

Projektmanagement im Softwarebereich (Softwarepraktikum) 1. Woche

Thema:
Einführung in die MicroArray-Analyse mittels R

Gruppe 2:
Michael Krivan
Silver Wolf
Yannic Lapawczyk

Gliederung

1. Einleitung: Bedeutung von MicroArrays in der Bioinformatik
2. Was ist ein Affymetrix MicroArray Chip?
 - 2.1 Affymetrix, Inc.
 - 2.2 Aufbau des Chips
 - 2.3 Begriffserklärungen: PM, MM, Probe, Probe Set, Probe Pair
3. Implementierung von MicroArray-Daten in R
 - 3.1 .CEL-Datei
 - 3.2 Normalisierung
4. Qualitätsanalyse von MicroArray-Daten in R
 - 4.1 Chip image
 - 4.2 Histogramme
 - 4.3 RNA-Degradation-Plot
 - 4.4 Quality-Control Plot
5. Quellen

1. Einleitung: Bedeutung von MicroArrays in der Bioinformatik

Frage: Wieviele Gene besitzt der Mensch?

Antwort: 20.000 bis 22.000

Frage: Wie können wir die Expression all dieser Gene überprüfen?

Antwort: MicroArrays!

Frage: Wieviele Gene sind auf einem MicroArray enthalten?
(Unseren gegebenen Microarrays?)

Antwort:

- Mind. 20.000
- Spezifisch: 54.675
- Genanzahl != Probeanzahl
(mehrere Probes für ein Gen)



2. Was ist ein Affymetrix MicroArray Chip?

2.1 Affymetrix, Inc.

- Amerikanische Firma (aus Santa Clara, California)
- Spezialisierung auf Herstellung von DNA-MicroArrays
- Wurde von Dr. Stephen Fodor 1992 gegründet
- Dr. Fodors Forschungsgruppe hatte in den späten 80ern erste DNA-MicroArrays ("GeneChips") entwickelt
- Erster Affymetrix GenChip war HIV-markierter GeneChip
- GeneChips sollen schnelles Scannen einer Probe auf ein bestimmtes Gen hin erlauben (durch Erkennung von mRNA-Teilen)
- Einzelner Chip kann mehrere Tausend Gene erkennen, kann jedoch nur einmalig verwendet werden
- Weitere MicroArray-Hersteller: Illumina, GE Healthcare, Applied Biosystems, Beckman Coulter, Eppendorf Biochip Systems, Agilent

2. Was ist ein Affymetrix MicroArray Chip?

2.2 Aufbau des Chips

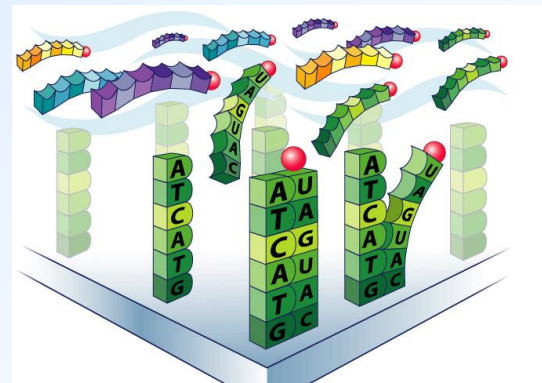
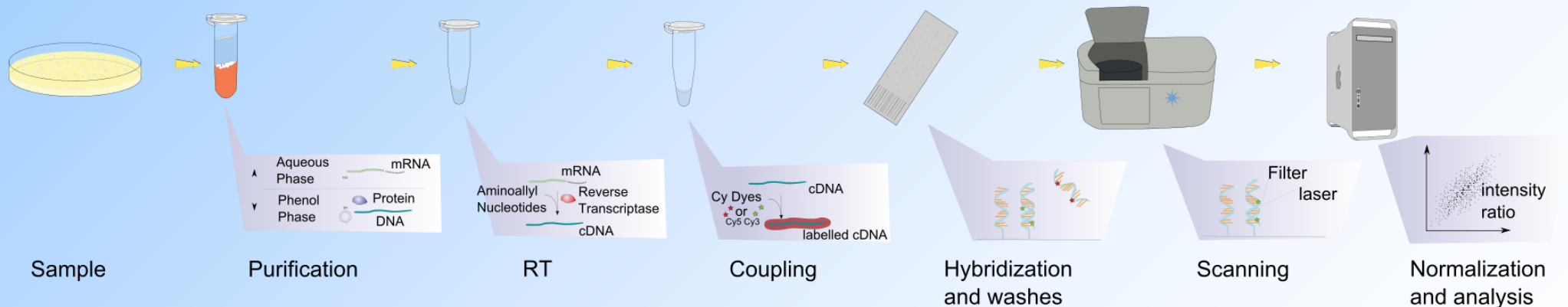
Unser spezifischer Typ: hgu133plus2

-> HG = Human Genome

-> DNA-MicroArray



Frage: Wie ist ein DNA-MicroArray aufgebaut?



2. Was ist ein Affymetrix MicroArray Chip?

2.3 Begriffserklärungen

Probe: "Oligonucleotides of 25 base pair length used to probe RNA targets."*

PM: "Probes intended to match perfectly the target sequence."*

- Bindet biologisch relevante transkribierte Information
- Intensität der Färbung -> Menge der gebundenen Transkripte quantifizieren

MM: "The probes having one base mismatch with the target sequence intended to account for non-specific binding."*

- Unspezifische Bindung
- Hintergrundrauschen wird quantifiziert -> Korrektur der Messungen am PM
- Dadurch können "falsche Bindungen abgezogen werden"

-> Unser spezifischer Typ: 11 PMs und 11 MMs

-> 1 Wert für affyID besteht aus diesen 22 Einzelwerten

Probe Set: "PMs and MMs related to a common affyID (an identification for a probe set which can be a gene or a fraction of a gene represented on the array.)"*

Probe Pair: "A unit composed of a perfect match and its mismatch."*

-> Frage: Wie mittels R Einzelwerte einsehen?

Dokumentation d. library "affy"

* Definitionen aus R

3. Implementierung von MicroArrays in R

3.1 .CEL-Datei

Frage: Was ist eine .CEL-Datei?

Antwort: Enthält Daten eines MicroArrays/Experimentes mit jeder Probe und der dazugehörigen Intensität der Hybridisierung

Frage: Wie können mit R .CEL-Dateien eingelesen werden?

Antwort:

- package affy: Funktionen für Affymetrix-Auswertungen

- Typspezifische cdf-Datei: Informationen was/wo auf Chip gespeichert ist

- Daten RMA-Normalisieren -> Dadurch: Entstehung der affyID (z.B. 1007_s_at)

- Zu jeder affy-ID das entsprechende Gen zuordnen (Gen Pakete)








- Gegeben: 7 .CEL-Dateien

3. Implementierung von MicroArrays in R

3.1 .CEL-Datei

```
9 # Installation:
10 source("https://bioconductor.org/biocLite.R")
11 biocLite("affy")
12 biocLite("hgu133plus2.db")
13 biocLite("affyQCReport")
14
15 # Lade affy package, hgu133plus2.db (Genzuordnung) und affyQCReport (QC-PDF) im Programm:
16 library("affy")
17 library("hgu133plus2.db")
18 library("affyQCReport")
19
20 # Lese alle .CEL-Dateien im "Working Directory" ein:
21 Data <- readAffy()
```

```
31 # Normalisierung:
32 Data.normalized <- normalize(Data)
33
34 # Diagramm-Erstellung:
35 vec <- c(1,3,4,5,6,7)
36 hist(Data)
37 hist(Data[,vec])
38 hist(Data.normalized)
39 hist(Data.normalized[,vec])
40
41 # RNA-Degradation-Plot:
42 degR1 <- AffyRNAdeg(Data)
43 degR2 <- AffyRNAdeg(Data[,vec])
44 degN1 <- AffyRNAdeg(Data.normalized)
45 degN2 <- AffyRNAdeg(Data.normalized[,vec])
46 plotAffyRNAdeg(degR1)
47 plotAffyRNAdeg(degR2)
48 plotAffyRNAdeg(degN1)
49 plotAffyRNAdeg(degN2)
```

 ND_1_CD14_TNF_90_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.237 KB
 ND_2_CD14_TNF_90_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.237 KB
 ND_3_CD14_TNF_90_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.241 KB
 ND_4_CD14_TNF_90_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.237 KB
 ND_51_CD14_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.243 KB
 ND_52_CD14_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.240 KB
 ND_53_CD14_133Plus_2.CEL	16.03.2016 22:56	CEL-Datei	13.238 KB

3. Implementierung von MicroArrays in R

3.1 .CEL-Datei

```
51 # RNA-Normalisierung:
52 eset <- rma(Data)
53
54 # affyIDs zu Genen zuordnen:
55
56 # Alle affyIDs aus ExpressionSet auslesen:
57 affyids <- featureNames(eset)
58
59 # 6 Beispiel-IDs ausgeben:
60 cat("\nAffyID Beispiele:\n")
61 print(affyids[1:6])
62
63 # Mit columns(hgu133plus2.db) kann abgefragt werden, was man aus den affyIDs erhalten moechte:
64 cat("\nBeispiele fuer ''Ausgaben'' mit select\n")
65 print(columns(hgu133plus2.db))
66
67 # wir entscheiden uns hier erstmal nur fuer "GENENAME" (mehrere koennten als vektor eingegeben werden):
68 genenames <- select(hgu133plus2.db, affyids, "GENENAME")
69
70 # wir erhalten eine Matrix mit folgenden Dimensionen:
71 cat("\nMatrixdimension:\n")
72 print(dim(genenames))
73 cat("1: AffyID, 2: Genname\n")
74
75 # Beispiel fuer Gennamen mit dazugehoeriger AffyID abfragen:
76 cat("\nBeispiele fuer Gennamen mit dazugehoeriger AffyID:\n")
77 print(genenames[1:5, ])
```

3. Implementierung von MicroArrays in R

3.2 Normalisierung

Frage: Was ist eine RMA-Normalisierung?

Antwort: Robust Multichip Average(RMA). Benutzt nur PMs mehrerer Arrays. Erst erfolgt eine Hintergrundkorrektur, $E[X | X + Y = \text{PM intensity}]$

anschließend werden die Werte normalisiert und addiert.

$$\log_2(PM_{ip}) = \theta_i + \psi_p + \text{error}$$

Frage: Warum eine Normalisierung?

Antwort: Zwischen zwei Arrays ändert sich die Expression einzelner Gene, allerdings nicht die globalen Eigenschaften der Verteilung der Expressionswerte -> Histogramme werden deshalb angeglichen.

Frage: Welche von den vorherigen Werten werden von der RMA-Normalisierung genutzt?

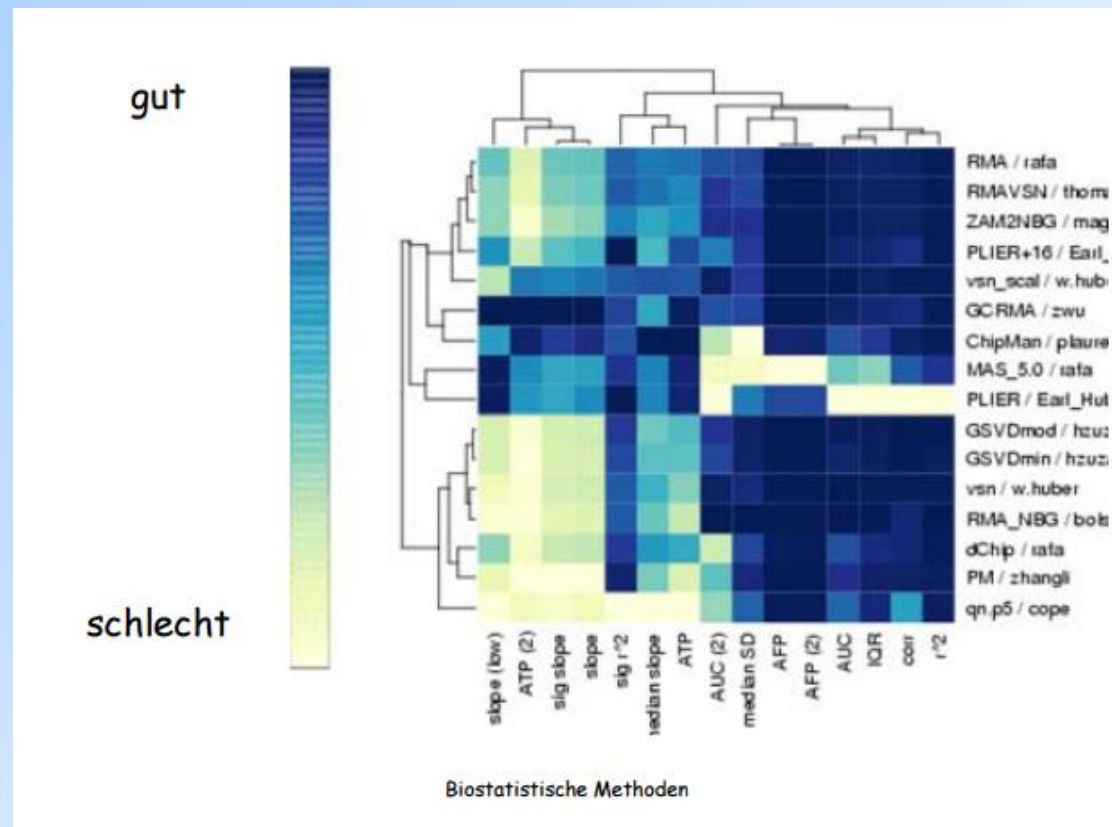
Antwort: PMs mehrerer Arrays.

3. Implementierung von MicroArrays in R

3.2 Normalisierung

Frage: Welche anderen Normalisierungen gibt es?

Antwort: z.B. mas5.



-> Es gibt keine beste Normalisierungsmethode. Die Auswahl eines optimalen Verfahrens hängt vom Ziel der angestrebten Analyse ab.

4. Qualitätsanalyse von MicroArrays in R

4.1 Chip image

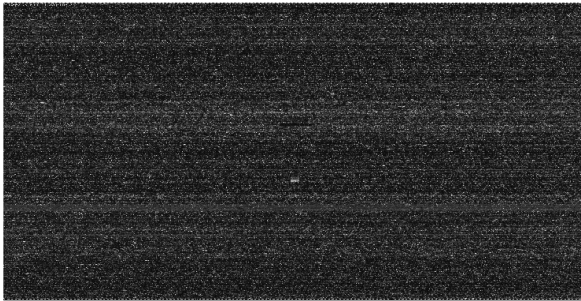
- Ziel Qualitätsanalyse: Überprüfung ob Hybridisierung fehlerfrei verlief
- Chip image = "Bild der Hybridisierung"
- Name oben links auf Chip image hybridisiert
- Weißes Kreuz mittig
- Karomuster in Ecken und Rand für Zuordnung der Probe-Koordinaten
- Verschiedene Farbmarkierungen des image möglich

```
51 # image auslesen:  
52 image(Data)
```

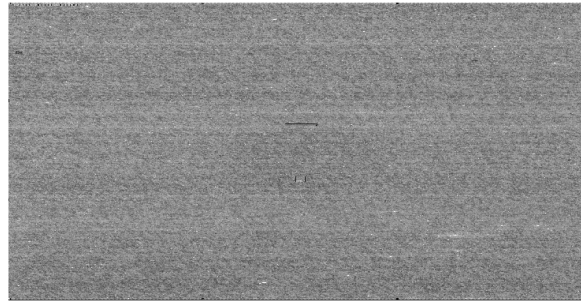

4. Qualitätsanalyse von MicroArrays in R

4.1 Chip image

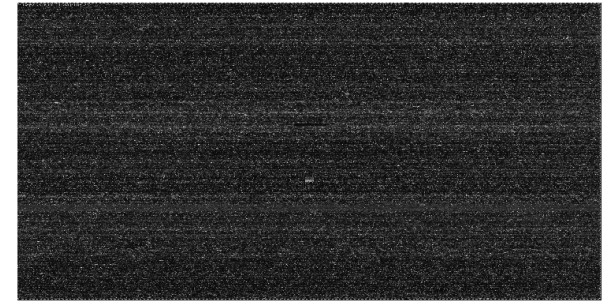
ND_1_CD14_TNF_90_133Plus_2.CEL



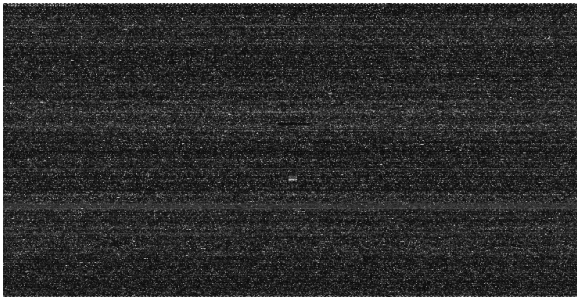
ND_2_CD14_TNF_90_133Plus_2.CEL



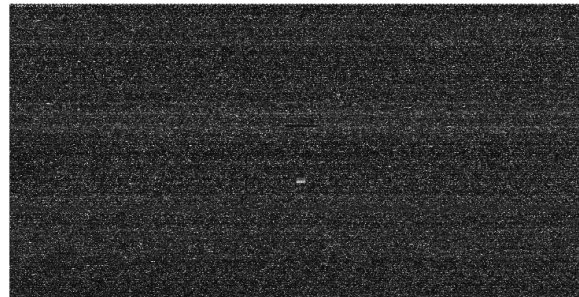
ND_3_CD14_TNF_90_133Plus_2.CEL



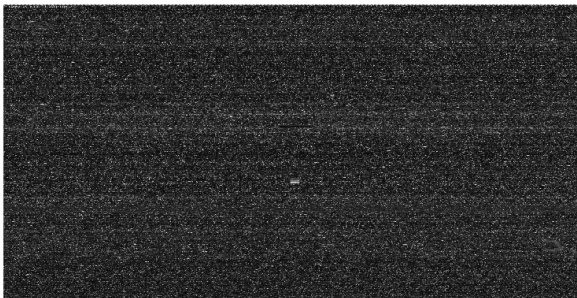
ND_4_CD14_TNF_90_133Plus_2.CEL



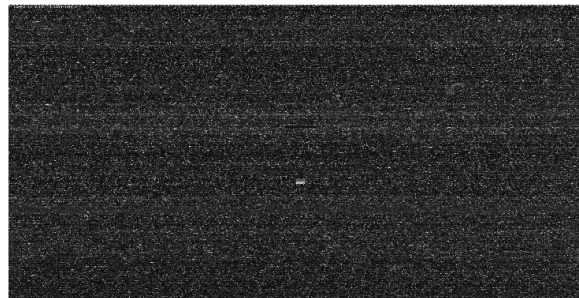
ND_51_CD14_133Plus_2.CEL



ND_52_CD14_133Plus_2.CEL



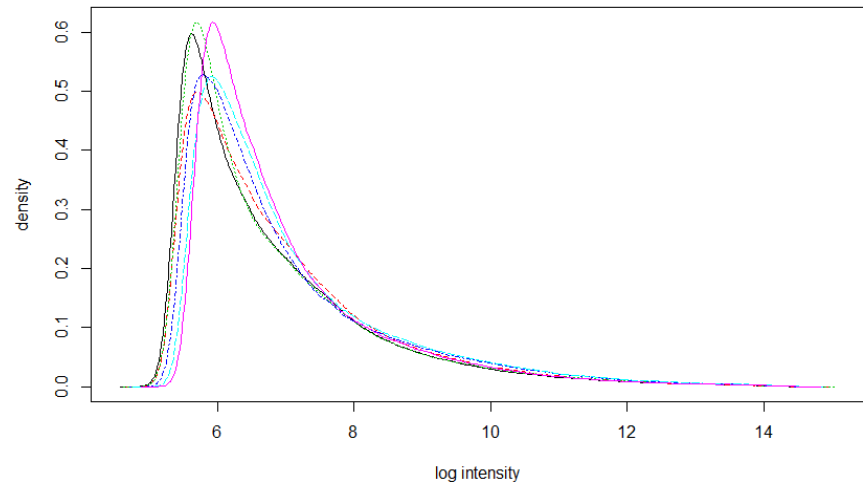
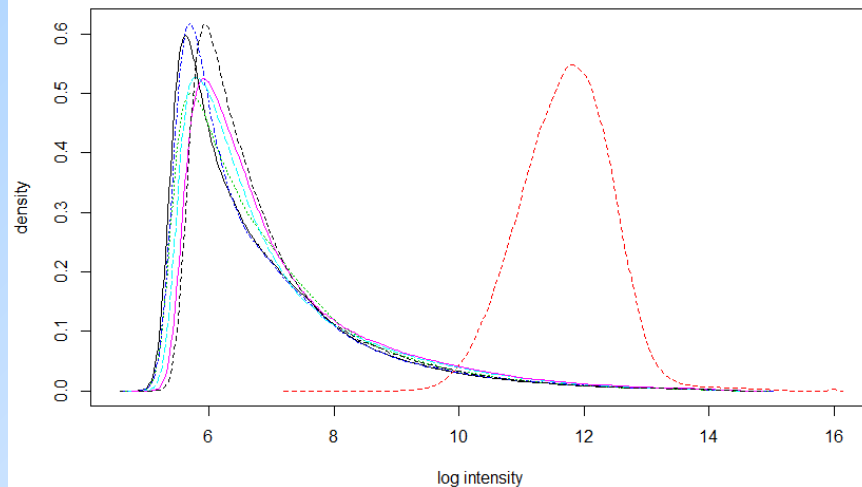
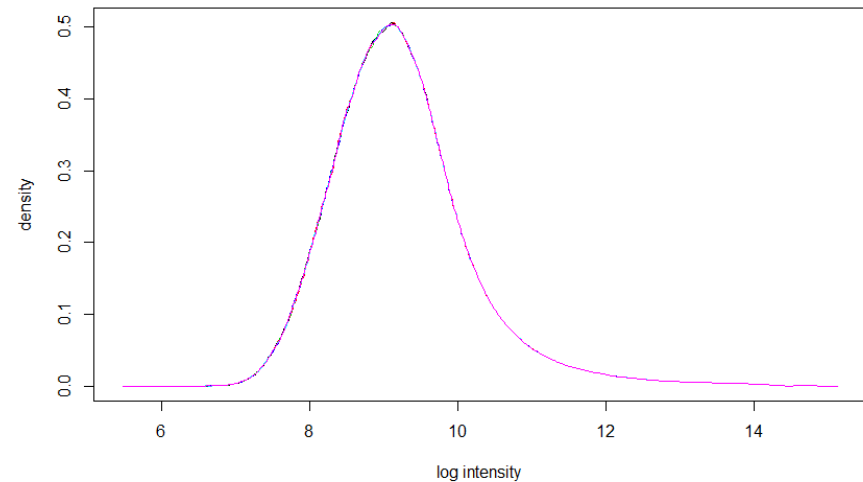
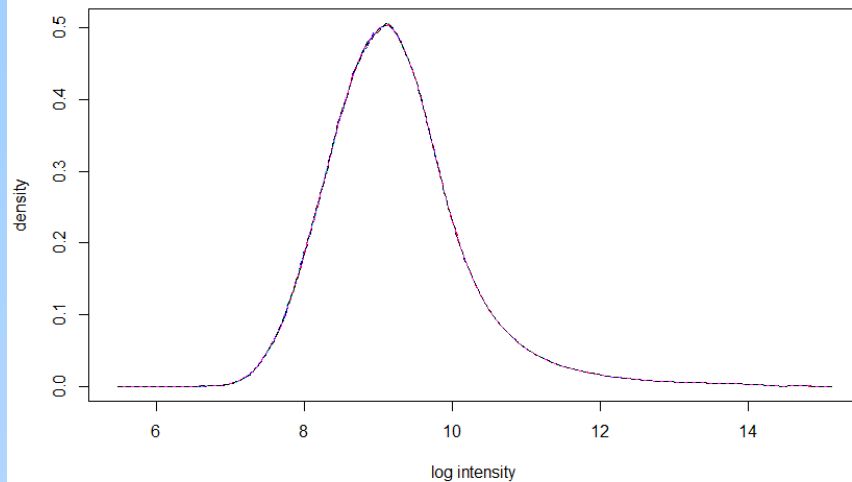
ND_53_CD14_133Plus_2.CEL



4. Qualitätsanalyse von MicroArrays in R

4.2 Histogramme

-> Histogramme = Signalverteilungen der PMs auf MicroArray

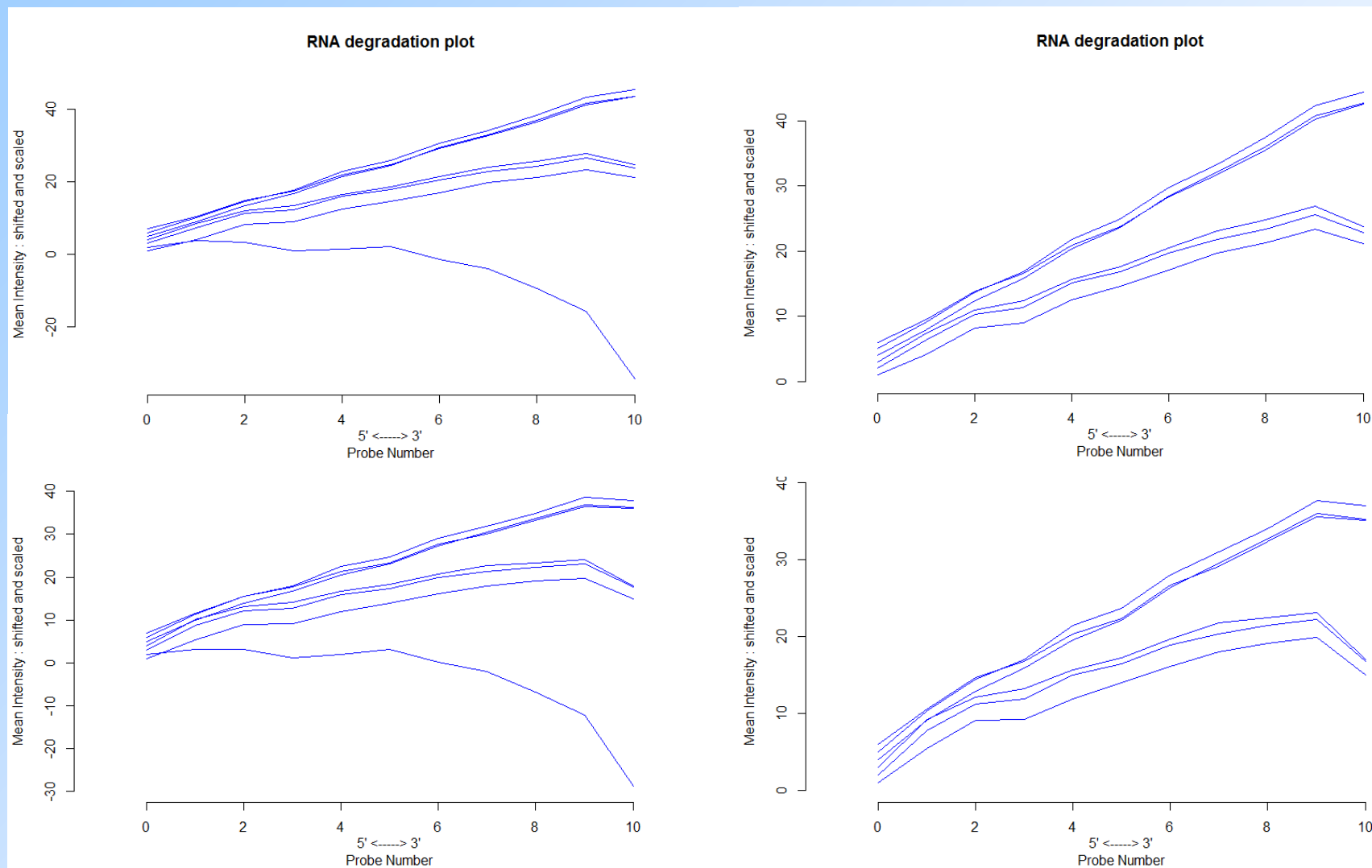


4. Qualitätsanalyse von MicroArrays in R

4.3 RNA-Degradation-Plot

Frage: Wie gut verlief die Hybridisierung von 5' nach 3'?

Im Plot: 45° Winkel Kurve -> gute Hybridisierung



4. Qualitätsanalyse von MicroArrays in R

4.4 Quality-Control Plot

Rohdaten

Normalisiert

Frage: Was sagt der Quality-Control Plot aus?

Antwort: Wie gut alle Chips im Vergleich zueinander sind/ob sie Vergleichbar sind/Bestimmte Markergene

```
79 # Quality-Control PDF erstellen:  
80 QCReport(Data,file="w1QC.pdf")  
81 QCReport(Data.normalized,file="w1QC_Normalisiert.pdf")
```

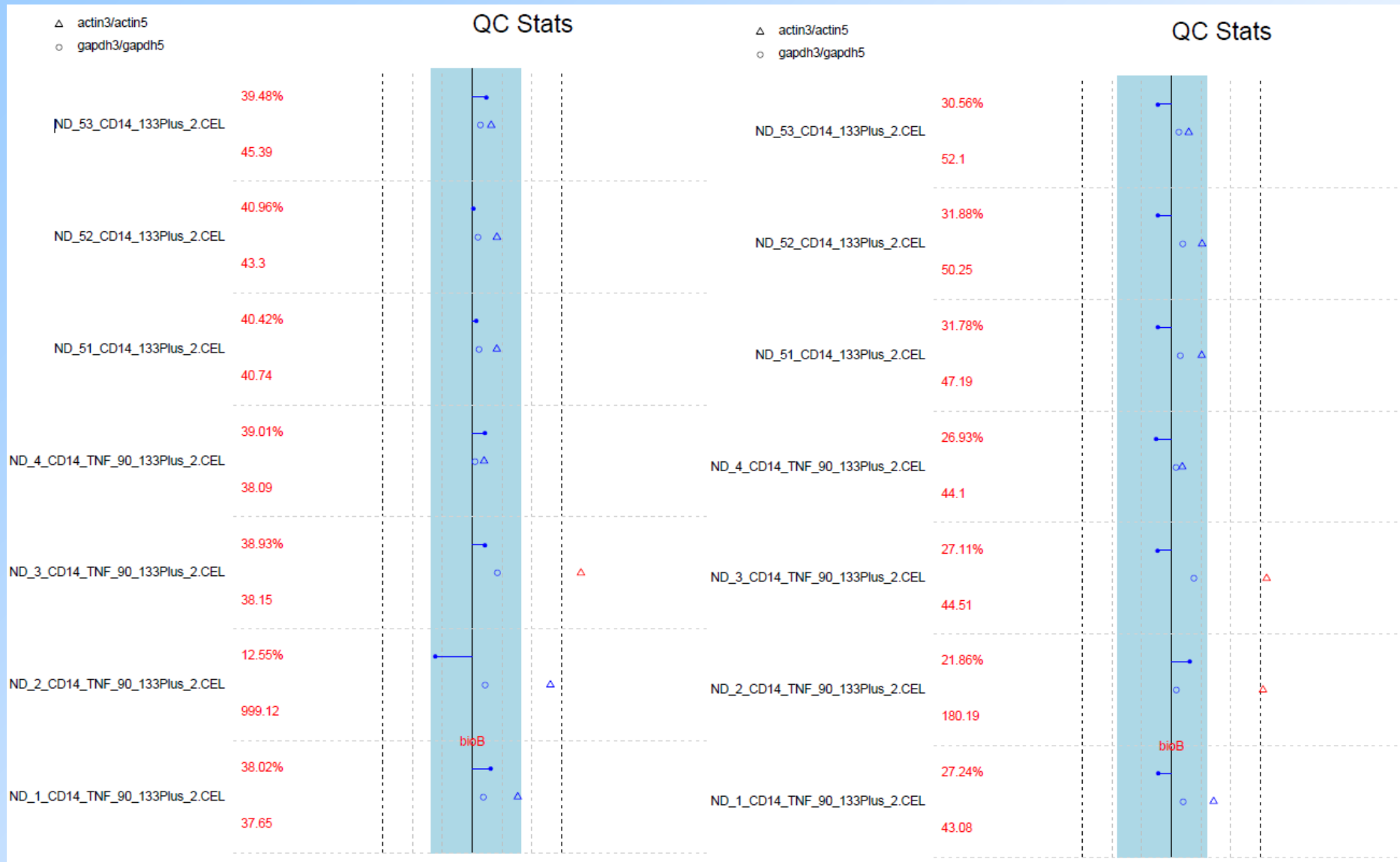
4. Qualitätsanalyse von MicroArrays in R

4.4 Quality-Control Plot

Rohdaten

Normalisiert

Frage: Was sagt der Quality-Control Plot aus?



Vielen Dank für Ihre Aufmerksamkeit!

5. Quellen

Daten:

<http://www.charite-bioinformatik.de/lehre.html>

Informationen:

<https://www.youtube.com/watch?v=oV7YFpyWg3E>

<https://www.youtube.com/watch?v=Bxl7p7MR95c>

<https://www.bioconductor.org/help/workflows/arrays/>

<http://bioinformatics.knowledgeblog.org/2011/06/20/analysing-microarray-data-in-bioconductor/>

<https://www.biostars.org/p/53870/>

http://www.biospektrum.de/blatt/d_bs_pdf&_id=932850

https://en.wikipedia.org/wiki/DNA_microarray

<https://bioconductor.org/packages/release/data/annotation/html/hgu133plus2cdf.html>

<http://homer.salk.edu/homer/basicTutorial/affymetrix.html>

<http://lectures.molgen.mpg.de/swp13/pres.pdf>

http://jura.wi.mit.edu/bio/education/bioinfo2007/arrays/array_exercises_1R.html

<https://en.wikipedia.org/wiki/Affymetrix>

<https://www.biostars.org/p/9677/>

[Astrand,Magnus\(2008\): Normalization and Differential Gene Expression of Microarray Data](#)

Bilder:

http://www.affymetrix.com/fa/images/gene_profile_array_large.jpg

<http://www.cell.com/cms/attachment/560021/4031703/gr1.jpg>https://en.wikipedia.org/wiki/DNA_microarray#/media/File:Microarray_exp_horizontal.svg

<http://docplayer.org/docs-images/24/3568552/images/27-0.png>