Preprocessing gene expression dataset for **VANGUARD** prospective study by `aroma.affymetrix`

Li Shen, Jing Wang, and Kevin R. Coombes
Dept. of Bioinformatics and Computational Biology
The University of Texas M.D. Anderson Cancer Center (MDACC)

January 28, 2011

Contents

1 Executive Summary

1.1 Introduction

This dataset was acquired using Affymetrix HuGene 1.0 ST arrays.

This is the first part of a series of related reports. In this report, we process the CEL files and perform QC using `aroma.affymetrix` package in *R*.

1.1.1 Objectives

The scientfic goals of this study are listed as below:

1. To identify common and distinct molecular profiles among the lung airways with respect to original location of resected tumors. Adjacent versus non-adjacent (same lung or contralateral) vs main carinas. Site-dependent modulation of the field of cancerization.
2. To identify genes that are modulated in a time-dependent fashion in the lung airways. Temporal modulation of the field of cancerization.
3. To correlate gene signatures signifying site and temporal modulation of the lung airway field of cancerization with patterns of recurrence or second primary tumor development.

1.2 Methods

1.2.1 Data Description

Patients with specimens available for analysis in a continuous time series up to either 24 months or 36 months and at least four specimens from four different sites in the lung airways at each time point. The patients are primary lung cancer only. A total of 391 airway samples were profiled from 19 patient cases (15 adenocarcinomas and 4 squamous cell carcinomas).

Please note that 71 samples were processed between 2007 and 2008 and 320 samples were processed during 11/2010 to 01/2011.

1.2.2 Statistical Methods

The datasets were handle using `aroma.affymetrix` package in R for initial QC. The .CEL files were quantified using the RMA background correction, quantile normalization, and RmaPlm (Probe-Level Models using RMA) summarization methods using original scale. After processing the data, we applied log base 2 transformation. All processes are conducted by aroma.affymetrix package in R.

For basic quality controls, we employed the plotting functions `mva.pairs`, `density` and `boxplot`. The function `boxplot` presents side-by-side graphical summaries of intensity information from each array and the function `mva.pairs` produces Bland-Altman (M-versus-A) pairwise plots which offer pairwise graphical comparison of intensities from randomly selected arrays.

1.3 Results

Figures 1 to 3 suggested the need for normalization of data. After data were background corrected, normalized and summarized, same figures were plotted (Figures 4 and 6).

We are aware that there is batch effect between the first (71 samples) and second (320 samples) from Figures 7 and 8.

List of Tables

List of Figures

2 Loading the Data

2.1 Setting working directory

We first set up directory we use for the analysis.

```
> getwd()
```

```
[1] "/data/bioinfo2/Lung-HN/Wistuba-VANGUARD/Analysis"
```

2.2 R Libraries

We begin by loading all the libraries we will need for this analysis. A list of the current versions of the libraries used for the analysis can be found in the appendix.

```
> library(affy)
> library(simpleaffy)
> library(geneplotter)
> library(xtable)
> library(ClassComparison)
> library(ClassDiscovery)
> library(PreProcess)
> library(aroma.affymetrix)  # aroma.affymetrix
> verbose <- Arguments$getVerbose(-8, timestamp=TRUE)
> library(preprocessCore)
> library(RColorBrewer)      # colorbrewer
```

2.3 Sample Information

In order to perform an analysis of microarray data, we need to know something about the samples that were hybridized to the microarrays. In standard R usage, this sort of "sample" information is typically stored in a data frame.

In this study, a total of 391 Affymetrix HuGene 1.0 ST arrays were employed. In order to create a shorter name to represent each sample, we combine the **Case.ID**, **Site of collection** and **Time.point** columns in the Sample Info file.

```
> si.dir <- "..//"
> si <- read.delim(file.path(si.dir, "VANGUARD 2009-2011 annotated.txt"),
+                  sep="\t", header=TRUE, as.is=TRUE)
> ### short name ###
> sn <- paste(si$"V_Case.ID.Inclusion_number.",
+             si$"site.of.collection",
+             si$"Time.point", sep="-")
> rownames(si) <- sn
> rm(sn)
```

Patients with specimens available for analysis in a continuous time series up to either 24 months or 36 months and at least four specimens from four different sites in the lung airways at each time point. The patients are primary lung cancer only. A total of 391 airway samples were profiled from 19 patient cases (15 adenocarcinomas and 4 squamous cell carcinomas).

```
> dim(si)
```

```
[1] 391  61
```

```
> colnames(si)
```

```
 [1] "Experiment.Names"
 [2] "Batch"
 [3] "Sample.Name..Lab."
 [4] "Prep.ID"
 [5] "CORE_ID"
 [6] "Row.ID"
 [7] "Sample.ID"
 [8] "V_Case.ID.Inclusion_number."
 [9] "Selected.Cases"
[10] "mRNA.Profiling.needed"
[11] "Evaluable..sample.at.B.12M._Reason"
[12] "Off.study_Reason"
[13] "MRN..MDAH."
[14] "Gender"
[15] "DOB..DOBirth."
[16] "DOSurgery"
[17] "DOInclusion"
[18] "Surgery"
[19] "Diagnosis..Histology."
[20] "Differentiation"
[21] "Leison.Site"
[22] "Anatomical_site"
[23] "site.of.collection"
[24] "Contralateral"
[25] "ContralateralADENO"
[26] "ContralateralSCC"
[27] "DetMap"
[28] "DetMapADENO"
[29] "DetMapSCC"
[30] "Timecarina"
[31] "Map"
[32] "MapAdeno"
```

```
[33] "MapSCC"
[34] "Time.point"
[35] "Code.4.time.point"
[36] "Time.code..annotated."
[37] "Lab."
[38] "BronchialBrush.No"
[39] "Code.4.Site.of.collection"
[40] "Site.code..annotated."
[41] "site.comment"
[42] "Accession..Primary.tumor_Pathology.number."
[43] "pT"
[44] "pN"
[45] "Final.Pat.Stage"
[46] "EGFR.status"
[47] "KRAS.status"
[48] "Note"
[49] "order"
[50] "CEL"
[51] "CEL.done"
[52] "T_Box..80C_RNA."
[53] "T_Row.80C_RNA."
[54] "T_Column..80C_RNA."
[55] "Cell.lysate.for.RNA.extraction"
[56] "T_Comment.cell.lysate.for.RNA.extraction."
[57] "R_Box..80C."
[58] "R_Row..80C."
[59] "R_Column..80C."
[60] "RNA..available."
[61] "Collected.Month.Sample."

> ### 19 patient cases (15 adenocarcinomas and 4 squamous cell carcinomas) ###
> with(si, table(Case.ID=V_Case.ID.Inclusion_number.,
+                Diagnosis=Diagnosis..Histology.))

       Diagnosis
Case.ID Adenocarcinoma Squamous
     1              25        0
     3              23        0
     6              24        0
     10              0       18
     16             15        0
     18             23        0
     20             23        0
     23              0       17
     30             23        0
     31             24        0
     35             17        0
     38              0       23
     40              0       24
     41             18        0
     44             24        0
     46             17        0
     47             18        0
     48             17        0
     50             18        0
```

```
> with(si, table(Batch))
```

```
Batch
  I  II
 71 320
```

```
> with(si, table(Time.point))
```

```
Time.point
  0  12  24  36
109 108 113  61
```

```
> with(si, table(site.of.collection))
```

```
site.of.collection
    Carina-dup           LB10            LB6            LB9 Left main stem
             1             61              1              1              1
           LLL            LUL      LUL-stump      Main stem             MC
             2             65              1              1             59
   Mid trachea           PLMS           RB10            RLL            RML
             1              1             65              1             66
           RUL
            64
```

```
> with(si, table(Batch, Time.point))
```

```
     Time.point
Batch   0  12  24  36
   I   53  10   6   2
   II  56  98 107  59
```

```
> with(si, table(Batch, site.of.collection))

     site.of.collection
Batch Carina-dup LB10 LB6 LB9 Left main stem LLL LUL LUL-stump Main stem MC Mid trachea
   I           0    4   0   1              0   0  11         0         0 34           0
   II          1   57   1   0              1   2  54         1         1 25           1
     site.of.collection
Batch PLMS RB10 RLL RML RUL
   I     0    4   0  11   6
   II    1   61   1  55  58

> with(si, table(Map))

Map
            ADJ      MC NON-ADJ
      1      62      60     268

> with(si, table(DetMap))

DetMap
            ADJ  CONTRA      MC NON-ADJ
      1      62     161      60     107
```

It is useful to display the sample information in a table inside a report. We then use the `xtable` command from the `xtable` package to generate a seperate HTML table (Please see attached).

3 Data Preprocessing

3.1 Setting up annotation files

For each chip type there is a unique chip definition file (CDF). A CDF contains information on which probes belong to what probeset, the (x,y) location of each probe, which the middle nucleotides in the target and the probe are (from which PM/MM status is inferred), and so on.

Aroma.affymetrix searches for CDF files in the annotationData/chipTypes directory of the current working directory. We first place the CDF in:
//annotationData/chipTypes/HuGene-1_0-st-v1

```
> chipType <- "HuGene-1_0-st-v1"
> cdf <- AffymetrixCdfFile$byChipType(chipType, tags = "r3")
> print(cdf)
```

```
AffymetrixCdfFile:
Path: annotationData/chipTypes/HuGene-1_0-st-v1
Filename: HuGene-1_0-st-v1,r3.cdf
Filesize: 16.67MB
Chip type: HuGene-1_0-st-v1,r3
RAM: 0.00MB
File format: v4 (binary; XDA)
Dimension: 1050x1050
Number of cells: 1102500
Number of units: 33252
Cells per unit: 33.16
Number of QC units: 0
```

3.2 Defining CEL set

Now we can actually read the CEL files, using the AffymetrixCelSet command. All the CEL files are stored under:
//bioinfo2/Lung-HN/Wistuba-VANGUARD/CEL 2009 and 2011

```
> cs <- AffymetrixCelSet$byName("IW-VANGUARD", cdf = cdf)
> ### Extract data as a matrix for a set of arrays
> ab <- extractMatrix(cs)   # not run
```

```
> all(colnames(ab)==si$Experiment.Names)
```

```
[1] FALSE
```

```
> colnames(ab) <- rownames(si)
> print(cs)
```

```
AffymetrixCelSet:
Name: IW-VANGUARD
Tags:
Path: rawData/IW-VANGUARD/HuGene-1_0-st-v1
Platform: Affymetrix
Chip type: HuGene-1_0-st-v1,r3
Number of arrays: 391
Names: 1188_IW_001_12-10-10_(HuGene-1_0-st-v1), 1188_IW_002_12-10-10_(HuGene-1_0-st-v1), ..., I
Time period: 2008-07-22 11:57:31 -- 2011-01-14 13:53:37
Total file size: 4128.77MB
RAM: 0.51MB
```

```
> dim(ab)
```

```
[1] 1102500     391
```

3.3 Quality assessment of raw data

The aroma.affymetrix package also contains some plotting routines that help us decide whether normalization is needed and, if so, whether it is behaving sensibly. We start with Bland-Altman (M-versus-A) pairwise plots of some randomly selected arrays (Figure 1). Using the smoothScatter method is highly recommended, since it provides much more efficient plotting routines. The need for normalization can also be assessed using density plots (Figure 2) and boxplots (Figure 3).

Additional assessments of the data will use colors to distinguish the different arrays. We prepare a standard vector of color assignments here, stored in an object that will be used by those plotting routines.

```
> colorSet.batch <- c("red", "gray")
> col.batch <- colorSet.batch[as.factor(si$Batch)]
```

Figures 1 and 3 suggested the need for normalization of data.

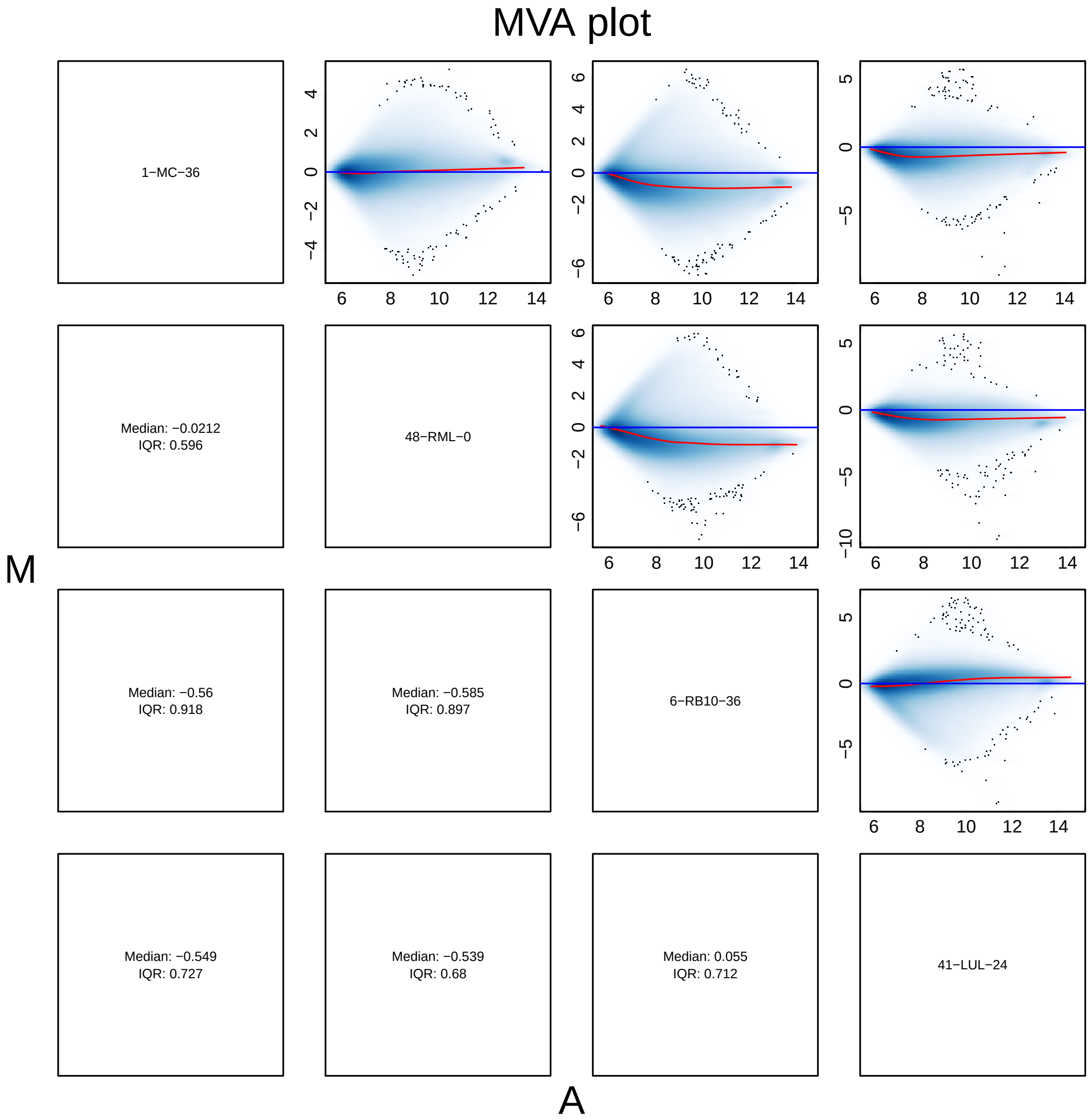

Figure 1: Pairwise Bland-Altman (M-vs-A) plots of the probe-level intensity data for four randomly selected arrays.

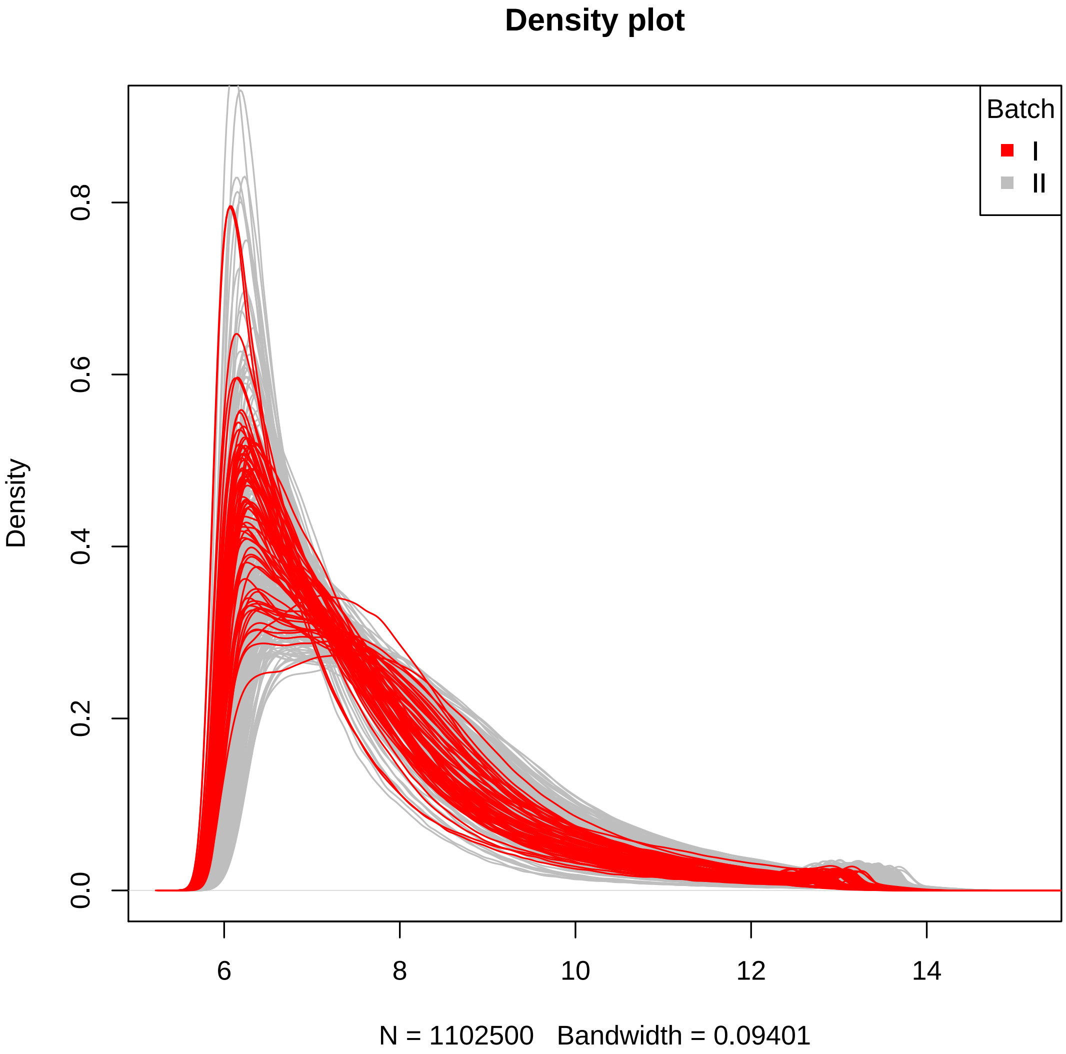

Figure 2: Density plots of the probe-level log intensity data in all arrays.

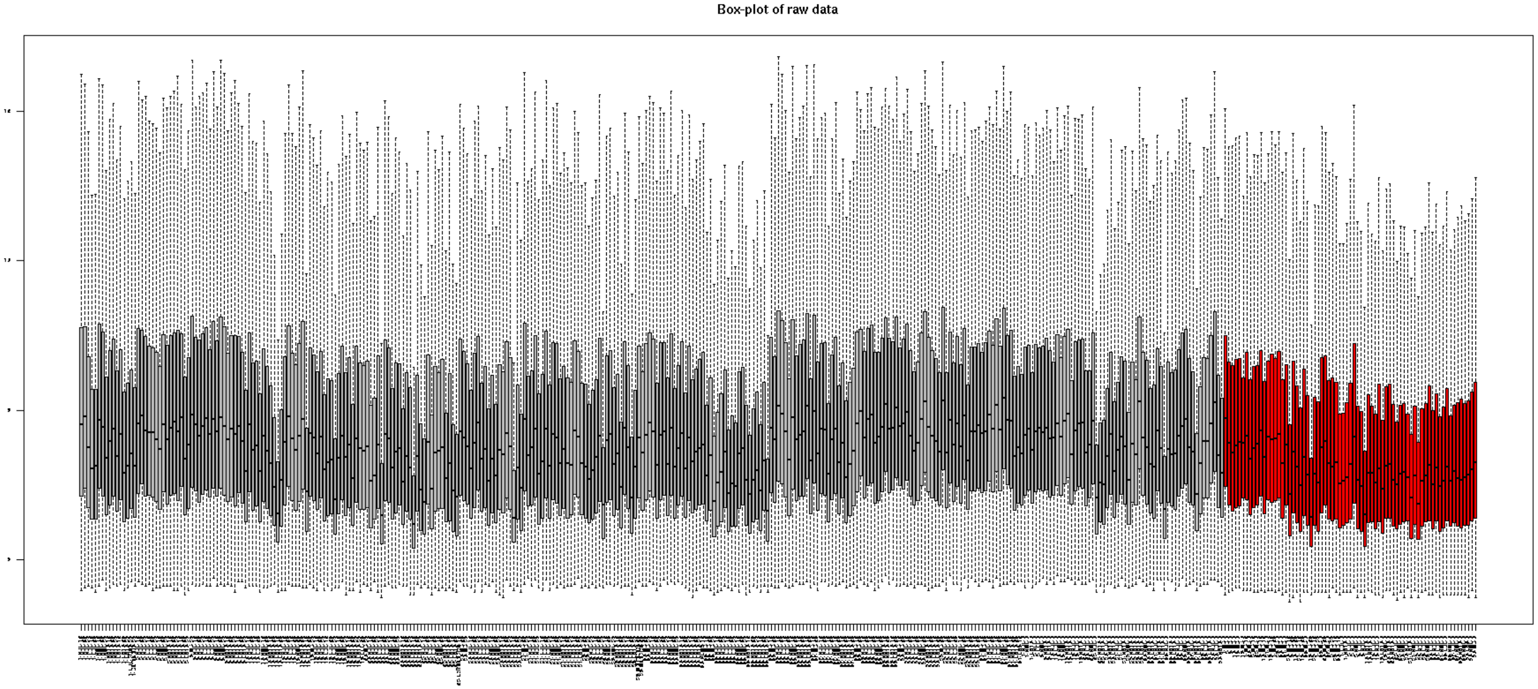

Figure 3: Box plots of the probe-level log intensity data in all arrays.

4 Quantification

The .CEL files were quantified using the RMA background correction, quantile normalization and RmaPlm summarization method. Then, we applied log base 2 transformation. All are processed by `aroma.affymetrix` package in R.

4.1 RMA Background Correction

We use the `RmaBackgroundCorrection` function applies the Robust Multiarray Analysis (RMA) algorithm to quantify the Affymetrix data. RMA background correction estimates the background by a mixture model where the background signals are assumed to be normally distributed and the true signals are exponentially distributed. This algorithm borrows strength across arrays (Note: This takes approximately 60 seconds per array).

```
> bc <- RmaBackgroundCorrection(cs)
> csBC <- process(bc, verbose=verbose)
```

```
> print(bc)
```

```
RmaBackgroundCorrection:
Data set: IW-VANGUARD
Input tags:
User tags: *
Asterisk ('*') tags: RBC
Output tags: RBC
Number of files: 391 (4128.77MB)
Platform: Affymetrix
Chip type: HuGene-1_0-st-v1,r3
Algorithm parameters: (subsetToUpdate: NULL, typesToUpdate: chr "pm", addJitter: logi FALSE, ji
Output path: probeData/IW-VANGUARD,RBC/HuGene-1_0-st-v1
Is done: TRUE
RAM: 0.00MB
```

4.2 Quantile Normalization

The normalization method is implemented as a two-pass procedure. First the target distribution is estimated by averaging the (ordered) signals over all arrays, then each array is normalized toward this target distribution. The implementation is such that data from at most two arrays are kept in memory at any time.

```
> qn <- QuantileNormalization(csBC, typesToUpdate ="pm")
> csN <- process(qn, verbose=verbose)
```

```
> print(qn)
```

```
QuantileNormalization:
Data set: IW-VANGUARD
Input tags: RBC
User tags: *
Asterisk ('*') tags: QN
Output tags: RBC,QN
Number of files: 391 (4113.54MB)
Platform: Affymetrix
Chip type: HuGene-1_0-st-v1,r3
Algorithm parameters: (subsetToUpdate: NULL, typesToUpdate: chr "pm", subsetToAvg: NULL, typesT
Output path: probeData/IW-VANGUARD,RBC,QN/HuGene-1_0-st-v1
Is done: TRUE
RAM: 6.02MB
```

4.3 Summarization of probe-level data

Probe-level models (PLMs) are models that describe the (observed or pre-processed) probe signals using statistical models consisting of effects and random noise. The PLM used in RMA is a log-additive model.

```
> plmTr <- RmaPlm(csN)
> ### To fit the PLM for all units (probe sets) ###
> fit(plmTr, verbose=verbose)
> ### Quality assessment of PLM fit ###
> qam <- QualityAssessmentModel(plmTr)
> png(file=file.path("Figures", paste(my.fig,"Nuse.png", sep="")),
+     width=1800, height=800)
> plotNuse(qam, col=col.batch[match(colnames(ab),
+                                    rownames(si))])
> dev.off()
> png(file=file.path("Figures", paste(my.fig,"RLE.png", sep="")),
+     width=1800, height=800)
> plotRle(qam)
> dev.off()
> ### Extract Normalized Data (Probe-summarized data) ###
> cesTr <- getChipEffectSet(plmTr)
> gExpr <- extractDataFrame(cesTr, units = NULL, addNames = TRUE)
> ### log2 transformed ###
> normData <- log2(gExpr[, 6:ncol(gExpr)])
> rownames(normData) <- gExpr$unitName
> colnames(normData) <- rownames(si)[match(colnames(normData), si$Experiment.Names)]
> save(si, gExpr, normData, file="gExpr-RMA-Aroma.RData")
```

```
> print(plmTr)
```

```
RmaPlm:
Data set: IW-VANGUARD
Chip type: HuGene-1_0-st-v1,r3
Input tags: RBC,QN
Output tags: RBC,QN,RMA
Parameters: (probeModel: chr "pm"; shift: num 0; flavor: chr "affyPLM"; treatNAsAs: chr "weight
Path: plmData/IW-VANGUARD,RBC,QN,RMA/HuGene-1_0-st-v1
RAM: 0.01MB
```

```
> dim(gExpr)            ### processed but not tranformed
```

```
[1] 33252   396
```

```
> colnames(gExpr)[1:7] ### first 5 columns are annotations
```

```
[1] "unitName"                                "groupName"
[3] "unit"                                    "group"
[5] "cell"                                    "1188_IW_001_12-10-10_(HuGene-1_0-st-v1)"
[7] "1188_IW_002_12-10-10_(HuGene-1_0-st-v1)"
```

```
> dim(normData)         ### processed and log2 transformed
```

```
[1] 33252   391
```

4.4 Quality assessment of summarized data

The overall quality of samples is fairly decent. However, in the density and box plot (Figures 5 and 6), we also can notice that densities of quantile normalized intensities are not identical.

4.5 RLE and NUSE

The RLE (Relative Log Expression) and NUSE (Normalized Unscaled Standard Error) plots are useful and sensitive measures to assess array quality. Both are derived from a probe-level model (PLM) that computes an expression measure using M-estimator robust regression.

RLE plots (Figure 7) are constructed using log-scale estimates for the expression of each probe set on each array. For each probe set and each array, ratios are calculated between the expression of a probe set and the median expression of this probe set across all arrays of the experiment. For each array, these relative expression values are displayed as a box plot. Since it is assumed that in most experiments only relatively few genes are differentially expressed, the boxes should be similar in range and be centered close to 0.

NUSE (Figure 8) represents normalized standard error (SE) estimates from the PLM fit. The SE estimates are normalized such that for each probe set, the median standard error across all arrays is equal to 1. A box plot of NUSE values is drawn for each array. On the NUSE plot, arrays with lower quality will have boxes that are centered higher and/or have a larger spread than the other good quality arrays from the same experiment. Typically, boxes centered above 1.1 represent arrays that have quality problems

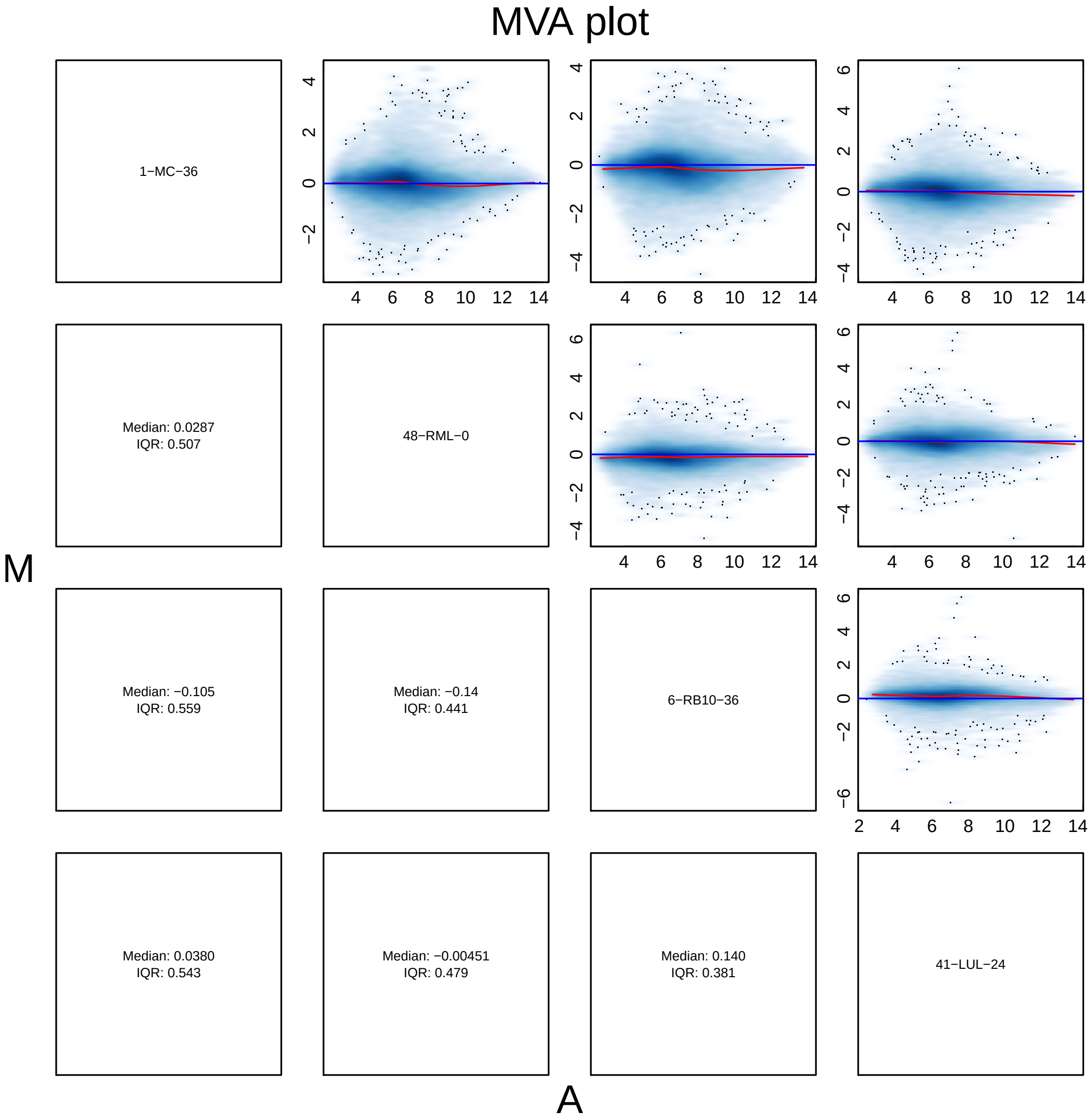

Figure 4: Pairwise Bland-Altman (M-vs-A) plots of the probe-level intensity data for randomly selected 4 arrays after processing data (background correction, normalization and summarization).

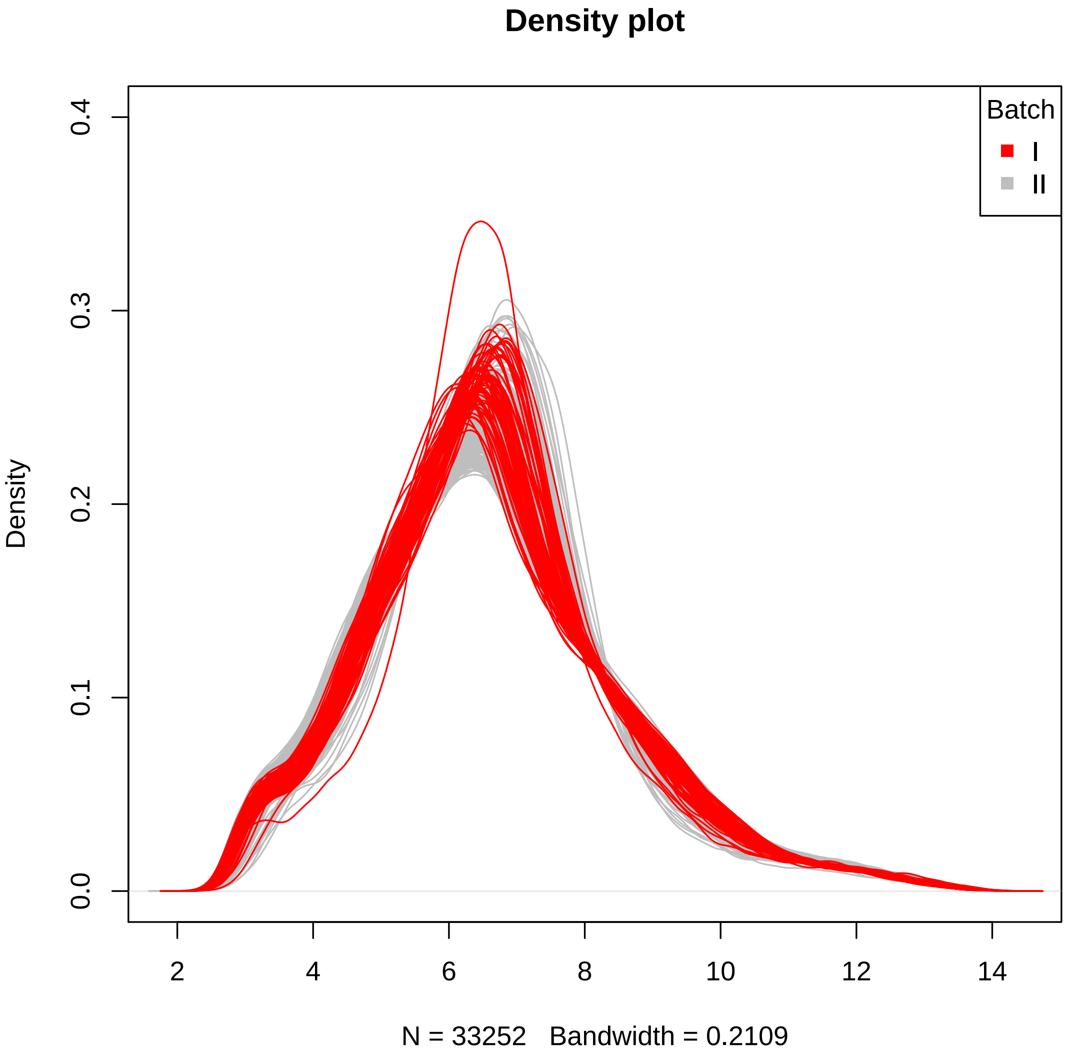

Figure 5: Density plots of the probe-level log intensity data in all arrays after processing data (background correction, normalization and summarization).

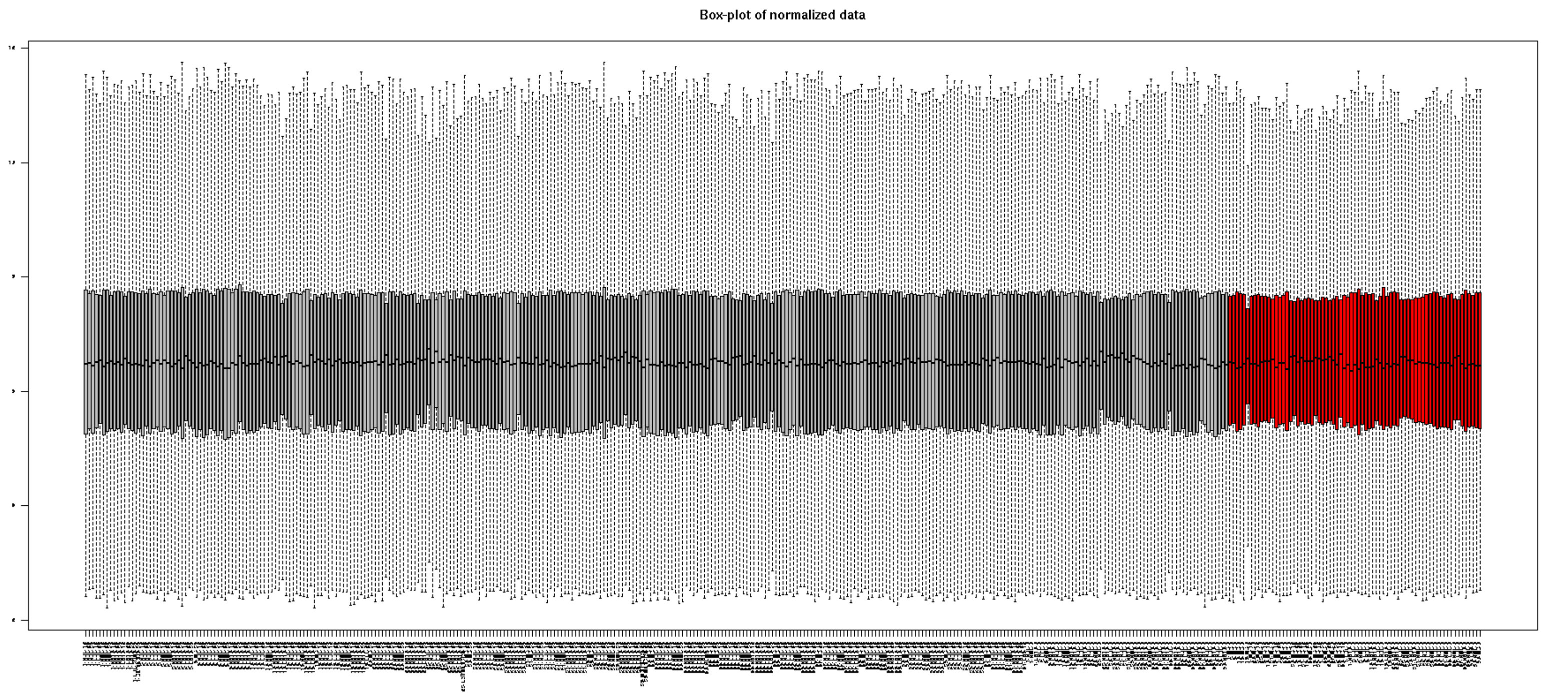

Figure 6: Box plots of the probe-level log intensity data in all arrays after processing data (background correction, normalization and summarization).

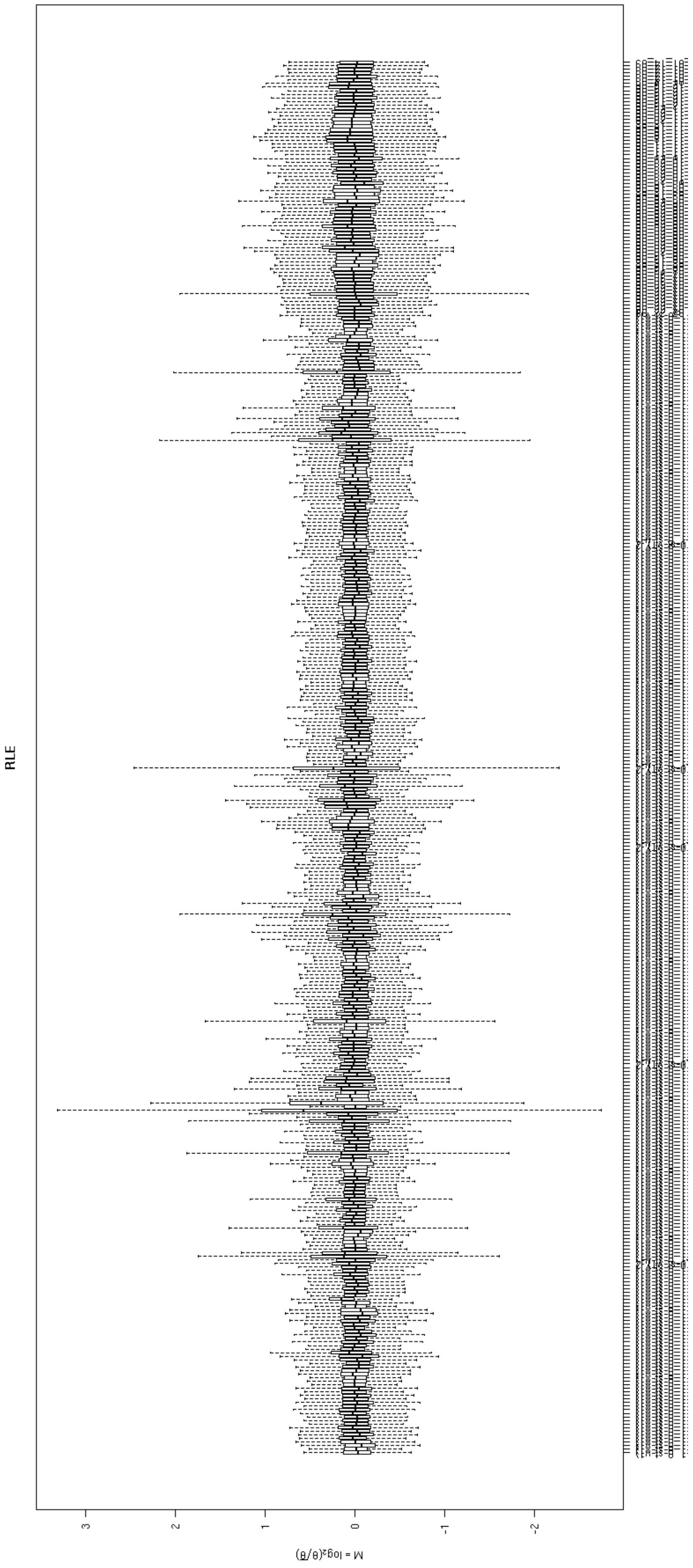

Figure 7: Box plots of RLE (Relative Log Expression)

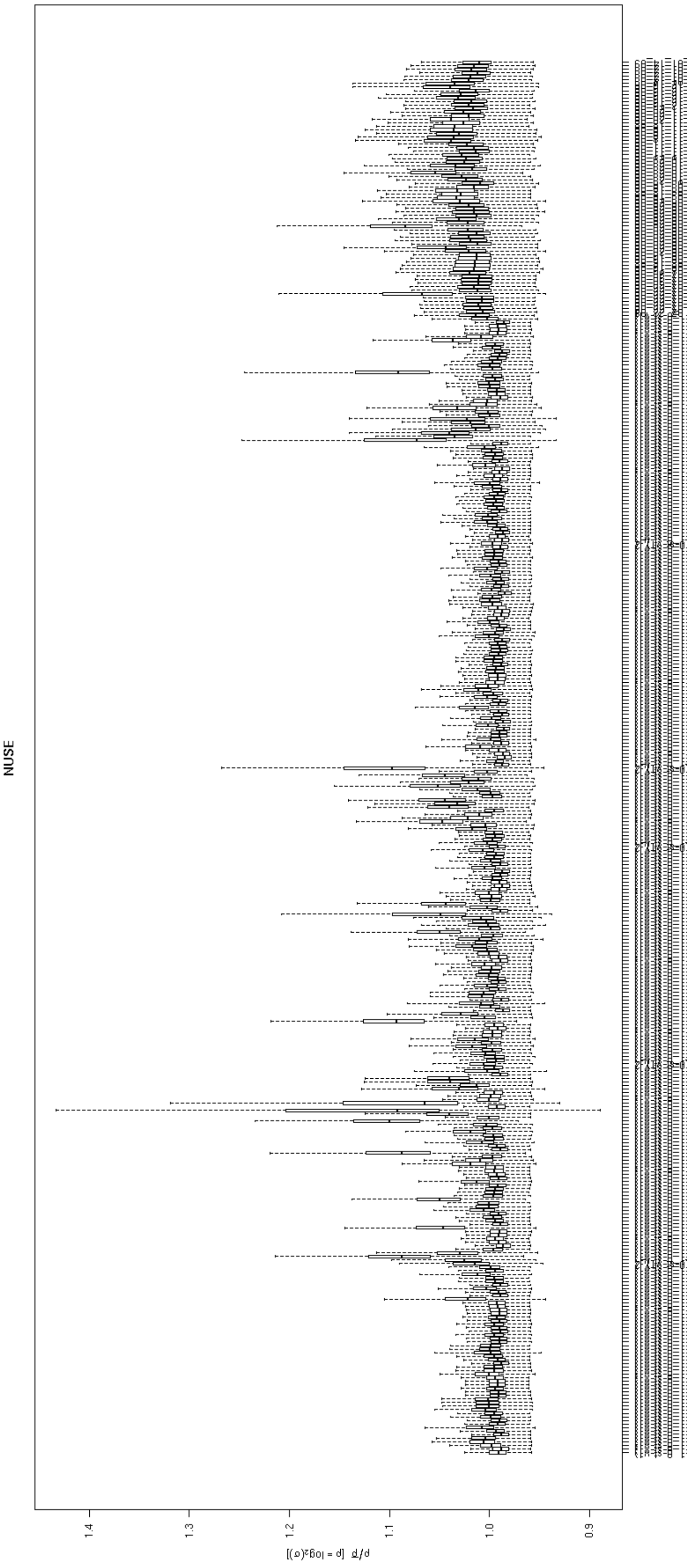

Figure 8: Box plots of NUSE (Normalized Unscaled Standard Error)

Appendices

A References

References

B Saving the Processed Data

The results of this analysis are saved in the following file.

```
> save.image("01_VANGUARD_HuGene1_0-st-Preprocess.RData")
```

C File Location

This analysis was run in the following directory:

```
> getwd()
```

```
[1] "/data/bioinfo2/Lung-HN/Wistuba-VANGUARD/Analysis"
```

D SessionInfo

This analysis was run in the following software environment:

```
> sessionInfo()
```

```
R version 2.12.0 (2010-10-15)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C                 LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=C                LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                    LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] splines   stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] RColorBrewer_1.0-2     preprocessCore_1.12.0  aroma.affymetrix_1.9.0
 [4] aroma.apd_0.1.7        affxparser_1.22.1      R.huge_0.2.0
 [7] aroma.core_1.9.1       aroma.light_1.18.2     matrixStats_0.2.2
[10] R.rsp_0.4.1            R.cache_0.3.0          R.filesets_0.9.1
```

```
[13] digest_0.4.2           R.utils_1.6.0          R.oo_1.7.4
[16] R.methodsS3_1.2.1      ClassDiscovery_2.10.1  mclust_3.4.7
[19] cluster_1.13.1         ClassComparison_2.10.1 PreProcess_2.10.0
[22] oompaBase_2.10.1       xtable_1.5-6           geneplotter_1.28.0
[25] lattice_0.19-13        annotate_1.28.0        AnnotationDbi_1.12.0
[28] simpleaffy_2.26.0      gcrma_2.22.0           genefilter_1.32.0
[31] affy_1.28.0            Biobase_2.10.0

loaded via a namespace (and not attached):
[1] affyio_1.18.0     Biostrings_2.18.2 DBI_0.2-5         grid_2.12.0
[5] IRanges_1.8.5     KernSmooth_2.23-4 RSQLite_0.9-4     survival_2.35-8
[9] tools_2.12.0
```