Preprocessing rat data

Hannah Meyer

2018-07-18

1. Dataset

The dataset consists of an outbred population of rats descended from eight inbred progenitors. 2,006 outbred rats were phenotypes for 195 traits of biomedical relevance including hematological, immunological, cardiovascular and fear-related phenotypes. 1,407 of the rats were genotyped [1],[2].

The phenotype and genotype data was downloaded from arrayexpress and figshare. Supplementary information about the data (distribution, type, etc) was taken from [1], [2].

```r
## directory ####
directory <- "~/data/LiMMBo/rat/rawdata"

## from figshare ####
measures <- data.table::fread(paste(directory, "/figshare/measures.txt", sep=""),
                  data.table=FALSE, stringsAsFactors = FALSE)

## from Baud et al (2014) ####
pheno_ScientificData <- data.table::fread(paste(directory,
                                 "/PublicationResults/ScientificDataS1.csv",
                                 sep=""),
                           data.table=FALSE, stringsAsFactors = FALSE,
                           skip=2)[, -c(6:7)]
colnames(pheno_ScientificData) <- gsub(" ", "", colnames(pheno_ScientificData))
pheno_ScientificData$gsub <- gsub("_normalized_by_batch", "",
                                  pheno_ScientificData$Measure)
## from Baud et al (2013) ####
pheno_NatGenet <- data.table::fread(paste(directory,
                              "/PublicationResults/",
                              "SupplementaryTable1_Phenotypes.csv", sep=""),
                              data.table=FALSE, stringsAsFactors = FALSE,
                        skip=3)
colnames(pheno_NatGenet) <- gsub(" ", "", colnames(pheno_NatGenet))
pheno_NatGenet$MeasureinGSCAN <- gsub("_bc", "", pheno_NatGenet$MeasureinGSCAN)
```

2. Sorting data

The rat phenotype and genotype will be used in a multi-variate association analysis. The data downloaded from figshare contains both, phenotypes and measures used as covariates in downstream analysis. In the following, `measures` is split based on their classification in the [2]. Only phenotypes with approximately normal distribution will be kept for this analysis, based on the original studies assumption of normality (as indicated with Mixed model in the analysis section of [2]).

a) Phenotypes

```
covariate_names <- dplyr::filter(pheno_ScientificData,
                                 Phenotypingtest == "Covariate")
phenotype_names <- dplyr::filter(pheno_ScientificData,
                                 Phenotypingtest != "Covariate")

phenotype_names <- merge(phenotype_names, pheno_NatGenet[, -c(1:2,8)],
                         by.x="gsub", by.y="MeasureinGSCAN", all=TRUE)

phenotype_names_notN <- dplyr::filter(phenotype_names,
                                           Mappingmethod != "Mixed models",
                                           !is.na(Measure))
phenotype_names_N <- dplyr::filter(phenotype_names,
                                  Mappingmethod == "Mixed models",
                                  !is.na(Measure))

phenotypes_normal <- measures[, colnames(measures) %in%
                                  phenotype_names_N$Measure ]
rownames(phenotypes_normal) <- measures$SUBJECT.NAME

phenotypes_notNormal <- measures[, colnames(measures) %in%
                                     phenotype_names_notN$Measure ]
rownames(phenotypes_notNormal) <- measures$SUBJECT.NAME


write.table(phenotypes_normal, paste(directory, "/phenotypes_normal.csv",
                                     sep=""),
          row.names=TRUE, col.names=NA, quote=FALSE, sep=",")
write.table(phenotypes_notNormal, paste(directory, "/phenotypes_notNormal.csv",
                                        sep=""),
            row.names=TRUE, col.names=NA, quote=FALSE, sep=",")
```

b) Covariates

```
covariates <- measures[, colnames(measures) %in%
                           c("BW_week11", "BW_week14", "BW_week17",
                             covariate_names$Measure) ]
covariates <- apply(covariates, 2, function(x) {
    if (! is.numeric(x)) return(as.numeric(as.factor(x)))
    if (is.numeric(x)) return(x)
})
rownames(covariates) <- measures$SUBJECT.NAME

# manually match covariate names from NatGenetc supp info to colnames of measure/
# covariates
#colnames(covariates)
#[1] "BW_week11"   "age_dissection"  "analyser_ear_hole"  "batch"
#[5] "blood_pressure_time"   "cage"  "date_IPGTT"   "date_of_birth"
#[9] "died_of_EAE"   "sibship"   "Haemalysis_serum"   "is.albino"
#[13] "litter_size"   "max_EAE_score"   "reliable_ear_hole"   "scanner_ear_hole"
#[17] "EAE_score_at_sacrifice" "sex"   "IPGTT_test_worked"   "BW_week14"
```

```
#[21] "BW_week17"   "Exp_novel_cage"   "Exp_zero_maze" "Exp_shuttlebox"

## information about conversion of these variables directly from from Amelie Baud
colnames(covariates)[colnames(covariates) == "BW_week11"] <- "BW_at_IPGTT"
colnames(covariates)[colnames(covariates) == "BW_week14"] <- "BW_at_day9_pi"
colnames(covariates)[colnames(covariates) == "BW_week17"] <- "BW_at_day28_pi"
colnames(covariates)[colnames(covariates) ==  "age_dissection"] <- "age"
colnames(covariates)[colnames(covariates) == "Haemalysis_serum"] <- "Haemalysis"
colnames(covariates)[colnames(covariates) == "IPGTT_test_worked"] <- "test_worked"
```

3. Map covariates to phenotypes

For downstream analysis, map covariates to the phenotypes they were associated with in the original study. Save relevant covariates and mapping.

```
phenotype_names_N <- phenotype_names_N[phenotype_names_N$Measure %in%
                                           colnames(phenotypes_normal),]
phenotype_names_N <- phenotype_names_N[  match(colnames(phenotypes_normal),
                                               phenotype_names_N$Measure),]

phenotype2covs <- lapply(1:nrow(phenotype_names_N), function(x) {
    unlist(strsplit(phenotype_names_N$Covariates[x], ","))
})
names(phenotype2covs) <- phenotype_names_N$Measure
saveRDS(phenotype2covs, paste(directory, "/phenotype2covs.rds", sep=""))

uniqueCovariatesTested <- unique(unlist(phenotype2covs))
#[1] "sex"             "batch"          "is.albino"      "Haemalysis"
#[5] "BW_at_day28_pi" "BW_at_day9_pi"  "age" "BW_at_IPGTT"    "test_worked"
relevantCovariates <- covariates[,colnames(covariates) %in% uniqueCovariatesTested]
write.table(relevantCovariates, paste(directory, "/covariates.csv", sep=""),
            row.names=TRUE, col.names=NA, quote=FALSE, sep=",")
```

1. Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nature Genetics. Nature Publishing Group; 2013;45: 767–775. doi:10.1038/ng.2644

2. Baud A, Guryev V, Hummel O, Johannesson M, Hermsen R, Stridh P, et al. Genomes and phenomes of a population of outbred rats and its progenitors. Scientific Data. Nature Publishing Group; 2014;1: 140011. doi:10.1038/sdata.2014.11