

Genetische Statistik

Präsenzübung 10: Metaanalysen (GWAMA)

Dr. Janne Pott (janne.pott@uni-leipzig.de)

January 25, 2022

Fragen

Gibt es Fragen zu

- Vorlesung?
- Übung?
- Seminar?

Plan heute

Genomweite Meta-Analyse (R-Blatt 5)

Ausgangslage

Daten von 6 Studien

Schritt 1: Daten anschauen

Screenshot

The screenshot shows a terminal window titled "Study1.out" containing the command-line arguments for a SNPTEST analysis. The output is a text file with the following content:

```
1  # Analysis: "SNPTEST analysis, started 2018-06-21 20:28:24"
2  #   started: 2018-06-21 20:28:32
3  #
4  # Analysis properties:
5  #   -cov_names sex age BMI (user-supplied)
6  #   -data Study1_chr14_chunk15.gz Study1_Phenotypes.sample (user-supplied)
7  #   -exclude_snps excludedSNPs_chr14_chunk15.txt (user-supplied)
8  #   -frequentist 1 (user-supplied)
9  #   -hwe (user-supplied)
10 #   -method expected (user-supplied)
11 #   -o Study1.out (user-supplied)
12 #   -pheno CORT (user-supplied)
13 #   -use_raw_phenotypes (user-supplied)
14 #
15 alternate_ids rsid chromosome position alleleA alleleB index average_maximum_I
all_AA all_AB all_BB all_NULL all_total all_maf missing_data_proportion cohort
frequentist_add_beta_1 frequentist_add_se_1 comment
16 --- rs10133404:92574620:G:A NA 92574620 G A 1 0.999865 0.81258 5593.41 3.585 (
0.353644 0.204239 NA
17 --- 14:92574638:T:A NA 92574638 T A 2 1 1 5597 0 0 0 5597 0 0 0 5597 0 0 1 NA
```

Abbildung 1: Header des SNPTEST outputs

Screenshot - Auswertung

- pheno CORT → Phänotyp Cortisol (Steroidhormon)
- cov_names sex age BMI in Study 1 & 4, cov_names age BMI in Study 2, 3, 5 & 6 → Unterschiedliche Adjustierung, mögliche Fehlerquelle
- frequentist 1 → additives Modell
- method expected → Gendoses

Studie	1	2	3	4	5	6
Fallzahl	5597	2943	2654	2070	1358	712

Ausgangslage

Daten von 6 Studien

Schritt 1: Daten anschauen

Schritt 2: Daten filtern

Filter (vor GWAMA)

Auf was kann muss man filtern?

Filter (vor GWAMA)

Auf was kann muss man filtern?

- Vollständigkeit der Daten (beta, se, maf, Allele, ...)
- Gleiche IDs (sonst keine Meta-Analyse)
- **KEINE** Filterung von MAF, p-Wert oder LD, das kommt erst **NACH** der GWAMA

Filter (vor GWAMA) in R

```
pathToData = ".../Exercises_R/data2/"  
  
# laden  
tab1<-fread(paste0(pathToData,"Study1.out"),skip = 14)  
tab2<-fread(paste0(pathToData,"Study2.out"),skip = 14)  
tab3<-fread(paste0(pathToData,"Study3.out"),skip = 14)  
tab4<-fread(paste0(pathToData,"Study4.out"),skip = 14)  
tab5<-fread(paste0(pathToData,"Study5.out"),skip = 14)  
tab6<-fread(paste0(pathToData,"Study6.out"),skip = 14)
```

Filter (vor GWAMA) in R

```
# filtern auf Schnittmenge
sharedIDs = tab1[is.element(rsid,tab4$rsid),rsid]
tab1<-tab1[is.element(rsid,sharedIDs),]
tab2<-tab2[is.element(rsid,sharedIDs),]
tab3<-tab3[is.element(rsid,sharedIDs),]
tab4<-tab4[is.element(rsid,sharedIDs),]
tab5<-tab5[is.element(rsid,sharedIDs),]
tab6<-tab6[is.element(rsid,sharedIDs),]
```

Insgesamt sind imm 156420 SNPs in den Studien, aber nur 156415 in der Schnittmenge.

Filter (vor GWAMA) in R

```
# prüfe Reihenfolge (hier nur am Beispiel von tab1, tab2, und  
stopifnot(tab1$rsid==tab2$rsid)  
stopifnot(tab1$rsid==tab4$rsid)
```

Filter (vor GWAMA) in R

```
# prüfe Allele (hier nur am Beispiel von tab1, tab2, und tab4)
stopifnot(tab1$alleleA==tab2$alleleA)
stopifnot(tab1$alleleA==tab4$alleleA)
```

Filter (vor GWAMA) in R

```
# filtern auf NA
filt3<-!is.na(tab1$comment) &
  !is.na(tab2$comment) &
  !is.na(tab3$comment) &
  !is.na(tab4$comment) &
  !is.na(tab5$comment) &
  !is.na(tab6$comment)
table(filt3)
```

```
## filt3
## FALSE  TRUE
## 59872 96543
```

```
table(filt3,!is.na(tab1$comment))
```

```
##
## filt3  FALSE  TRUE
## FALSE 57145  2727
```

Filter (vor GWAMA) in R

```
tab1<-tab1[!filt3,]  
tab2<-tab2[!filt3,]  
tab3<-tab3[!filt3,]  
tab4<-tab4[!filt3,]  
tab5<-tab5[!filt3,]  
tab6<-tab6[!filt3,]
```

Insgesamt sind 96543 SNPs immer NA. Diese werden gefiltert. Für die Meta-Analyse stehen daher 59872 SNPs zur Verfügung.

Ausgangslage

Daten von 6 Studien

Schritt 1: Daten anschauen

Schritt 2: Daten filtern

Schritt 3: Meta-Analyse durchführen

Was ist mein Input?

Was ist mein Output?

GWAMA

Was ist mein Input?

- Beta & SE pro Studie
- Für Post-GWAMA-QC: EAF, info

Was ist mein Output?

- Statisiken von FEM & REM (beta, SE, p, I^2)
- Minimale Info
- Gewichtete EAF

GWAMA in R

```
j<-seq(1,60000,10000)
#dumTab = foreach(i = 1:dim(tab1)[1])%do%{
dumTab = foreach(i = 1:50)%do%{
  #i=1
  if(is.element(i,j)==T)message(paste0("working on SNP ", i))
  myBetas<-c(tab1$frequentist_add_beta_1[i],tab2$frequentist_a
  mySEs<-c(tab1$frequentist_add_se_1[i],tab2$frequentist_add_s
  mod<-metagen(myBetas,mySEs,studlab = c("Study 1","Study 2","
  n = c(tab1$all_total[i], tab2$all_total[i], tab3$all_total[i]
  totalN = sum(n)
  nWeight = n/totalN
  maf = c(tab1$all_maf[i] , tab2$all_maf[i], tab3$all_maf[i], t
  nWeightedMAF = sum(nWeight * maf)

  minInfo = min(c(tab1$info[i], tab2$info[i], tab3$info[i], tab4$info[i]))
```

GWAMA in R

```
## working on SNP 1

## working on SNP 10001

## working on SNP 20001

## working on SNP 30001

## working on SNP 40001

## working on SNP 50001

## [1] "beta_FEM"          "se_FEM"           "pval_FEM"         "beta_REM"
## [6] "pval_REM"          "k"                 "I2"                "MAF_weig
## [11] "chr"               "pos"              "SNP"
```

GWAMA in R

```
# QQ und Manhattan Plot vorbereiten  
table(myTab$chromosome)
```

```
## < table of extent 0 >
```

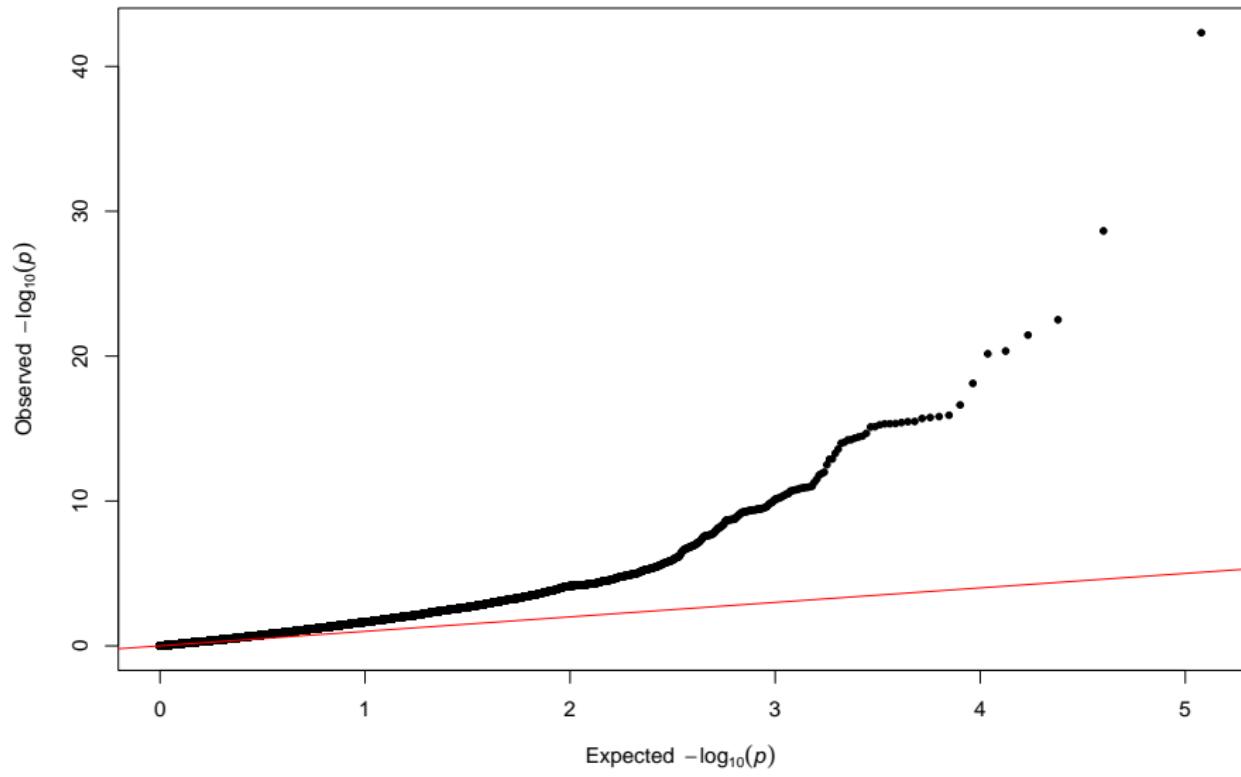
```
Y1<-qchisq(1-myTab$pval_FEM,df = 1)  
lambda1<-median(Y1)/0.456  
Y2<-qchisq(1-myTab$pval_Rem,df = 1)  
lambda2<-median(Y2)/0.456  
lambda1;lambda2
```

```
## [1] 1.75964
```

```
## [1] 1.611853
```

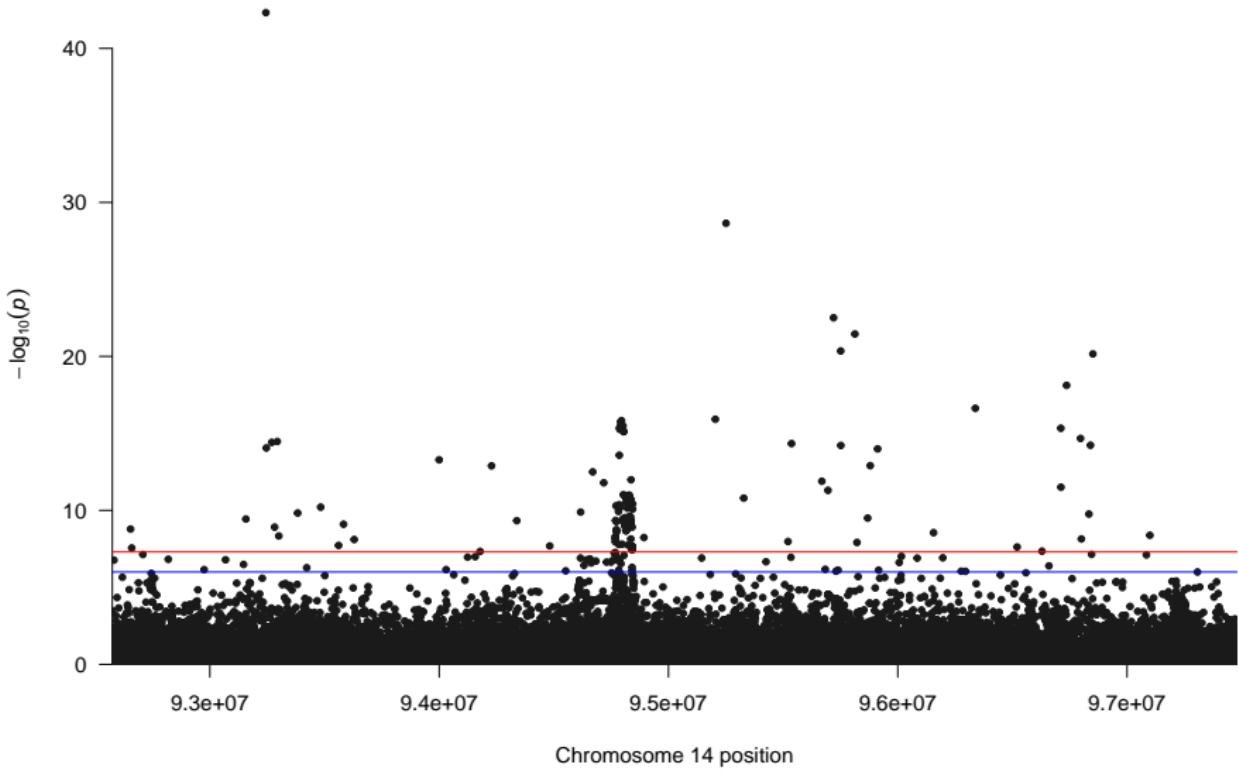
GWAMA in R: QQ - FEM

QQ-Plot FEM, vor QC, Inflation: 1.76



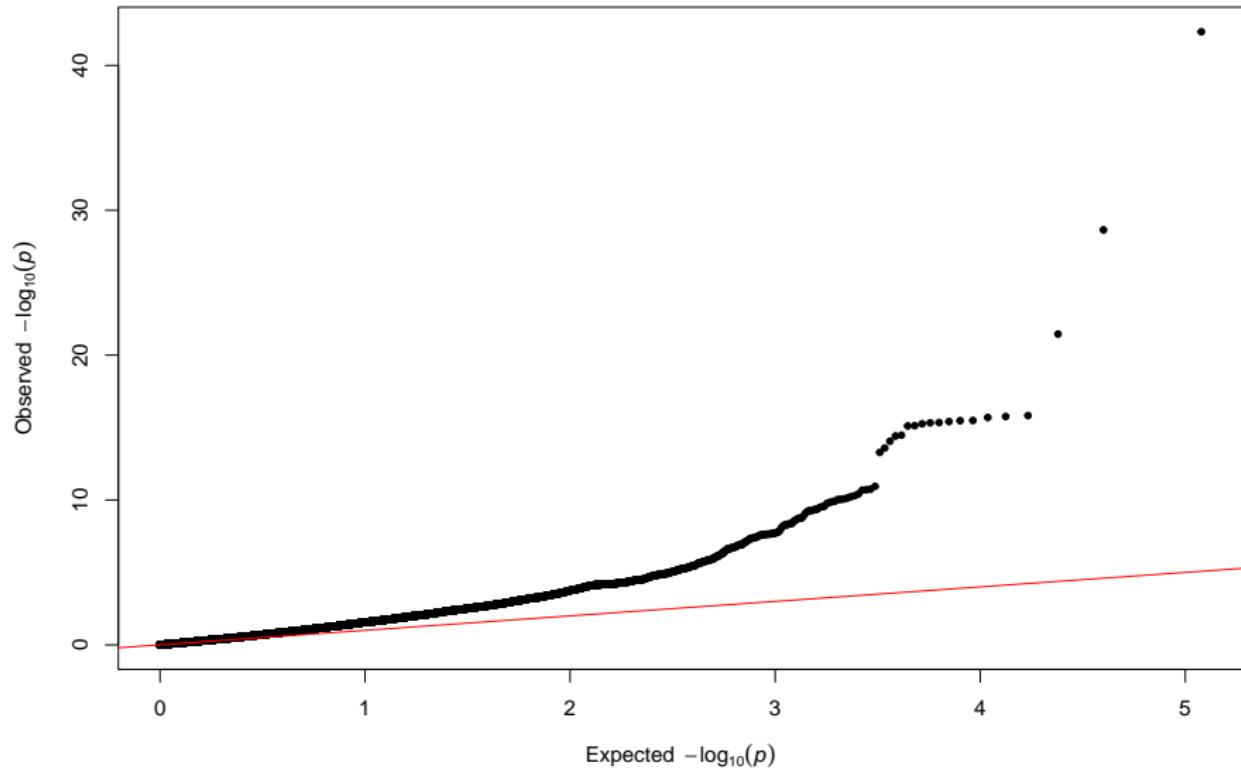
GWAMA in R: Manhattan - FEM

Manhattan-Plot FEM, vor QC



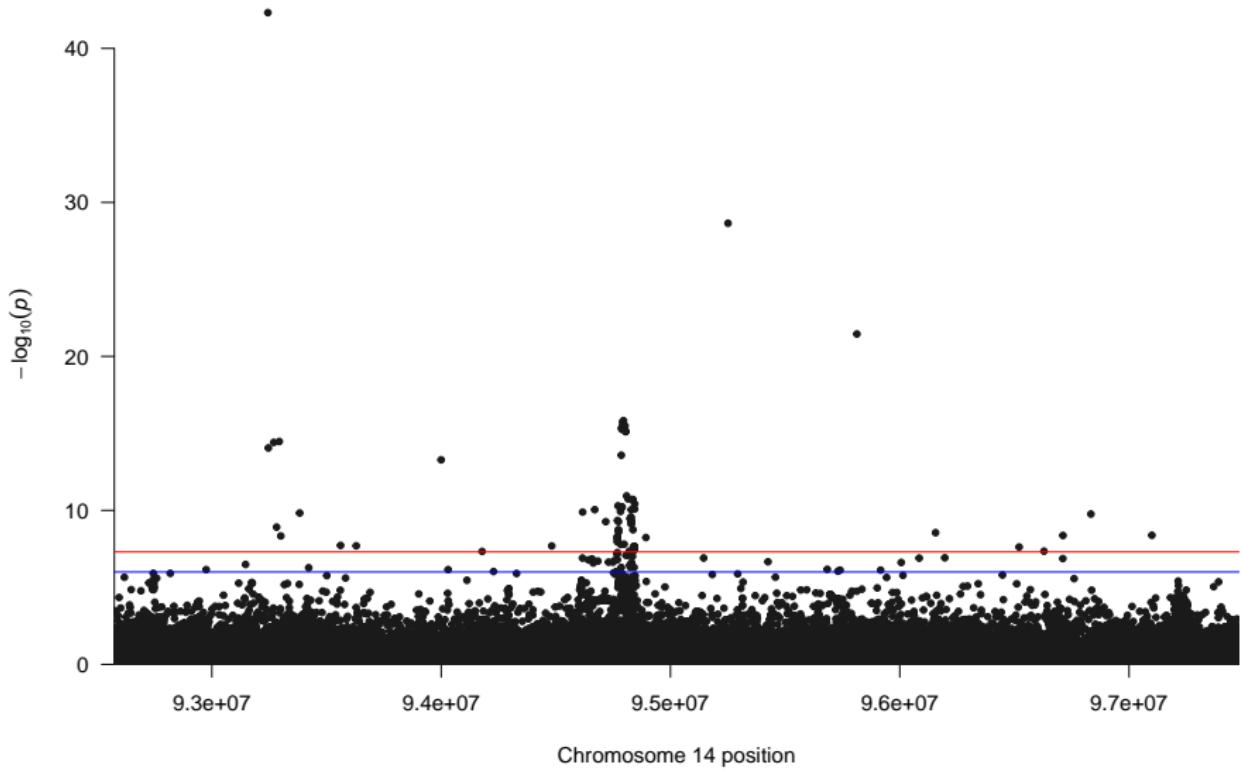
GWAMA in R: QQ - REM

QQ-Plot REM, vor QC, Inflation: 1.612



GWAMA in R: Manhattan - REM

Manhattan-Plot REM, vor QC



Problem: hohe Inflation

Ursache?

Problem: hohe Inflation

Ursache:

- unterschiedliche Modelle
- Populationsstruktur in den Studien! Study 1 = Study 2 (Männer) + Study 3 (Frauen)

Lösung:

- Nur mit Study 2, 3, 5, und 6 arbeiten!
- Anschließend λ neu ausrechnen und bewerten
- Falls ok, dann SNPs filtern

Filter (nach GWAMA)

- MAF<0.01: minor allele frequency, man filtert meistens die ganz seltenen Varianten raus
- info<0.5: Imputationsqualität, abhängig von Analyseplan, wird oft auch höher angesetzt, z.B. 0.8
- $p(Q)<0.05$: Cochrans Q Statistik, Heterogenitätsmaß, Summe der gewichteten Differenzen, p-Wert mit $k-1$ Freiheitsgraden; bei wenigen Studien wird die Grenze noch oben korrigiert, z.B 0.1
- $I^2>75\%$: I^2 Statistik, Heterogenitätsmaß, Prozentsatz der Variation zwischen Studien
- $k<2$: Anzahl der Studien
- $p < 5 \times 10^{-8}$: p-Wert der Meta-Analyse (Standard = FEM)
- $r^2>0.5$: LD-Maß, statistische Abhängigkeit der Marker, der bestassoziierte SNP bleibt, SNPs in LD zu diesem werden gefiltert

Online Annotation

Lead SNP: rs9989237

- dbSNP
- GeneCards
- GTEx
- GWAS Catalog
- LD-Tool
- KEGG Pathway