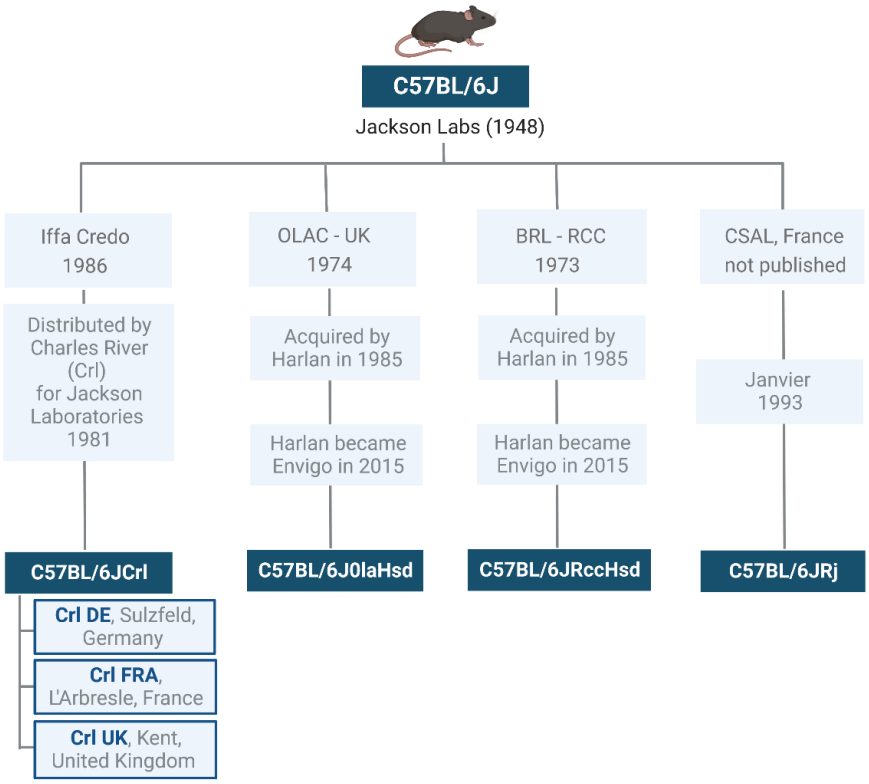
Using mice from different breeders fails to improve replicability of results

Ivana Jaric[1*], Bernhard Voelkl[1], Irmgard Amrein[2], David P. Wolfer[2], Janja Novak[1], Carlotta Detotto[3], Ulrike Weber-Stadlbauer[4], Urs Meyer[4], Francesca Manuella[5], Isabelle Mansuy[5], Hanno Würbel[1*]

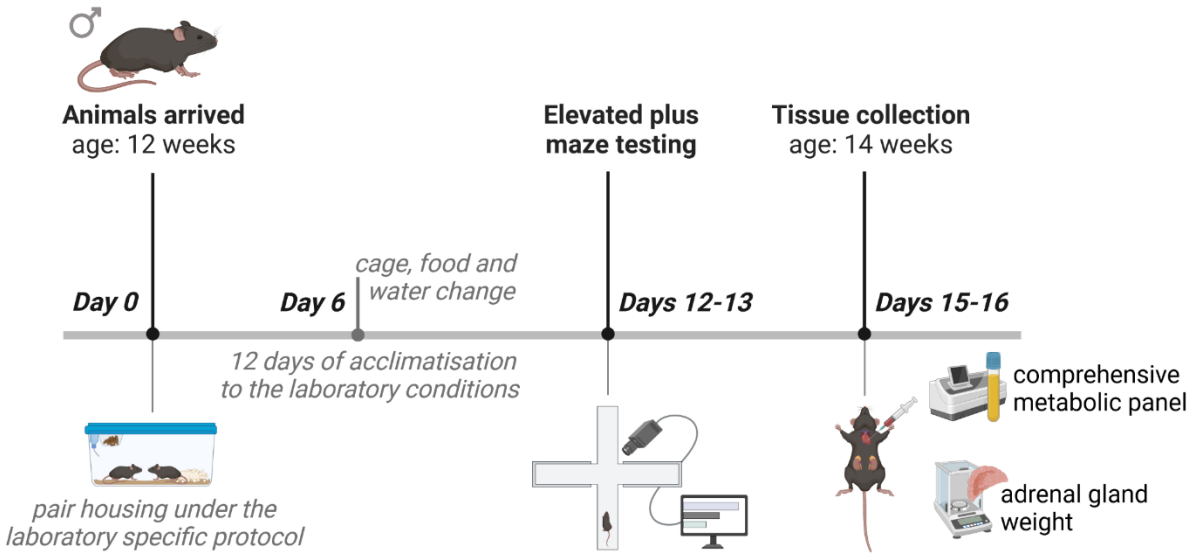
Supplementary Information

Supplementary Figures

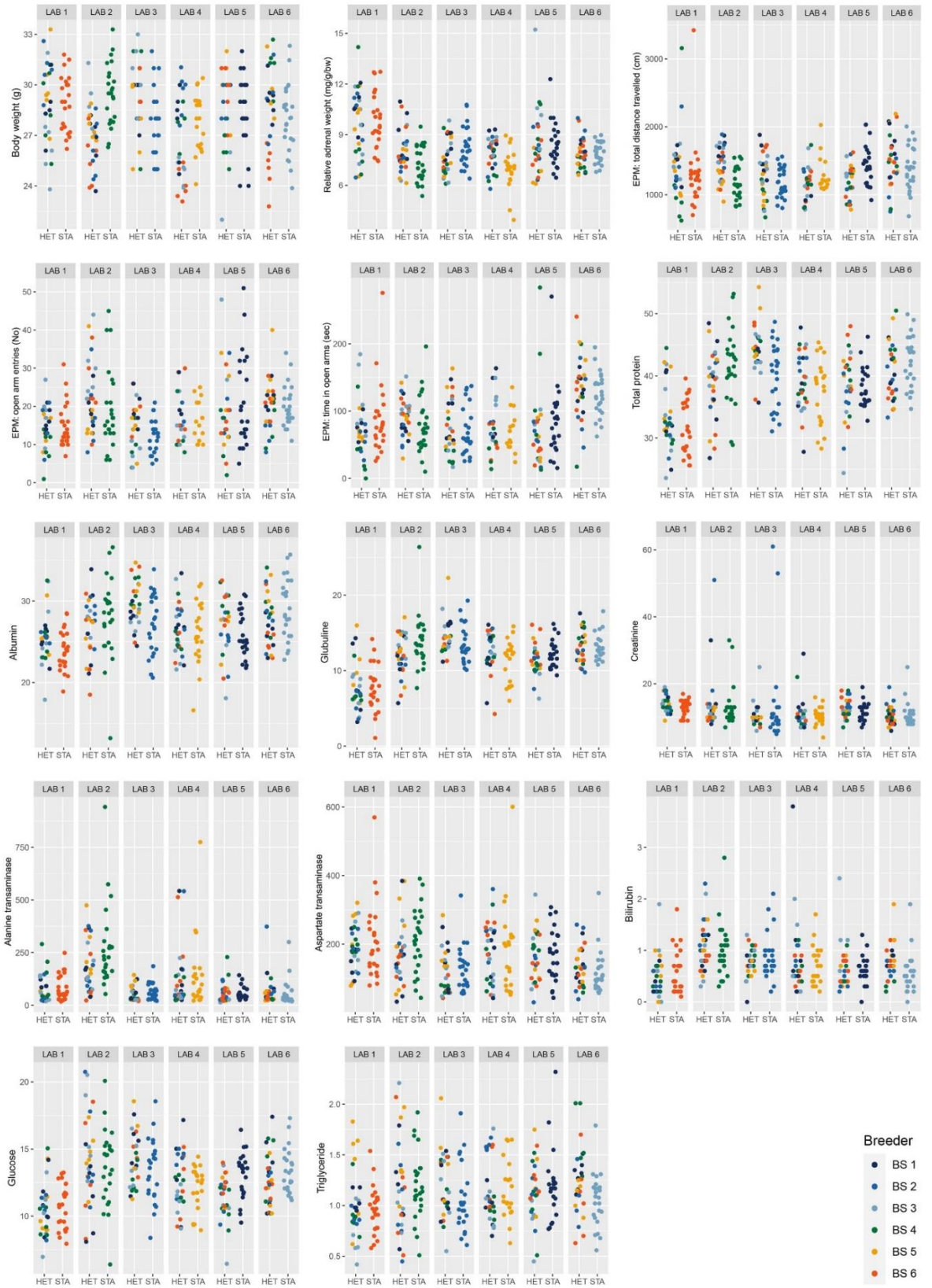
Supplementary Figure 1: The origin of B57BL/6J mice we used in our experimental setup as heterogenization factor.



Supplementary Figure 2: Timeline of the study



Supplementary Figure 3: Raw data plotted for each outcome measure and testing laboratory for both STA and HET study designs



Supplementary Table 1: Detailed housing and husbandry conditions in each testing laboratories (given as a separate excel file)

Sample size calculation, Simulated Sampling with the R markdown code

The aim of the simulation was to get an estimate for the frequency with which we can find a difference in the variances between the STA and HET experimental design. Sample size are calculated for the case of 6 participating testing laboratories. We assume 12 independent variables. For two third (8) variables we assume a feeble effect (sd of the latent variable = 0.1) of testing laboratory, for two variables we assume a medium sized effect (sd=2) and for two variables we assume a strong effect (sd=15); for the residual variation we assume a strong individual effect (sd=15). The critical statistic is the f-ratio of the means squares of the ANOVAs for the factor "lab" of the STA and HET groups with 1 and 5 df. For the HET group we set the sample size to 30 per laboratory, which allows a complete balanced design (with regard to rearing lab and breeder) within each STA group. For the STA group we started with a sample size of 30 per testing laboratory. As this resulted in a power of 0.884, which was above the aspired threshold level of 0.8, we subsequently reduced the sample size for the STA group in increments of two until the estimated power fell below 0.8. The recommended sample size for the STA group is the smallest sample size with an estimated power above 0.8, which is 24 subjects per group (*i.e.* testing laboratory) with an estimated power of 0.825.

```
#Initialization

setwd("M:/REP FAIL/Experiment B4")

library(lmerTest)

# Functions

simhom<- function(sdlab, sderr,n){ labeffect<-
   rnorm(6, mean=0, sd=sdlab) test<-c()

   for(k in 1:6){

      test<-c(test,rep(labeffect[k],n)+rnorm(n, mean=100, sd=sderr))

   }

   test

}

simhet<- function(sdlab, sderr){ labeffect<-rnorm(6,
   mean=0, sd=sdlab) test<-c()

   for(k in 1:6){
```

```
      labtest<-c(labtest,rep(labeffect[i],6)+rnorm(6, mean=100, sd=sderr))
    }
    pos<-c(6*k-5,6*k-4,6*k-3,6*k-2,6*k-1,6*k)
    test<-c(test,labtest[-pos])
  }
  test
}

# Simulation N(hom=30)
n=30
labhom<-as.factor(rep(1:6,each=n))
labhet<-as.factor(rep(1:6,each=30))
set.seed(1066)

f1<-c()
for(i in 1:1000){
  outcomes<-mapply(simhom,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
,n)
  pca<-prcomp(outcomes)
  model<-lm(pca$x[,1]~labhom)
  atab<-anova(model)
  ms1<-atab$`Mean Sq`[1]
  outcomes<-mapply(simhet,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
)
  pca<-prcomp(outcomes)
  model<-lm(pca$x[,1]~labhet)
  atab<-anova(model)
  ms2<-atab$`Mean Sq`[1]
  f1<-c(f1,ms1/ms2)
}
length(which(f1 >6.6))/1000

## [1] 0.884 power estimate

# Simulation N(hom=24)
n=24
labhom<-as.factor(rep(1:6,each=n))
labhet<-as.factor(rep(1:6,each=30))
set.seed(1066)

f1<-c()
for(i in 1:1000){
  outcomes<-mapply(simhom,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
,n)
  pca<-prcomp(outcomes)
  model<-lm(pca$x[,1]~labhom)
  atab<-anova(model)
  ms1<-atab$`Mean Sq`[1]
  outcomes<-mapply(simhet,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
)
```

```
  pca<-prcomp(outcomes)
  model<-
  lm(pca$x[,1]~labhet)
  atab<-anova(model)
  ms2<-atab$`Mean
  Sq`[1] f1<-
  c(f1,ms1/ms2)
}
length(which(f1 >6.6))/1000

## [1] 0.825 power estimate

# Simulation N(hom=22)
n=22
labhom<-
as.factor(rep(1:6,each=n))
labhet<-
as.factor(rep(1:6,each=30))
set.seed(1066)

f1<-c()
for(i in 1:1000){
  outcomes<-
  mapply(simhom,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
,n)
  pca<-prcomp(outcomes)
  model<-
  lm(pca$x[,1]~labhom)
  atab<-anova(model)
  ms1<-atab$`Mean Sq`[1]
  outcomes<-
  mapply(simhet,sample(c(rep(0.1,8),rep(1,2),rep(15,2))),rep(15,12)
)
  pca<-prcomp(outcomes)
  model<-
  lm(pca$x[,1]~labhet)
  atab<-anova(model)
  ms2<-atab$`Mean
  Sq`[1] f1<-
  c(f1,ms1/ms2)
}
length(which(f1 >6.6))/1000

## [1] 0.757 power estimate
```

R-markdown: Statistical analysis (given as a separate pdf file).