

Genexpressionsanalyse II

R-Kurs 2014-01-28

Markus Kreuz

Datensatz:

lymphoma_data.RData

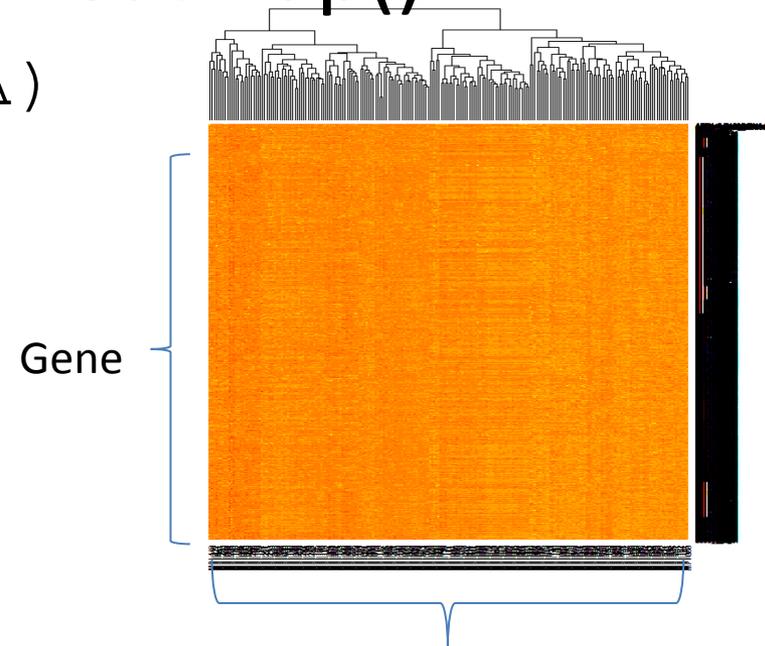
- 220 Lymphomfälle
 - Affymetrix hgu133a Array
 - 4457 Probesets mit höchster Streuung selektiert
- Annotation der Fälle (case_annotation):
 - MPINR (CHIP_ID)
 - AGE
 - GENDER
 - Diagnose (PANELDIAG1)
 - MYC-Translokation (MYC_STATUS: IG-MYC; non-IG-MYC; neg)
 - Genexpressionsbasierte Klassifikatoren:
 - GCBABC (GCB; ABC; unclassified)
 - MOLDIAG (mBL; non-mBL; intermediate)
 - COMAP (BL-PAP; mind-L; PAP-1; PAP-2; PAP-3; PAP-4)

- Daten entstammen der Publikation:

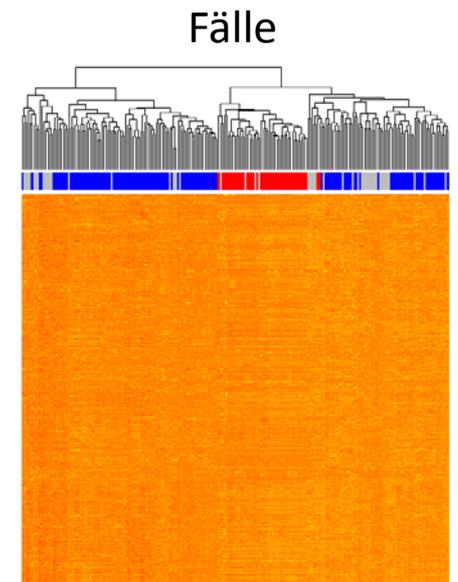
„A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling” Hummel M. et al. NEJM 2006

Visualisierung der Daten mit heatmap():

- `heatmap(GE_MMML, Rowv=NA)`
 - `Rowv=NA` verhindert Clustering der Zeilen (zeitaufwendig)
 - Fälle sind hierarchisch geclustert



- `cl<-rep("2",dim(case_annotation)[1])`
- `cl[case_annotation$MOLDIAG=="intermediate"] <-"8"`
- `cl[case_annotation$MOLDIAG=="non-mBL"] <-"4"`
- `heatmap(GE_MMML, Rowv=NA, ColSideColors=cl, labCol="", labRow="")`
 - Färbung der Fälle gemäß Molekularer Diagnose
rot=mBL, grau=intermediate,blau=non-mBL
 - `labCol/labRow` Labels für Zeilen und Spalten



Einfache Clusterverfahren:

K-Means:

```
library(cluster)
x<-kmeans(t(GE_MMML),centers=2)
table(x$cluster,case_annotation$MOLDIAG)
```

	intermediate	mBL	non-mBL
1	36	0	127
2	12	44	1

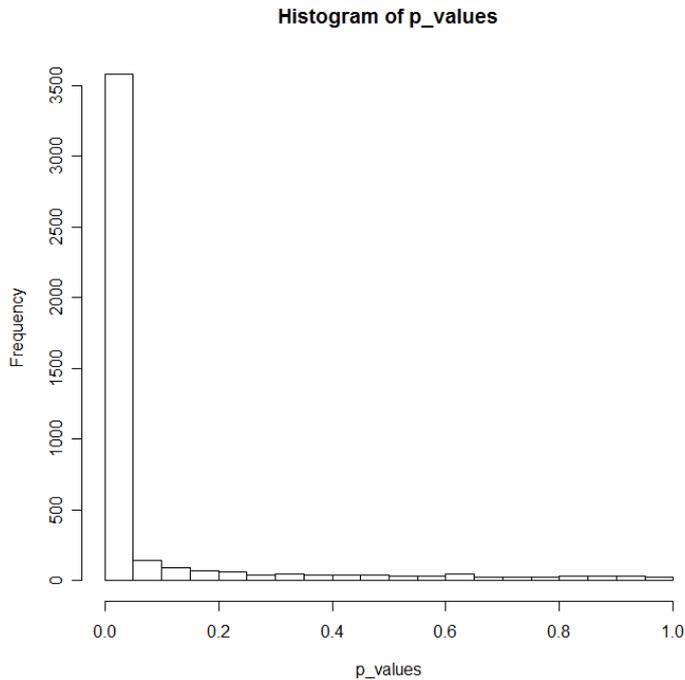
PAM:

```
x<-pam(t(GE_MMML),2)
table(x$clustering,case_annotation$MOLDIAG)
```

	intermediate	mBL	non-mBL
1	14	43	1
2	34	1	127

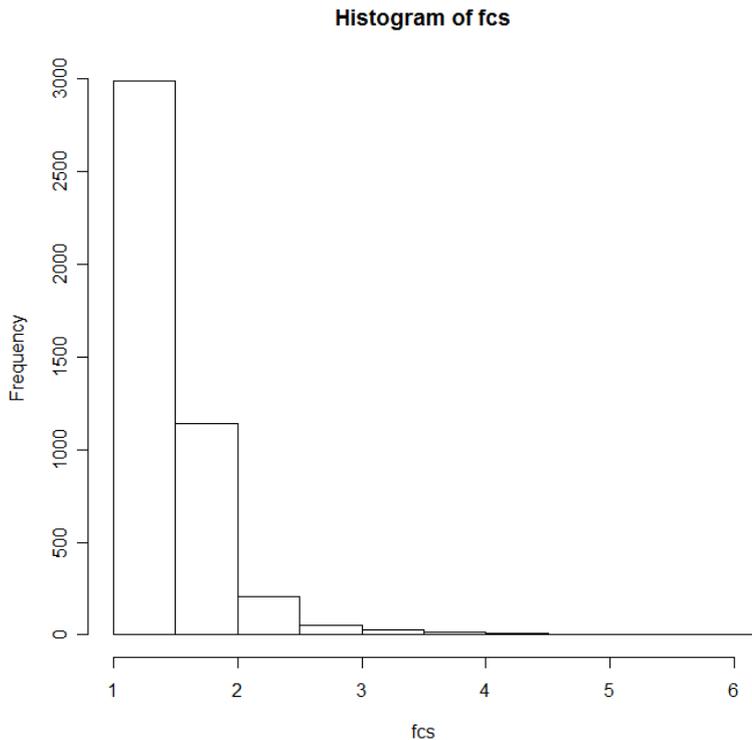
Differentielle Genexpression:

```
p_values<-rep(NA,dim(GE_MMML)[1])
for (i in 1:dim(GE_MMML)[1]){
  p_values[i]<-
    t.test(GE_MMML[i,case_annotation$MOLDIAG=="mBL"],
           GE_MMML[i,case_annotation$MOLDIAG=="non-mBL"])$
      p.value
}
hist(p_values,breaks=30)
```



Fold-Changes:

```
fcs<-rep(NA,dim(GE_MMML)[1])
for (i in 1:dim(GE_MMML)[1]){
  fcs[i]<-
  2^(abs(mean(GE_MMML[i,case_annotation$MOLDIAG=="mBL"])-
  mean(GE_MMML[i,case_annotation$MOLDIAG=="non-mBL"])))
}
hist(fcs,breaks=30)
```



Deskription differentielle Gene:

```
#Adjustierung für mult. Testen
```

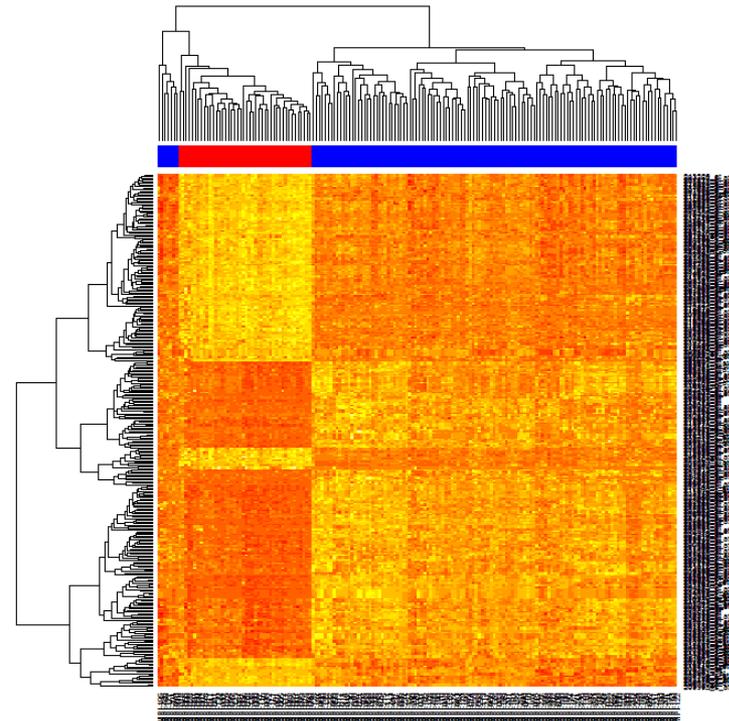
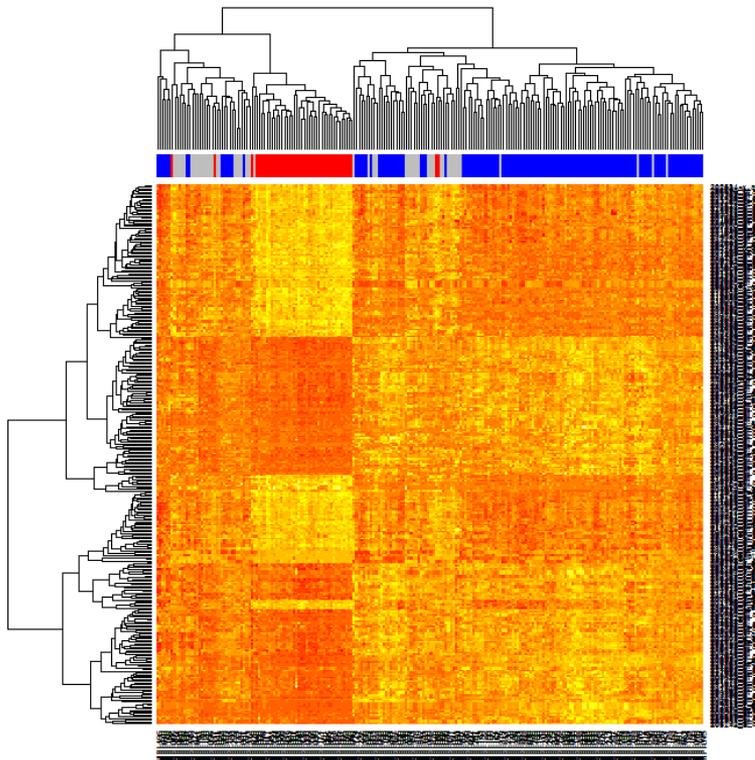
```
adjp=p.adjust(p_values, method="bonferroni")
```

```
#Darstellen differentieller Gene (signifikant und FC>2)
```

```
heatmap(GE_MMML[adjp<0.01 & fcs>2,],ColSideColors=cl)
```

```
#Ausschluss von „intermediate“ Fällen
```

```
heatmap(GE_MMML[adjp<0.01 &  
  fcs>2,case_annotation$MOLDIAG!="intermediate"],  
  ColSideColors=cl[case_annotation$MOLDIAG!="intermediate"])
```



Vergleich mit „zufälligem“ Kontrast:

```
p_values_2<-rep(NA,dim(GE_MMML)[1])
fcs_2<-rep(NA,dim(GE_MMML)[1])

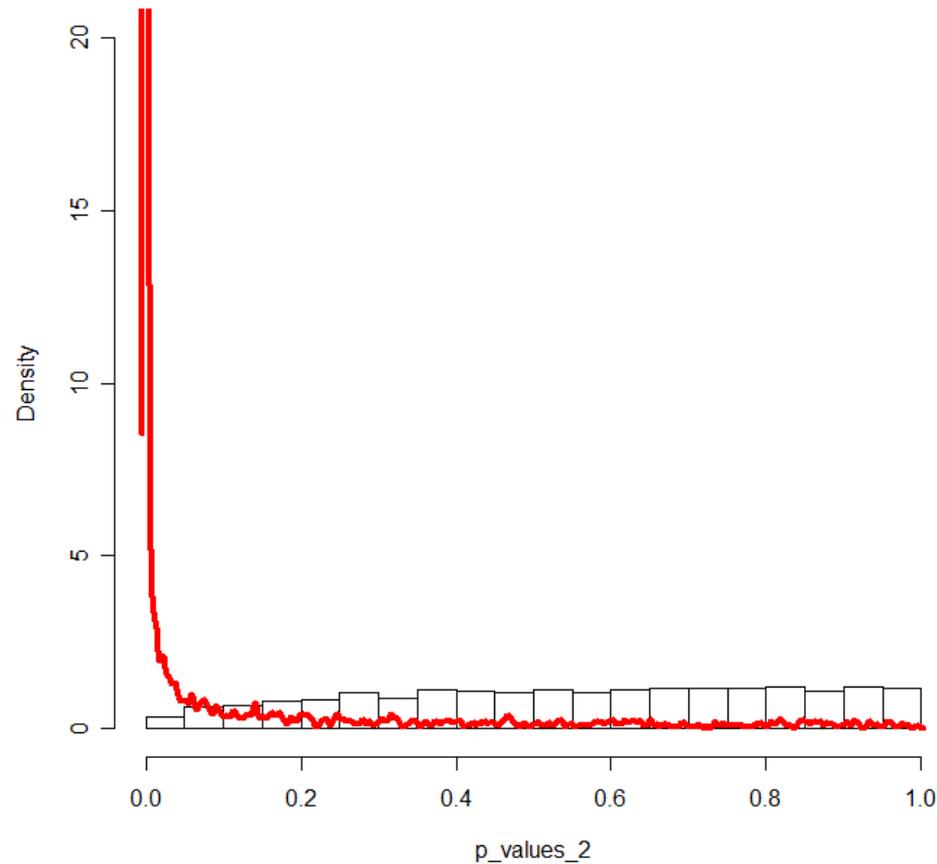
#Teile Datensatz zufällig in 2 Teile
sel<-sample(1:220,110)

#Berechne Differentielle GE und FCs zwischen den Gruppen
for (i in 1:dim(GE_MMML)[1]){
  p_values_2[i]<-t.test(GE_MMML[i,sel],GE_MMML[i,-
  sel])$p.value
  fcs_2[i]<-2^(abs(mean(GE_MMML[i,sel])-mean(GE_MMML[i,-
  sel])))
}

#p-Wert Verteilung
hist(p_values_2,breaks=30)
hist(fcs_2,breaks=30,xlim=c(1,6))
```

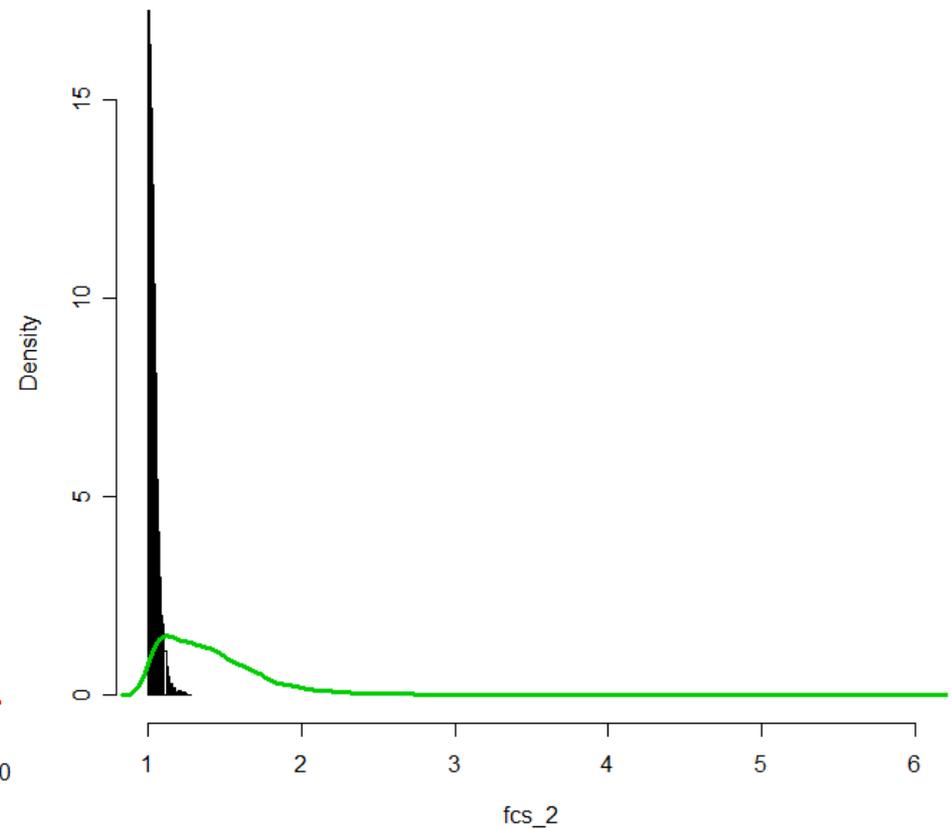
Vergleich mit „zufälligem“ Kontrast (II):

Histogram of p_values_2



**Rot = Dichte der p-Werte für
mBL vs. non-mBL Kontrast**

Histogram of fcs_2



**Grün = Dichte der FCs für
mBL vs. non-mBL Kontrast**

Aufgaben:

1. Beschreibe „deskriptiv“ und „induktiv“ (Testen) den Zusammenhang der Gruppe mBL mit Alter, Geschlecht, MYC-Status, GCBABC, COMAP und Paneldiagnose. Durch welche Eigenschaften kann die mBL Gruppe charakterisiert werden?
2. Untersuche innerhalb der non-mBL Gruppe den Kontrast ABC gegen GCB. Unterscheiden sich die Subgruppen bezüglich Genexpression? Korrelieren die Gruppe ABC und GCB mit anderen Merkmalen (vgl. Aufgabe 1)?
Achtung, für diese Aufgabe nur die Fälle mit non-mBL Signatur betrachten!
3. Zerlegen den kompletten Datensatz in 2 Teile, so dass mBL, intermediate und non-mBL möglichst gleich auf die Gruppen verteilt werden. Entwickle mit Hilfe des ersten Datensatzes eine einfache Klassifikationsregel, die die Fälle in 2 Gruppen (mBL und non-mBL) einteilt. Validiere die Regel im 2. Datensatz.
Erweitere die Klassifikationsregel, so dass unsichere Fälle (intermediate) in eine 3. Gruppe eingeteilt werden. Validiere erneut mit dem 2. Datensatz und schätze die Sensitivität/Spezifität für die Detektion von mBL mit dem Klassifikator ab.