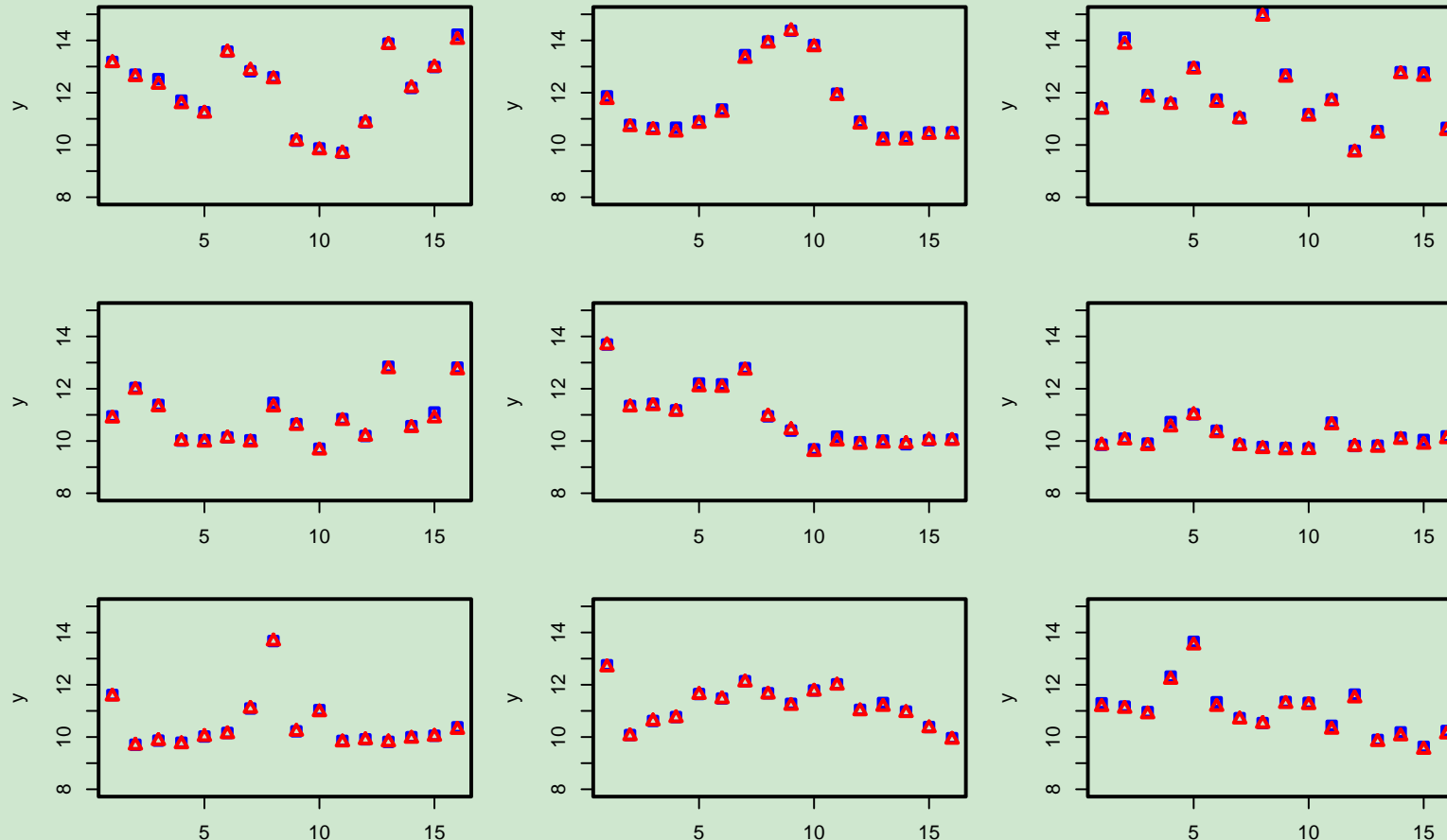
SNP chips

Summary

Fitting probeset models

Comparison between the multiplicative model of Li & Wong (2001) and the log-additive model of Irizarry et al. (2003):

Example



```
+ points(eLogAdd[[kk]], pch=2, col="red")  
+ }
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Using the CDF structure

For an 100K XbaI SNP sample, read probesets 56-60:

```
> cel <- readCelUnits("array1.CEL", units=56:60, stratifyBy="pm")
```

An alternative way (it will become clear later why):

```
> chiptype <- readCelHeader("array1.CEL")$chiptype
```

```
> chiptype
```

```
[1] "HG_U95Av2"
```

```
> cdfFile <- findCdf(chiptype)
```

```
> cdf <- readCdfUnits(cdfFile, units=56:60, stratifyBy="pm")
```

```
> cel2 <- readCelUnits("array1.CEL", cdf=cdf)
```

This reads the same data:

```
> identical(cel1, cel2)
```

```
[1] TRUE
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

**CDF
structure**

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Using the CDF structure (continued)

A CDF structure is a list of probesets, in turn lists of elements type, direction, and groups. The groups element is the interest one. For expression arrays there is only one group, whereas for SNP arrays there are four (allele A & allele B on both strands).

```
> cdf[["31323_r_at"]]$groups
$'31323_r_at'
$'31323_r_at'$x
 [1] 323 201 358 611 612 409 479 106 607  88  45  44 452 503 528 390
$'31323_r_at'$y
 [1] 545 113 581 237 237 447 215 291 425 463 497 497 315 467  43 471
$'31323_r_at'$pbase
 [1] "G" "T" "G" "T" "G" "G" "G" "G" "G" "T" "G" "T" "G" "A" "G" "A"
$'31323_r_at'$tbase
 [1] "C" "A" "C" "A" "C" "C" "C" "C" "C" "A" "C" "A" "C" "T" "C" "T"
$'31323_r_at'$expos
 [1] 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

When passing this `cdf` object to `readCelUnits()`, the latter uses the group fields `x` and `y` to identify the probe intensities:

```
> cel2 <- readCelUnits("array1.CEL", cdf=cdf)
```

Note: Only PMs will be read since we already stratified on PM when reading the CDF!

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Reading minimal CDF structure of probe indices

Internally probe (cell) indices are calculated from x and y in order to identify the signals in the CEL files. There is a function to get the probe indices directly;

```
> cdf <- readCdfCellIndices(cdfFile, units=80:84, stratifyBy="pm")
> cdf[["31323_r_at"]]$groups
$'31323_r_at'
$'31323_r_at'$indices
 [1] 349124 72522 372199 152292 152293 286490 138080 186347 272608
[10] 296409 318126 318125 202053 299384 28049 301831
```

This can be used to read CEL files (this is actually what is done internally):

```
> files <- c("array1.CEL", "array2.CEL")
> cel <- readCelUnits(files, cdf=cdf)
> cel[["31323_r_at"]]
$'31323_r_at'
$'31323_r_at'$intensities
      [,1] [,2]      [,1] [,2]
[1,] 4137.0 2741.0 [9,] 15113.5 10725.8
[2,] 1960.5 1465.0 [10,] 14543.3 8279.0
[3,] 1899.0 1562.0 [11,] 4384.0 2179.8
[4,] 2440.0 1107.8 [12,] 2191.0 1467.0
[5,] 2822.0 1769.5 [13,] 1269.0 1086.0
[6,] 3632.5 2456.0 [14,] 1450.0 930.3
[7,] 12030.0 8050.0 [15,] 1941.0 1338.0
[8,] 19525.0 9252.8 [16,] 1747.8 1261.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Restructure CDF structure and apply to CEL files

1. Rearrange CDF so that the values in each probeset is read into a 4x4 matrix:

```
> cdf2 <- applyCdfGroups(cdf, lapply, lapply, matrix, nrow=4)
> cdf2[["31323_r_at"]]$groups
$`31323_r_at`
$`31323_r_at`$indices
      [,1] [,2] [,3] [,4]
[1,] 349124 152293 272608 202053
[2,]  72522 286490 296409 299384
[3,] 372199 138080 318126  28049
[4,] 152292 186347 318125 301831
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

**Rearranging
CDF
structure**

SNP chips

Summary

Restructure CDF structure and apply to CEL files

1. Rearrange CDF so that the values in each probeset is read into a 4x4 matrix:

```
> cdf2 <- applyCdfGroups(cdf, lapply, lapply, matrix, nrow=4)
> cdf2[["31323_r_at"]]$groups
$`31323_r_at`
$`31323_r_at`$indices
      [,1] [,2] [,3] [,4]
[1,] 349124 152293 272608 202053
[2,]  72522 286490 296409 299384
[3,] 372199 138080 318126  28049
[4,] 152292 186347 318125 301831
```

2. Read the two CEL files using this structure:

```
> cel2 <- readCelUnits(files, cdf=cdf2)
> cel2[["31323_r_at"]]
$`31323_r_at`
$`31323_r_at`$intensities
, , 1
      [,1] [,2] [,3] [,4]
[1,] 4137.0 2822.0 15113.5 1269.0
[2,] 1960.5 3632.5 14543.3 1450.0
[3,] 1899.0 12030.0 4384.0 1941.0
[4,] 2440.0 19525.0 2191.0 1747.8
, , 2
      [,1] [,2] [,3] [,4]
[1,] 2741.0 1769.5 10725.8 1086.0
[2,] 1465.0 2456.0 8279.0 930.3
[3,] 1562.0 8050.0 2179.8 1338.0
[4,] 1107.8 9252.8 1467.0 1261.0
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset

0. Find all CEL files, read the CDF structure for SNPs 81-84:

```
> files <- list.celfiles(path="cel/Xba/", full.names=TRUE)
> cdfFile <- findCdf(".*Xba.*") # Find the CDF for the Xba chip
> cdf <- readCdfCellIndices(cdfFile, units=81:84, stratifyBy="pmmm")
> names(cdf)
[1] "SNP_A-1649924" "SNP_A-1699383" "SNP_A-1732654" "SNP_A-1702431"
```

1. Rearrange CDF indices into probe quartets (PM_A , MM_A , PM_B , MM_B):

```
> cdf2 <- applyCdfGroups(cdf, cdfMergeToQuartets)
> cdf2[["SNP_A-1699383"]]$groups
$forward
$forward$indices
      [,1]  [,2]  [,3]  [,4]  [,5]
pmA 1551095 1551468 805028 792499 1433269
mmA 1552695 1553068 806628 794099 1434869
pmB 1551096 1551467 1547229 792500 1433270
mmB 1552696 1553067 1548829 794100 1434870
$reverse
$reverse$indices
      [,1]  [,2]  [,3]  [,4]  [,5]
pmA 1448961 827956 1539037 1023370 1442668
mmA 1450561 829556 1540637 1024970 1444268
pmB 1448960 1583757 1539038 1023369 1442667
mmB 1450560 1585357 1540638 1024969 1444267
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

2. Using the CDF structure with quartets, read the CEL files:

```
> files <- list.celfiles(path="cel/Xba/", full.names=TRUE)
> cel <- readCelUnits(files, cdf=cdf2)
> names(cel)
[1] "SNP_A-1649924" "SNP_A-1699383" "SNP_A-1732654" "SNP_A-1702431"
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

2. Using the CDF structure with quartets, read the CEL files:

```
> files <- list.celfiles(path="cel/Xba/", full.names=TRUE)
> cel <- readCelUnits(files, cdf=cdf2)
> names(cel)
[1] "SNP_A-1649924" "SNP_A-1699383" "SNP_A-1732654" "SNP_A-1702431"

> y <- cel[["SNP_A-1699383"]]$forward$intensities
> str(y)
 num [1:4, 1:5, 1:90] 5272 1239 243 433 4184 ...
> y[, , 2]
, , 1
      [,1] [,2] [,3] [,4] [,5]
[1,] 5272 4184 4103 5036 4165
[2,] 1239 610 1182 1149 1227
[3,] 243 419 422 720 979
[4,] 433 427 498 731 933
> y[, , 90]
      [,1] [,2] [,3] [,4] [,5]
[1,] 2256 1870 1897 2661 1909
[2,] 560 495 358 624 894
[3,] 1644 1345 1412 1397 1344
[4,] 422 281 549 684 703
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

Task: Plot (PM_A , PM_B) for SNPs 56:59 across all 90 samples.

```
cdfFile <- findCdf(".*Xba.*") # Find the CDF for the Xba chip
cdf <- readCdfCellIndices(cdfFile, units=56:59, stratifyBy="pm")
cdf <- applyCdfGroups(cdf, cdfMergeStrands)
cdf <- applyCdfGroups(cdf, cdfMergeToQuartets)
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

Task: Plot (PM_A , PM_B) for SNPs 56:59 across all 90 samples.

```
cdfFile <- findCdf(".*Xba.*") # Find the CDF for the Xba chip
cdf <- readCdfCellIndices(cdfFile, units=56:59, stratifyBy="pm")
cdf <- applyCdfGroups(cdf, cdfMergeStrands)
cdf <- applyCdfGroups(cdf, cdfMergeToQuartets)

files <- list.celfiles(path="cel/Xba/", full.names=TRUE)
cel <- readCelUnits(files, cdf=cdf)
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

Task: Plot (PM_A , PM_B) for SNPs 56:59 across all 90 samples.

```
cdfFile <- findCdf(".*Xba.*") # Find the CDF for the Xba chip
cdf <- readCdfCellIndices(cdfFile, units=56:59, stratifyBy="pm")
cdf <- applyCdfGroups(cdf, cdfMergeStrands)
cdf <- applyCdfGroups(cdf, cdfMergeToQuartets)

files <- list.celfiles(path="cel/Xba/", full.names=TRUE)
cel <- readCelUnits(files, cdf=cdf)

layout(matrix(1:4, nrow=2))
par(mar=c(4,4,1,1)+0.2, pch=176)
lim <- c(7,16)
xlab <- expression(log[2](y[A]))
ylab <- expression(log[2](y[B]))
for (kk in 1:4) {
  y <- cel[[kk]][[1]]$intensities
  yA <- y[1,,]
  yB <- y[2,,]
  plot(log(yA,2), log(yB,2), xlab=xlab, ylab=ylab, xlim=lim, ylim=lim)
  abline(a=0, b=1, lty=2)
  legend("topleft", legend=names(cel)[kk], bty="n")
}
```

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

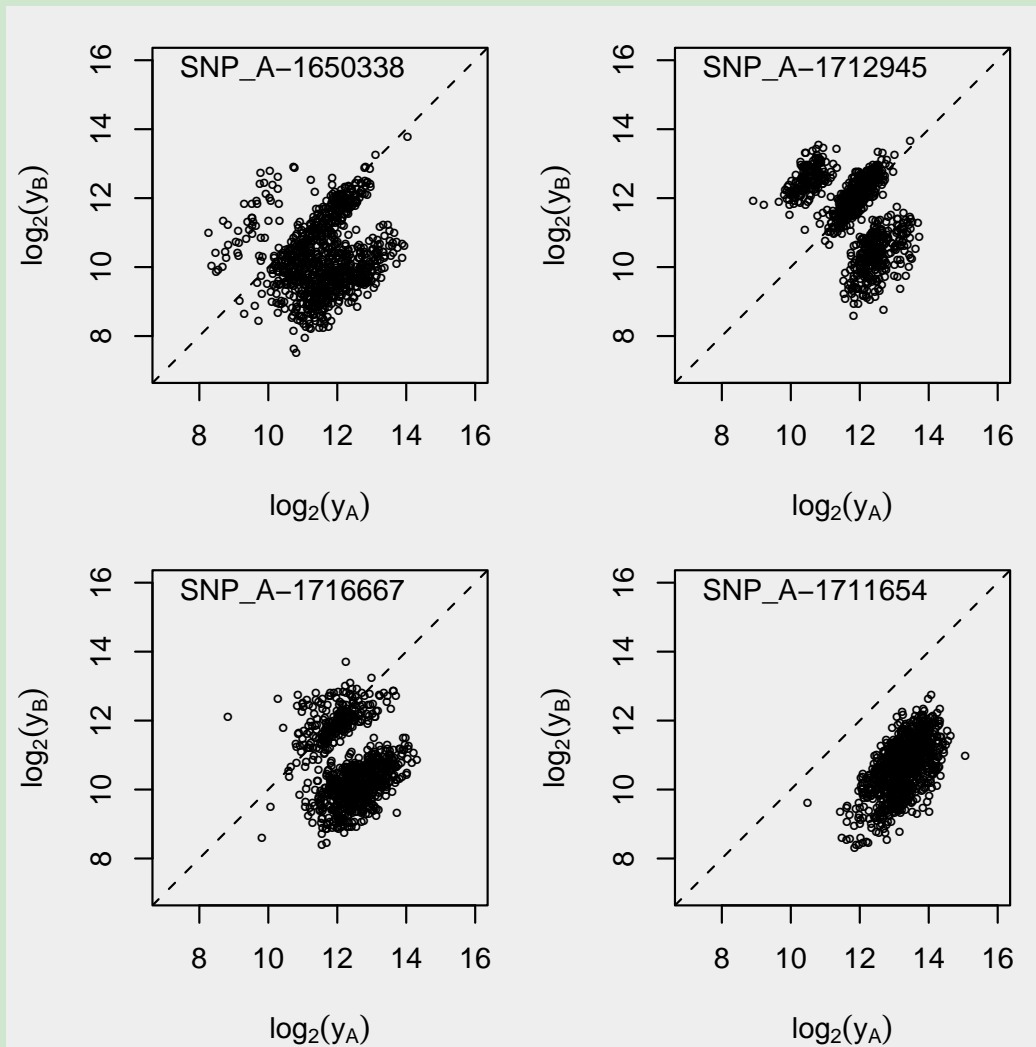
Rearranging
CDF
structure

SNP chips

Summary

Example using the 90 CEPH 100K SNP chip dataset (continued)

Example



Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary

- ▶ Use it!
- ▶ Work with binary rather than ASCII files; much faster!
Affymetrix' CEL File Conversion Tool:
<http://www.affymetrix.com/support/developer/tools/devnettools.affx>
(Fon't know about CDF converters)
- ▶ Learn how to restructure CDF files.
- ▶ Read the help.
- ▶ Look at the examples.

Introduction
to the
affxparser
package

H Bengtsson
J Bullard
K Hansen

Introduction
Overview

CEL files

Reading one
array

Reading
multiple
arrays

Reading
multiple
probesets on
multiple
arrays

Settings

Probeset
models

CDF file

CDF
structure

Reading
CEL files
using CDF

Rearranging
CDF
structure

SNP chips

Summary