1 Reworked analyses and figure drafts for paper1 after JEB reviews

```
Loading required package: Hmisc
Loading required package: grid
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2
   Attaching package: 'Hmisc'
   The following objects are masked from 'package:base':
       format.pval, round.POSIXt, trunc.POSIXt, units
   Loading required package: lme4
Loading required package: Matrix
Loading required package: compiler
Loading required package: car
[1] 0
```

```
[1] 133
```

1.1 Some obvious comparisons

1.1.1 Germination rate as a function of gene-family

And here's another test of the same hypothesis using permutation approach (may be less affected by unequal var).

```
[1] "typeI prob:"
[1] 0.012
```

1.2 Survival to harvest as a function of gene family

Here are analyses that examine the effect of gene family upon survival:

```
                Df Sum Sq Mean Sq F value Pr(>F)
GeneFamilyName  12 0.1829 0.01524   0.758  0.692
Residuals      105 2.1122 0.02012
```

1.3 Read in the salk copy numbers and process

```
[1] 121   9
```

2 Distribution of mutations

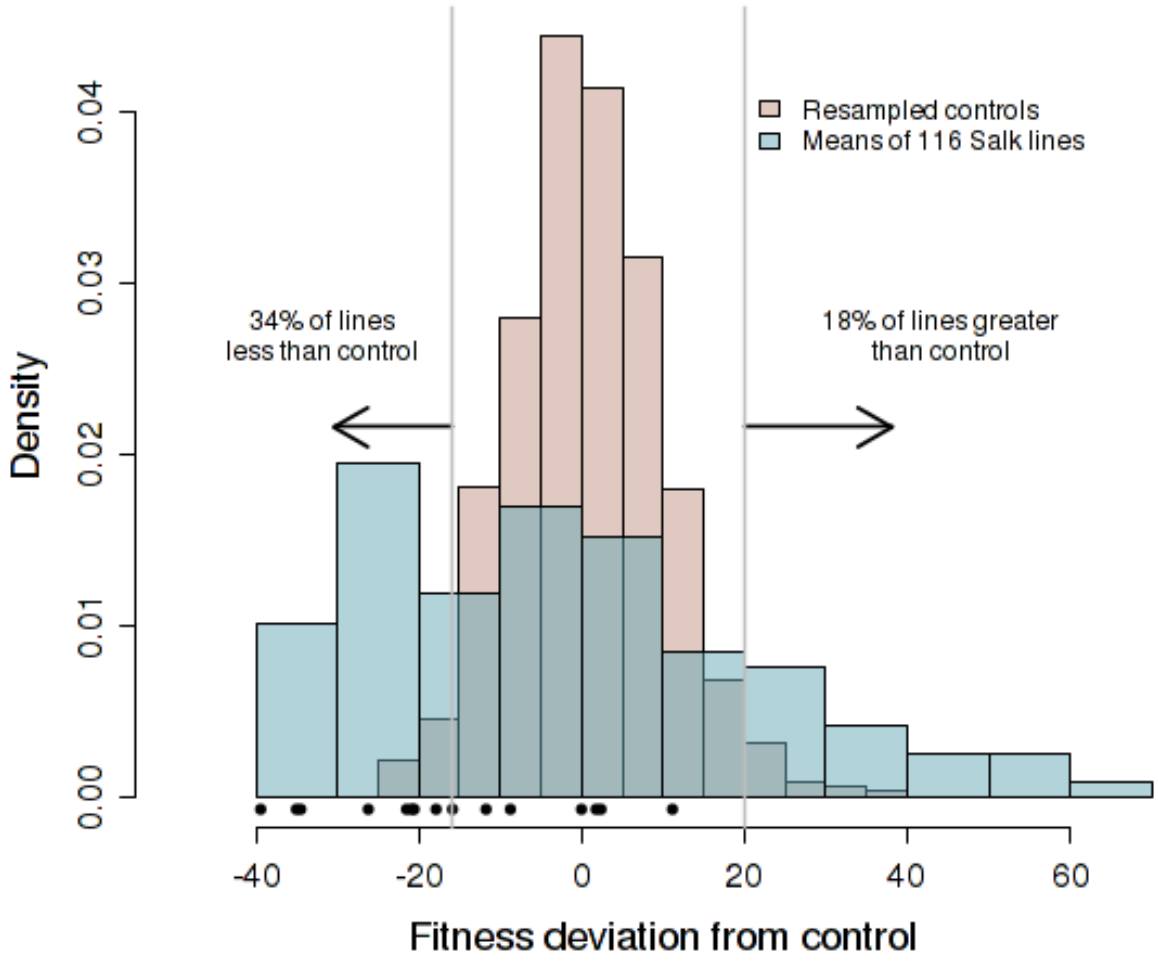
This is a redraw of the theoretical figure (fig 1) using R code:

Here are various figures emphasizing variation among mutant lines compared to controls. In this first figure, the distribution of line means is plotted with the distribution of all control plants as well as 116 resampled means of control replicates equal in size to the average number of reps per SALK line

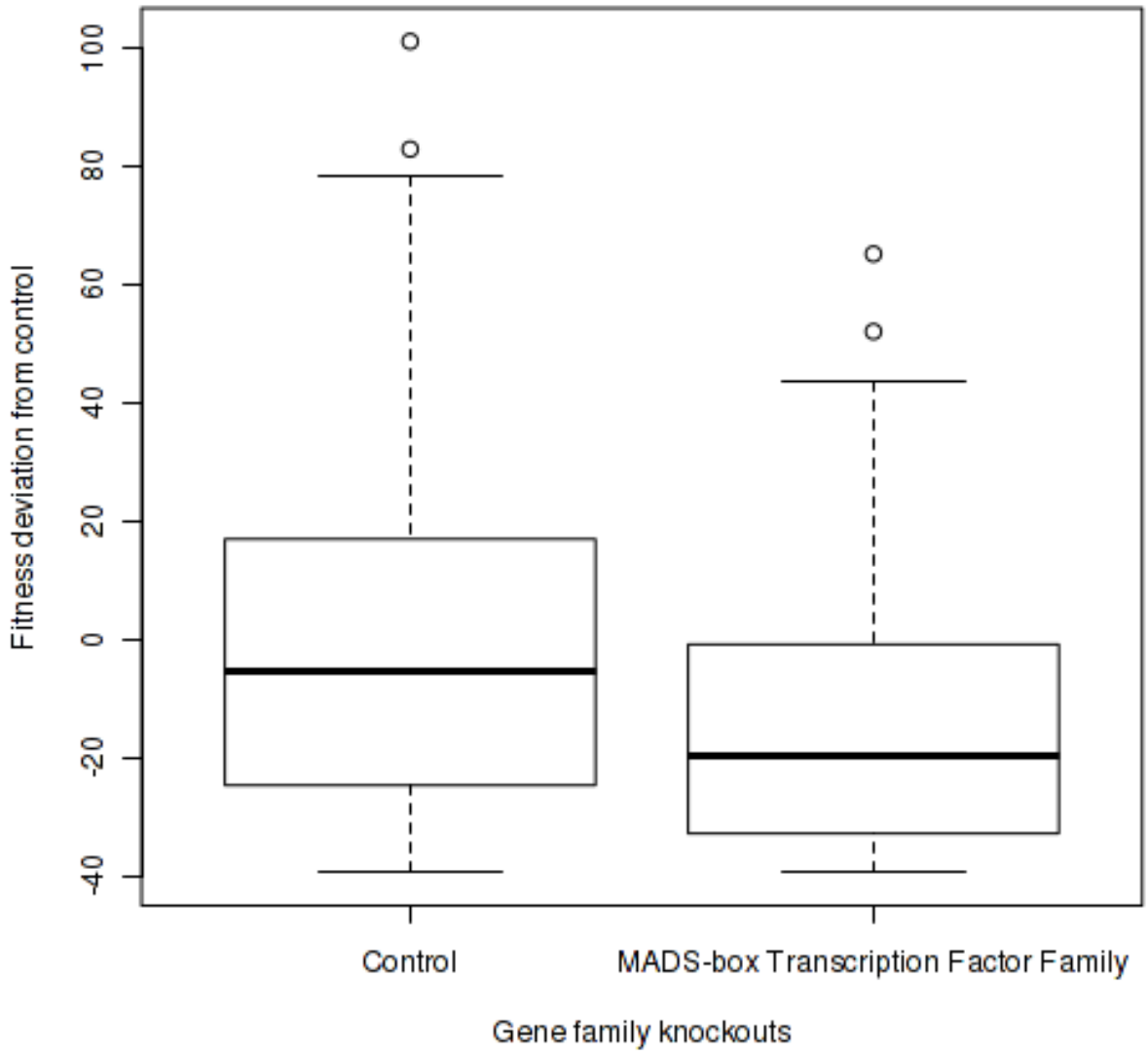
```
[1] "pdf"
```

This figure compares just the resampled means to the SALK line means

```
[1] "pdf"
```



Here are the number of line means less than control

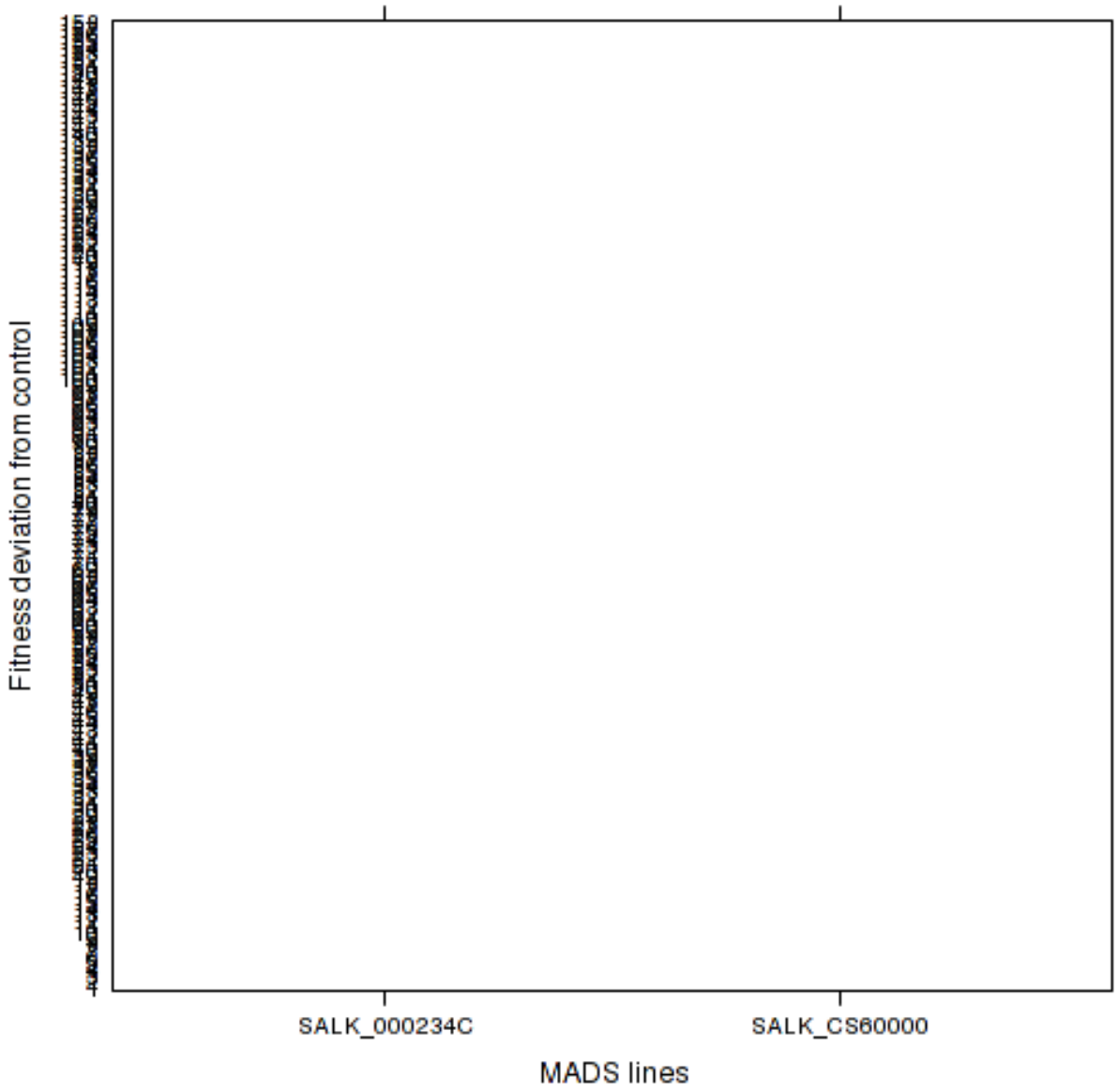



```
## Note: no visible binding for global variable '.Data'
## Note: no visible binding for global variable '.Data'
## Note: no visible binding for global variable '.Data'
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## GeneFamilyName   1   8841    8841   10.38 0.00146 **
## Residuals      227 193427     852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 72 observations deleted due to missingness
```

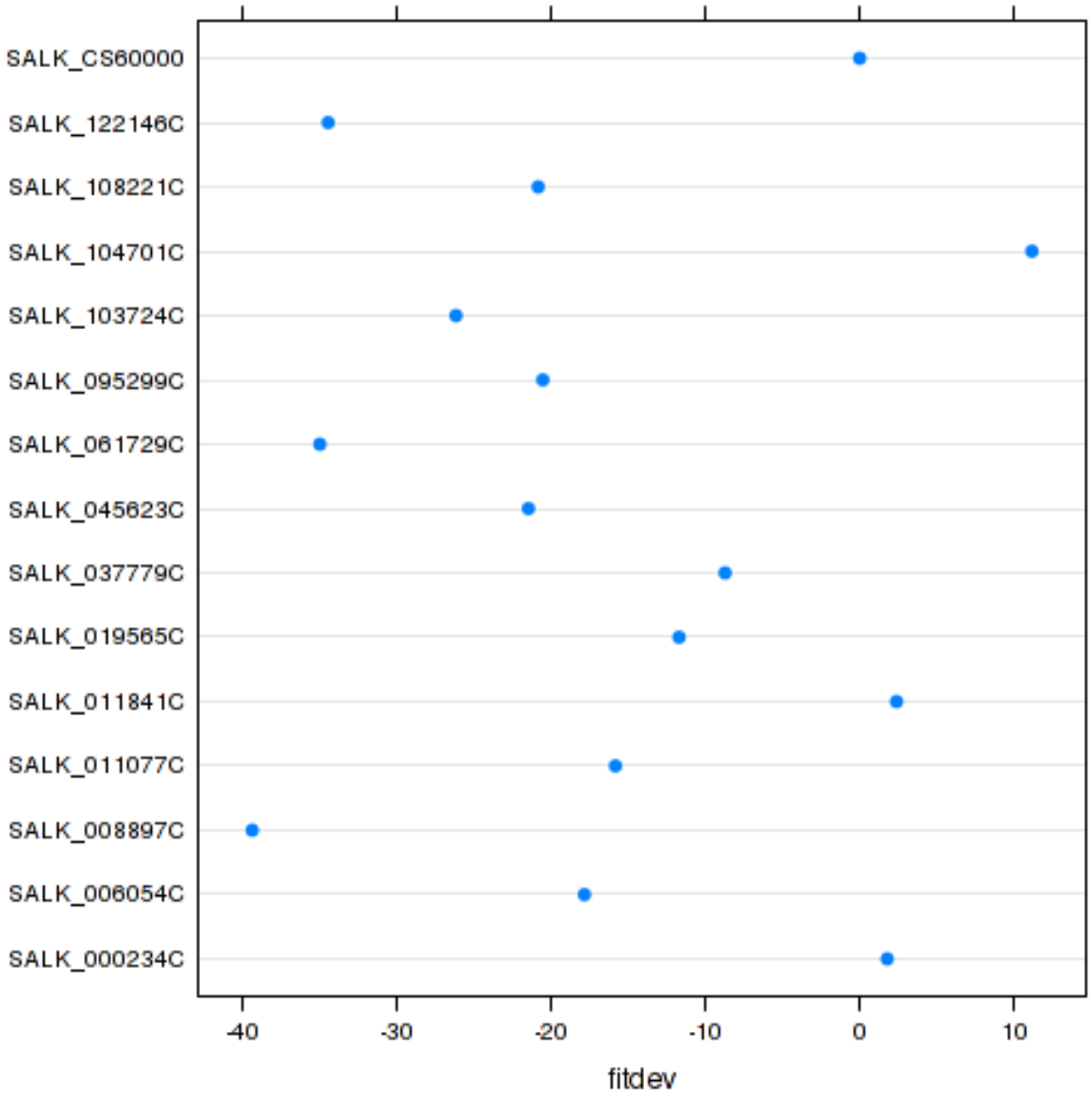
```
##
##  Welch Two Sample t-test
##
## data:  fitdev by GeneFamilyName
## t = 3.263, df = 224.06, p-value = 0.001275
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.930097 19.964495
## sample estimates:
##                                 mean in group Control
##                                          -1.935837e-15
## mean in group MADS-box Transcription Factor Family
##                                          -1.244730e+01
```

Here is the distribution of fitness effects among mads box genes

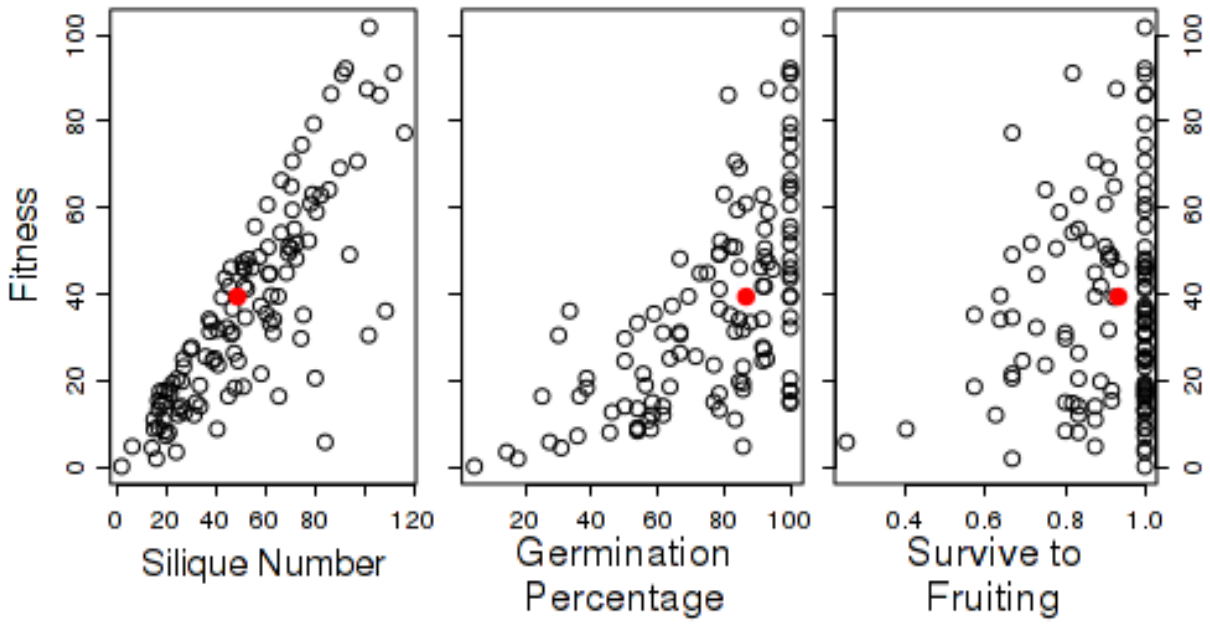
```
## Warning in (function (x, y, box.ratio = 1, box.width = box.ratio/(1 + box.ratio),
: NAs introduced by coercion
```




```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## SALK_Line    14  26332  1880.9   2.288 0.00616 **
## Residuals   214 175937   822.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 72 observations deleted due to missingness
```



2.1 Fitness components



3 Analysis of fitness

Taking all SALK lines and pooling them and comparing to the control:

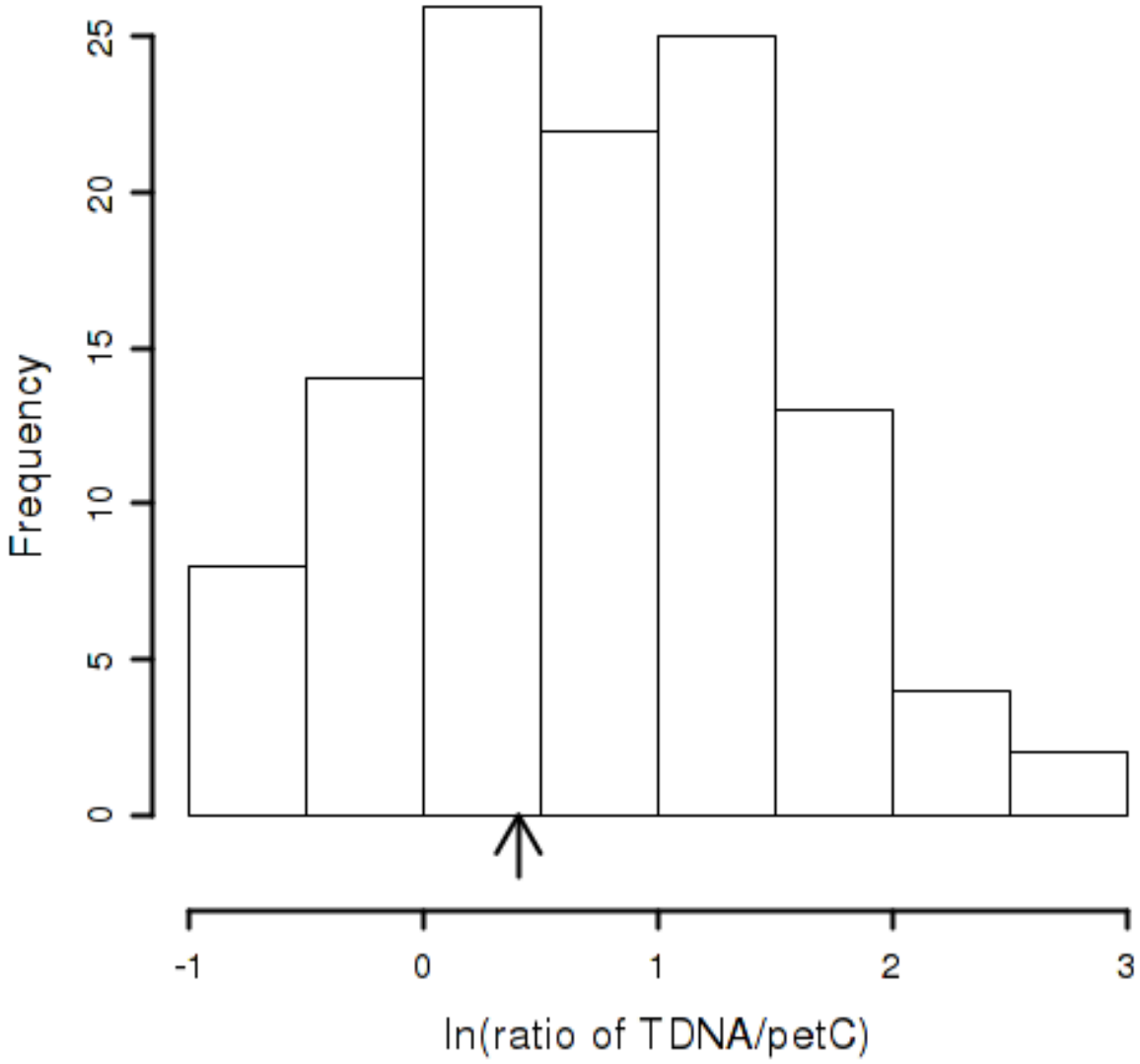
You can see from the resampled line distributions that salk lines and controls have similar mean fitnesses, with definite differences in variance among groups

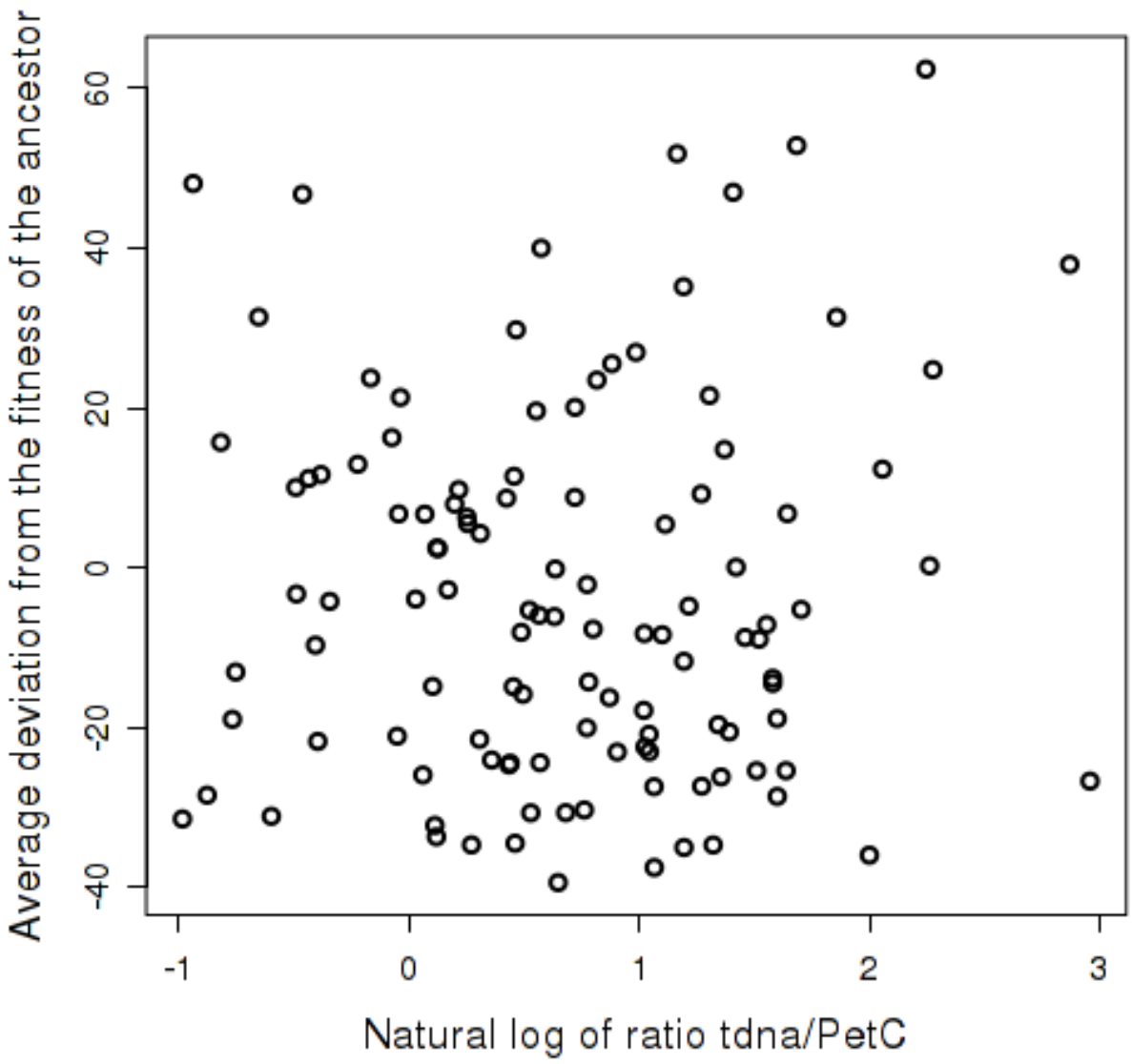
```
##
##  Welch Two Sample t-test
##
## data:  fitness by treat
## t = -0.63525, df = 155.85, p-value = 0.5262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.141375  4.179152
## sample estimates:
## mean in group control   mean in group treat
##              39.39545              41.37657
##
##  Kruskal-Wallis rank sum test
##
## data:  fitmerg$treat and fitmerg$fitness
## Kruskal-Wallis chi-squared = 1106.5, df = 763, p-value = 4.391e-15
```

```
## [1] "permutation typeI prob:"
## [1] 0.742
```

4 Effects of tdna insert number

Here is a plot that relates our total measure of fitness in the lines used in the original pilot study to the ratio of tdna to endogenous genes






```
Call:
lm(formula = fitdev ~ log(area.ratio), data = fruit.by.line)

Residuals:
    Min      1Q  Median      3Q     Max
-35.920 -19.680  -3.513  14.541  64.439

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)      -4.0472     2.8887  -1.401    0.164
log(area.ratio)   0.8829     2.6995   0.327    0.744

Residual standard error: 23.52 on 112 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.0009542,Adjusted R-squared:  -0.007966
F-statistic: 0.107 on 1 and 112 DF,  p-value: 0.7442
```

Clearly no pattern there and a regression confirms.

Just to make sure, I also lumped the ratios into categories and looked for a pattern again:

When focusing on medians, it looks a little bit like there might be some variation across categories

```
summary(aov(fitdev~as.factor(ratiocat),fruit.by.line))

                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(ratiocat)   2    187    93.3   0.168  0.846
Residuals           111  61805   556.8
7 observations deleted due to missingness
```

Alas, no pattern there either. I might suggest that, at the least, we worry about copy number less than other factors when choosing lines for UnPAK projects

5 Fitness as a function of gene family size

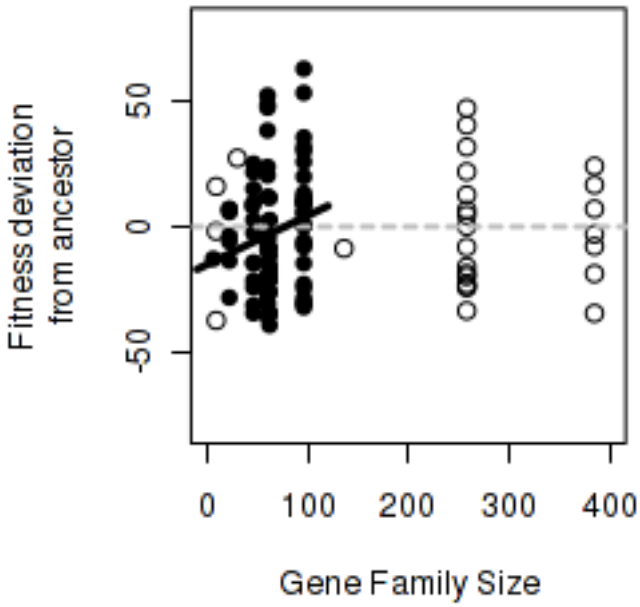
In the original pilot study we chose genes from different families with differing sizes. Again, there does not seem to be a significant relationship between gene family size and reproductive output in the non-regulatory genes, but there does seem to be a slight pattern in the regulatory genes.

This figure is based on the number of genes in the Gene Family data on tair

This is the same figure but with the means of each line plotted instead of all the reps for each line.

```
Call:
lm(formula = fitdev ~ FamilySize * Regulatory, data = fruit.by.line)

Residuals:
   Min     1Q Median     3Q    Max
-37.50 -17.46  -1.68  15.80  60.80

Coefficients:
```

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.031853  10.582719   0.003    0.998
FamilySize                -0.004502   0.038440  -0.117    0.907
Regulatoryyes            -16.152593  12.906961  -1.251    0.213
FamilySize:Regulatoryyes   0.188779   0.112813   1.673    0.097 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.2 on 112 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.02922,Adjusted R-squared:  0.003212
F-statistic: 1.124 on 3 and 112 DF,  p-value: 0.3427

Call:
lm(formula = fitdev ~ FamilySize, data = fruit.by.line, subset = Regulatory ==
    "yes")

Residuals:
    Min      1Q  Median      3Q     Max
-34.700 -16.930  -1.639  15.799  60.804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.1207     7.3637  -2.189   0.0313 *
FamilySize    0.1843     0.1057   1.743   0.0848 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.12 on 86 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.03414,Adjusted R-squared:  0.0229
F-statistic: 3.039 on 1 and 86 DF,  p-value: 0.08484
```



And some analyses on the means of each line's fitness deviation from the ancestor: First OLS ancova. Then regressions for non-regulatory and then regulatory genes

```
summary(aov(fitdev~FamilySize*Regulatory,fruit.by.line))

                      Df Sum Sq Mean Sq F value Pr(>F)
FamilySize             1    306   306.3   0.569  0.452
Regulatory             1      1     0.7   0.001  0.972
FamilySize:Regulatory  1   1507  1507.0   2.800  0.097 .
Residuals            112  60277   538.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
5 observations deleted due to missingness

summary(lm(fitdev~FamilySize,subset=Regulatory=="no",fruit.by.line))


Call:
lm(formula = fitdev ~ FamilySize, data = fruit.by.line, subset = Regulatory ==
    "no")

Residuals:
    Min      1Q  Median      3Q     Max
-37.504 -18.043  -1.843  16.262  47.859
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.031853  10.701052   0.003    0.998
FamilySize  -0.004502   0.038869  -0.116    0.909

Residual standard error: 23.46 on 26 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.0005156,Adjusted R-squared:  -0.03793
F-statistic: 0.01341 on 1 and 26 DF,  p-value: 0.9087
```

```
summary(lm(fitdev~FamilySize,subset=Regulatory=="yes",fruit.by.line))
```

```

Call:
lm(formula = fitdev ~ FamilySize, data = fruit.by.line, subset = Regulatory ==
    "yes")

Residuals:
    Min      1Q  Median      3Q     Max
-34.700 -16.930  -1.639  15.799  60.804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.1207     7.3637  -2.189   0.0313 *
FamilySize    0.1843     0.1057   1.743   0.0848 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.12 on 86 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.03414,Adjusted R-squared:  0.0229
F-statistic: 3.039 on 1 and 86 DF,  p-value: 0.08484
```

```
summary(lm(log(fitdev+40)~FamilySize,subset=Regulatory=="yes",fruit.by.line))
```

```

Call:
lm(formula = log(fitdev + 40) ~ FamilySize, data = fruit.by.line,
    subset = Regulatory == "yes")

Residuals:
    Min      1Q  Median      3Q     Max
-3.7990 -0.5099  0.2162  0.6402  1.2325
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.02358    0.27102  11.156   <2e-16 ***
FamilySize   0.00439    0.00389   1.128    0.262
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8509 on 86 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.01459,Adjusted R-squared:  0.00313
F-statistic: 1.273 on 1 and 86 DF,  p-value: 0.2623
```

Here is a little more sophisticated analysis of the family size effect

```
library(nlme)

##
## Attaching package:  'nlme'
##
## The following object is masked from 'package:lme4':
##
##    lmList

fm1 <- lme(fixed=fitdev~FamilySize,random=~1|SALK_Line,method="ML",subset=Regulatory=="y
#plot(fm1)
vfix <- varFixed(~FamilySize)
fm2 <- lme(fixed=fitdev~FamilySize,random=~1|SALK_Line,method="ML",subset=Regulatory=="y
#plot(resid(fm2,type="pearson")~fitted(fm2))
fm3 <- lme(fixed=fitdev~1,random=~1|SALK_Line,method="ML",subset=Regulatory=="yes",na.ac

anova(fm3,fm1,fm2) #looks like fm1 is the best model

##     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## fm3     1  3 7933.316 7947.407 -3963.658
## fm1     2  4 7931.925 7950.713 -3961.962 1 vs 2 3.390619  0.0656
## fm2     3  4 7985.516 8004.304 -3988.758

anova(fm1)

##             numDF denDF  F-value p-value
## (Intercept)     1   722 1.105524  0.2934
## FamilySize      1    86 3.460323  0.0663
```

```
anova(fm2)

##             numDF denDF  F-value p-value
## (Intercept)     1   722 1.276404  0.2589
## FamilySize      1    86 3.534842  0.0635

anova(fm3)

##             numDF denDF  F-value p-value
## (Intercept)     1   722 1.100184  0.2946
```

Ok, here is the analysis of family size with line as random intercept. Terrible heteroscedacity, repaired using a fixed variance structure. No signal of family size in the final model.

Now, the joint categories approach: Two tests. The first assumes that the rows and columns are independent, but the expected values come from the marginal totals. The second assumes that the number of fitneses in each of the four categories is equal.

```
Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 2.3999, df = 1, p-value = 0.1213
[1] 6.025056e-07
```

Here is a test of the change in variance through time:

```
brks=c(0,seq(10,150,10),166)
family.size.cat <- cut(fitmerg$FamilySize,breaks=brks)
sds <- with(fitmerg,tapply(fitdev,family.size.cat,sd))
brkmid <- (brks+(c(brks[-1],166)-brks)/2)[-length(sds)]
bartlett.test(fitmerg$fitdev,family.size.cat)

##
##  Bartlett test of homogeneity of variances
##
## data:  fitmerg$fitdev and family.size.cat
## Bartlett's K-squared = 38.058, df = 6, p-value = 1.095e-06

plot(sds~brkmid)
summary(lm(sds~brkmid))

##
```

```
## Call:
## lm(formula = sds ~ brkmid)
##
## Residuals:
##    (0,10]   (20,30]   (40,50]   (50,60]   (60,70]  (90,100] (130,140]
##    -7.426    -5.119     2.664    11.247    -2.590    14.254   -13.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.7467     7.4525   5.199  0.00347 **
## brkmid       -0.1638     0.1023  -1.602  0.17017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 5 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.339,Adjusted R-squared:  0.2069
## F-statistic: 2.565 on 1 and 5 DF,  p-value: 0.1702
```

So not much change in variance, though not a lot of power either.

```
Welch Two Sample t-test

data:  fitdev by Regulatory
t = 1.2283, df = 398.63, p-value = 0.2201
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.900829  8.231144
sample estimates:
 mean in group no mean in group yes
        4.215966          1.050808
```

6 Multiple environment experiment

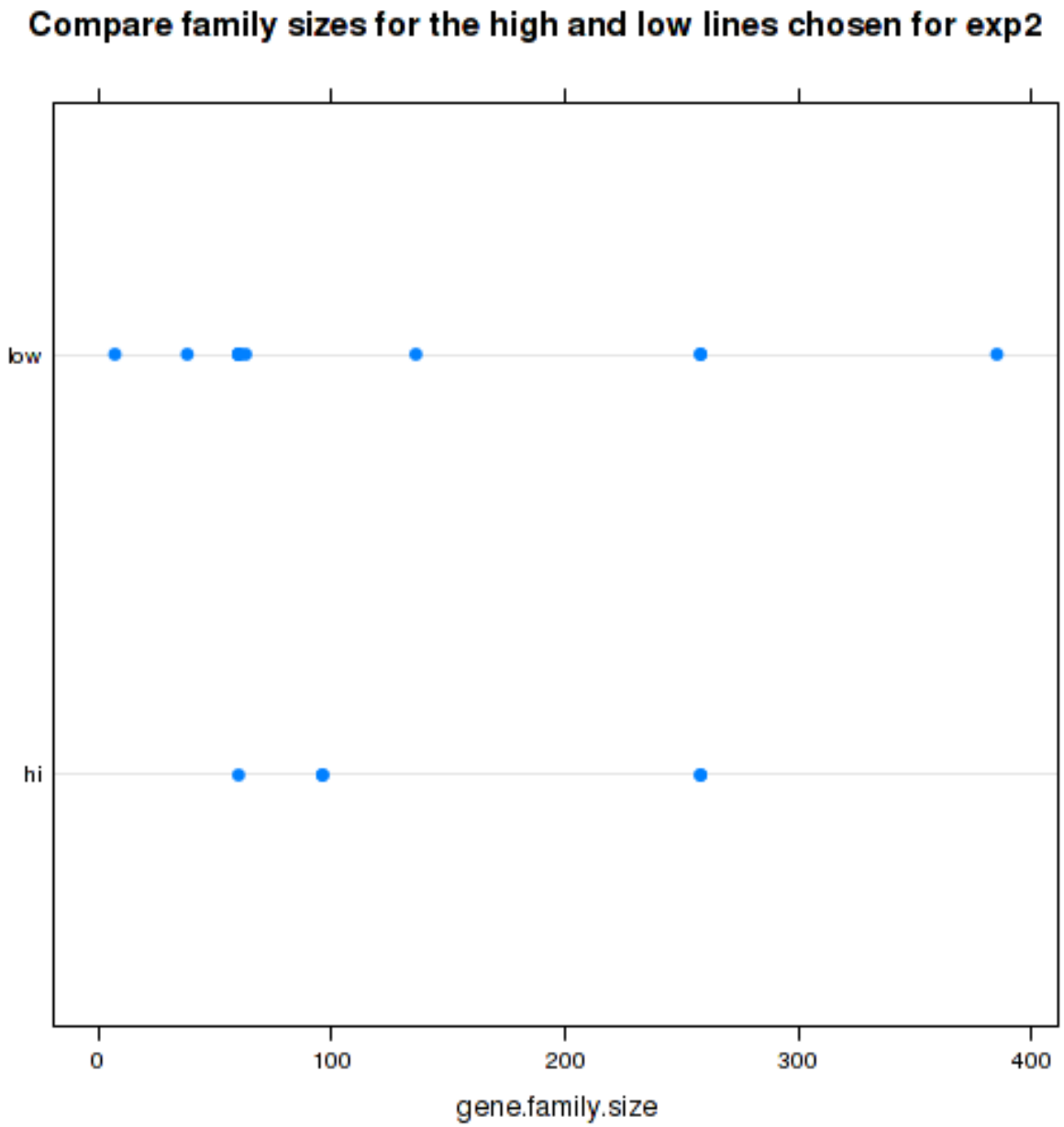
6.1 Figures for the 1st multi-environment experiment

This figure is fitness deviation ignoring germination (we don't have per-sowed seed estimates (in other words, replicated) of germination for the first experiment, just average germination for that line for that germination effort. We do have those data for the second three treatments/time points

```
Note: no visible binding for global variable 'SALK_Line'
[1] "SALK_017933C"
[1] "SALK_033462C"
[1] "SALK_038957C"
[1] "SALK_042704C"
[1] "SALK_050488C"
[1] "SALK_054680C"
[1] "SALK_059835C"
[1] "SALK_063722C"
[1] "SALK_094332C"
[1] "SALK_126600C"
[1] "SALK_134535C"
[1] "SALK_150522C"
[1] "CS60000"
```

I'm going to try and address courtney's question about the gene family size of high and low lines

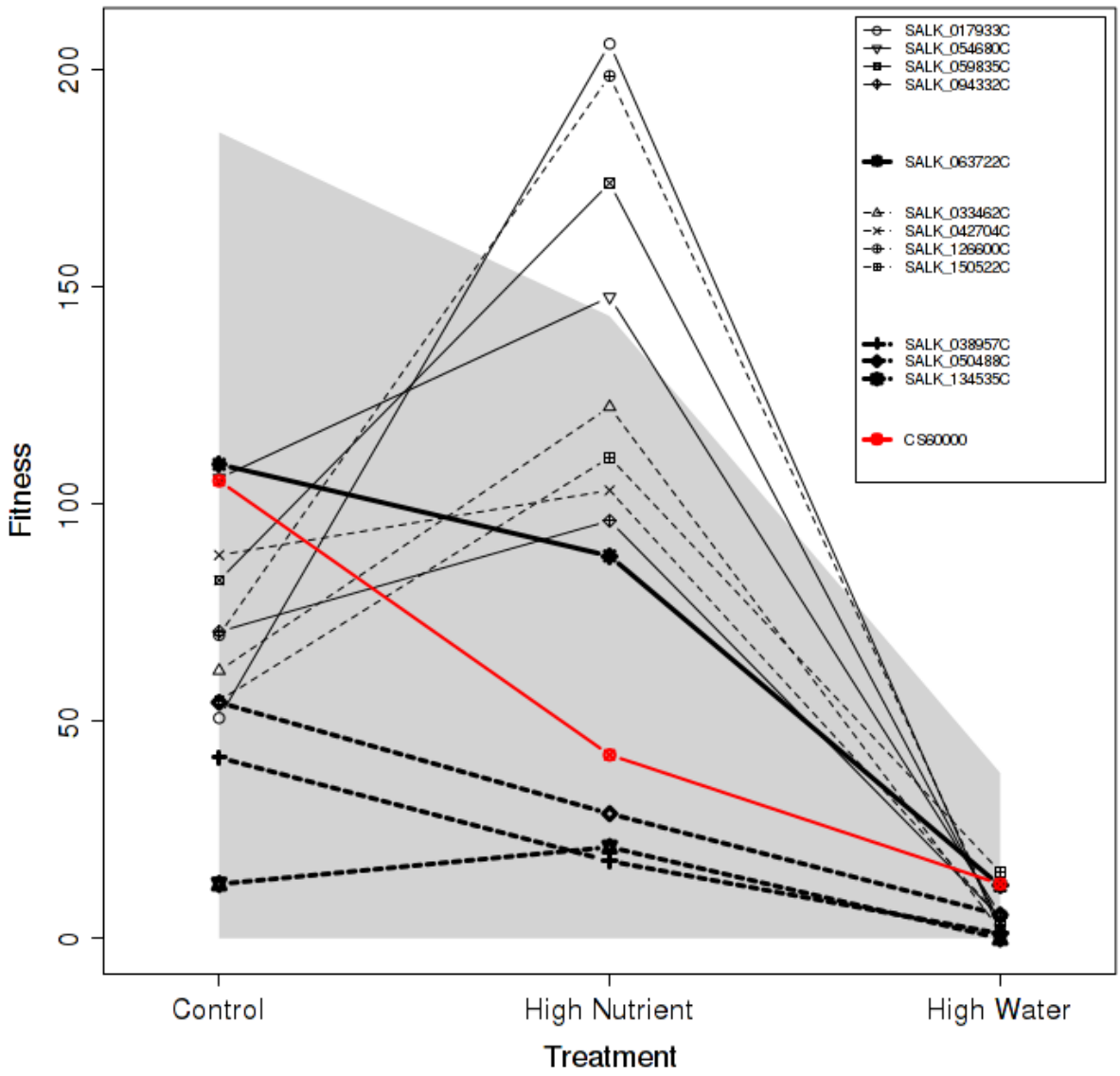
```
## [1] "SALK_Line"        "fitdev"           "gene.family.size"
## [4] "lohi"
```



Now here is the figure with straight mean fruit number per line:

```
Note: no visible binding for global variable 'SALK_Line'
Note: no visible binding for global variable 'treattype'
[1] "SALK_017933C"
[1] 1
[1] "SALK_033462C"
[1] 1
[1] "SALK_038957C"
[1] 1
```

```
[1] "SALK_042704C"
[1] 1
[1] "SALK_050488C"
[1] 1
[1] "SALK_054680C"
[1] 1
[1] "SALK_059835C"
[1] 1
[1] "SALK_063722C"
[1] 1
[1] "SALK_094332C"
[1] 1
[1] "SALK_126600C"
[1] 1
[1] "SALK_134535C"
[1] 1
[1] "SALK_150522C"
[1] 1
[1] "CS60000"
[1] 2
```



6.1.1 Copy number

In the following figure line width is proportional to copy number category

6.2 Various tests of GxE

```
#MixedEffects
fit1 <- lmer(fitdevng~1+(1|SALK_Line),subset=SALK_Line!="CS60000",data=intresults)
```

```
Note: no visible binding for global variable 'x'
Note: no visible binding for global variable 'x'
Note: no visible binding for global variable '.xData'
Note: no visible binding for global variable '.->Lambdat'
Note: no visible binding for global variable '.->LamtUt'
Note: no visible binding for global variable '.->Lind'
Note: no visible binding for global variable '.->Ptr'
Note: no visible binding for global variable '.->RZX'
Note: no visible binding for global variable '.->Ut'
Note: no visible binding for global variable '.->Utr'
Note: no visible binding for global variable '.->V'
Note: no visible binding for global variable '.->VtV'
Note: no visible binding for global variable '.->Vtr'
Note: no visible binding for global variable '.->X'
Note: no visible binding for global variable '.->Xwts'
Note: no visible binding for global variable '.->Zt'
Note: no visible binding for global variable '.->beta0'
Note: no visible binding for global variable '.->delb'
Note: no visible binding for global variable '.->delu'
Note: no visible binding for global variable '.->theta'
Note: no visible binding for global variable '.->u0'
Note: no visible binding for '<<-' assignment to 'RZX'
Note: no visible binding for '<<-' assignment to 'Utr'
Note: no visible binding for '<<-' assignment to 'V'
Note: no visible binding for '<<-' assignment to 'VtV'
Note: no visible binding for '<<-' assignment to 'Vtr'
Note: no visible binding for '<<-' assignment to 'beta0'
Note: no visible binding for '<<-' assignment to 'delb'
Note: no visible binding for '<<-' assignment to 'delu'
Note: no visible binding for '<<-' assignment to 'u0'
Note: no visible binding for '<<-' assignment to 'Ut'
Note: no visible binding for global variable 'Ut'
Note: no visible binding for '<<-' assignment to 'LamtUt'
Note: no visible binding for '<<-' assignment to 'Xwts'
Note: no visible global function definition for 'initializePtr'
Note: no visible binding for '<<-' assignment to 'Ptr'
Note: no visible binding for global variable 'X'
Note: no visible binding for global variable 'Lambdat'
Note: no visible binding for global variable 'LamtUt'
Note: no visible binding for global variable 'Lind'
Note: no visible binding for global variable 'RZX'
Note: no visible binding for global variable 'Ut'
```

```
Note: no visible binding for global variable 'Utr'
Note: no visible binding for global variable 'V'
Note: no visible binding for global variable 'VtV'
Note: no visible binding for global variable 'Vtr'
Note: no visible binding for global variable 'Xwts'
Note: no visible binding for global variable 'Zt'
Note: no visible binding for global variable 'beta0'
Note: no visible binding for global variable 'delb'
Note: no visible binding for global variable 'delu'
Note: no visible binding for global variable 'theta'
Note: no visible binding for global variable 'u0'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable 'theta'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable 'Xwts'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable '.->Ptr'
Note: no visible binding for global variable '.->mu'
Note: no visible binding for global variable '.->offset'
Note: no visible binding for global variable '.->sqrtXwt'
Note: no visible binding for global variable '.->sqrtrwt'
Note: no visible binding for global variable '.->weights'
Note: no visible binding for global variable '.->wtres'
Note: no visible binding for global variable '.->y'
Note: no visible binding for global variable '.->REML'
Note: no visible binding for '<<-' assignment to 'REML'
Note: no visible binding for global variable 'REML'
Note: no visible binding for '<<-' assignment to 'REML'
Note: no visible global function definition for 'callSuper'
Note: no visible binding for '<<-' assignment to 'mu'
Note: no visible binding for '<<-' assignment to 'sqrtXwt'
Note: no visible binding for '<<-' assignment to 'sqrtrwt'
Note: no visible binding for '<<-' assignment to 'wtres'
Note: no visible binding for global variable 'sqrtrwt'
Note: no visible binding for global variable 'mu'
Note: no visible global function definition for 'callSuper'
Note: no visible binding for global variable 'u0'
Note: no visible binding for global variable 'beta0'
Note: no visible binding for global variable 'u0'
Note: no visible global function definition for 'ptr'
Note: no visible binding for global variable 'theta'
Note: no visible binding for global variable 'Ptr'
```

```
Note: no visible global function definition for 'initializePtr'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable 'Ptr'
Note: no visible global function definition for 'initializePtr'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for '<<-' assignment to 'Ptr'
Note: no visible binding for global variable 'mu'
Note: no visible binding for global variable 'sqrtXwt'
Note: no visible binding for global variable 'sqrtrwt'
Note: no visible binding for global variable 'wtres'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable 'mu'
Note: no visible binding for global variable 'Ptr'
Note: no visible binding for global variable 'REML'
Note: no visible global function definition for 'ptr'
Note: no visible global function definition for 'ptr'
Note: no visible global function definition for 'ptr'
Note: no visible global function definition for 'ptr'
Note: no visible global function definition for 'ptr'
Note: no visible global function definition for 'ptr'
```

```
fit2 <- lmer(fitdevng~treattype+(1|SALK_Line),subset=SALK_Line!="CS60000",data=intresult
```

```
Note: no visible global function definition for 'callSuper'
```

```
fit3 <- lmer(fitdevng~treattype+treattype:SALK_Line+(1|SALK_Line),subset=SALK_Line!="CS6
```

```
Note: no visible global function definition for 'callSuper'
```

```
anova(fit1,fit2,fit3)
```

```
refitting model(s) with ML (instead of REML)
```

```
Note: no visible global function definition for 'callSuper'
Note: no visible binding for global variable '.refClassDef'
Note: no visible global function definition for 'callSuper'
Note: no visible global function definition for 'callSuper'
Data: intresults
Subset: SALK_Line != "CS60000"
Models:
fit1: fitdevng ~ 1 + (1 | SALK_Line)
fit2: fitdevng ~ treattype + (1 | SALK_Line)
fit3: fitdevng ~ treattype + treattype:SALK_Line + (1 | SALK_Line)
     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
```

```
fit1  3 4643.4 4655.4 -2318.7   4637.4
fit2  6 4589.9 4613.9 -2288.9   4577.9 59.548      3  7.343e-13 ***
fit3 50 4578.4 4778.1 -2239.2   4478.4 99.466     44  3.555e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fit using OLS
fitaov1 <- aov(fitdevng~treattype*SALK_Line,subset=SALK_Line!="CS60000",data=intresults)
summary(fitaov1)
```

```
                     Df  Sum Sq Mean Sq F value   Pr(>F)
treattype             3  350733  116911  24.814 1.37e-14 ***
SALK_Line            11  191431   17403   3.694 5.42e-05 ***
treattype:SALK_Line  33  329706    9991   2.121 0.000476 ***
Residuals           353 1663151    4711
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
63 observations deleted due to missingness
```

```
#fit using ML
fitaov1.glm <-glm(fitdevng~treattype*SALK_Line,subset=SALK_Line!="CS60000",data=intresul
anova(fitaov1.glm,test="Chisq")
```

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: fitdevng

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  400    2535021
treattype            3   350733       397    2184288 4.771e-16 ***
SALK_Line           11   191431       386    1992857 2.789e-05 ***
treattype:SALK_Line 33   329706       353    1663151 0.0001816 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#
#now look in each environment and ask if there are line differences
#
```

```
for (trt in unique(intresults$treattype))
  {
    fit.line <- glm(fitdevng~1+as.factor(SALK_Line),subset=SALK_Line!="CS60000",data=int
    fit.intercept <- glm(fitdevng~1,subset=SALK_Line!="CS60000",data=intresults[intresul
    cat(rep("-",30));cat("\n")
    print (paste("Effect of including line in environment: ",trt))
    print(anova(fit.intercept,fit.line,test="Chisq"))
    cat(rep("-",30));cat("\n")
  }
```

```
Note: no visible binding for global variable 'SALK_Line'
Note: no visible binding for global variable 'SALK_Line'
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
[1] "Effect of including line in environment:  nutrient"
Analysis of Deviance Table

Model 1: fitdevng ~ 1
Model 2: fitdevng ~ 1 + as.factor(SALK_Line)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        98    1417594
2        87    1125616 11   291978  0.02033 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
[1] "Effect of including line in environment:  control"
Analysis of Deviance Table

Model 1: fitdevng ~ 1
Model 2: fitdevng ~ 1 + as.factor(SALK_Line)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       101     532531
2        90     439049 11    93482  0.05823 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
[1] "Effect of including line in environment:  highwater"
Analysis of Deviance Table

Model 1: fitdevng ~ 1
Model 2: fitdevng ~ 1 + as.factor(SALK_Line)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
1        95      14671
2        84      11170 11   3500.2 0.005812 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
[1] "Effect of including line in environment:  FIRST.EXP"
Analysis of Deviance Table

Model 1: fitdevng ~ 1
Model 2: fitdevng ~ 1 + as.factor(SALK_Line)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       103     219492
2        92      87315 11   132177 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Here is an analysis with Prior information and ecotypes included

```
results$genotype.cls = rep("control",dim(results)[1])
results$genotype.cls[results$fitdev<0] = "low"
results$genotype.cls[results$fitdev>0] = "high"
results$genotype.cls[results$SALK_Line=="CS60000"] = "control"
results$genotype.cls[grep("ECO",results$SALK_Line)] = "ecotype"

fit1 <- lm(fitness~treattype*genotype.cls,subset=treattype!="FIRST.EXP",data=results)
Anova(fit1,contrasts = list(treattype=contr.sum,genotype.cls=contr.sum),type=3)

Anova Table (Type III tests)

Response: fitness
                      Sum Sq  Df F value    Pr(>F)
(Intercept)           110881   1  49.774 7.596e-12 ***
treattype              44786   2  10.052 5.497e-05 ***
genotype.cls          490361   3  73.373 < 2.2e-16 ***
treattype:genotype.cls 330825   6  24.751 < 2.2e-16 ***
Residuals             895536 402
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fit2 <- lm(fitness~treattype*genotype.cls,subset=(treattype!="FIRST.EXP")&(genotype.cls!
Anova(fit2,contrasts = list(treattype=contr.sum,genotype.cls=contr.sum),type=3)
```

```
Anova Table (Type III tests)

Response: fitness
                      Sum Sq  Df F value    Pr(>F)
(Intercept)           110881   1  54.756 9.043e-13 ***
treattype              44786   2  11.058 2.157e-05 ***
genotype.cls          490177   2 121.031 < 2.2e-16 ***
treattype:genotype.cls 328440   4  40.548 < 2.2e-16 ***
Residuals             759380 375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis just looks for GxE and main effects in the ecotypes. Not much signal Ecotype lines only

```
fitaov.genotype.cls.nomut <- aov(fitness~treattype*SALK_Line,subset=(treattype!="FIRST.E
summary(fitaov.genotype.cls.nomut)

##                      Df Sum Sq Mean Sq F value Pr(>F)
## treattype             2  20455   10228   1.714  0.234
## SALK_Line             9  23496    2611   0.438  0.883
## treattype:SALK_Line   9  58971    6552   1.098  0.446
## Residuals             9  53689    5965
```

```
fitall <- glm(fitdev~treattype*SALK_Line,subset=treattype!="FIRST.EXP",data=intresults)
(anova(fitall,test="F"))

## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: fitdev
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
## NULL                                   383    3103317
## treattype             2   647029       381    2456288 59.5646 < 2.2e-16 ***
## SALK_Line            12   234750       369    2221538  3.6018 4.360e-05 ***
## treattype:SALK_Line  24   347732       345    1873806  2.6676 5.317e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#add in a classifier for early experiment performance
firstmns <- with(intresults[intresults$treattype=="FIRST.EXP",c("fitdev","SALK_Line")],
                 aggregate(cbind(first.fitdev=fitdev),by=list(SALK_Line=SALK_Line),mean,
intresults <- merge(intresults,firstmns,all.x=T)

fitlo <-  glm(fitdev~treattype*SALK_Line,subset=((treattype!="FIRST.EXP")&(first.fitdev<
(anova(fitlo,test="F"))
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: fitdev
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
## NULL                                  236    1805730
## treattype            2   278944       234    1526786 25.0251 1.731e-10 ***
## SALK_Line            7   146974       227    1379812  3.7673 0.0007137 ***
## treattype:SALK_Line 14   192704       213    1187108  2.4697 0.0029519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fithi <-  glm(fitdev~treattype*SALK_Line,subset=((treattype!="FIRST.EXP")&(first.fitdev>
(anova(fithi,test="F"))
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: fitdev
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
## NULL                                  146    1232611
## treattype            2   428976       144     803635 41.2297 1.228e-14 ***
## SALK_Line            4    18145       140     785490  0.8720   0.48267
## treattype:SALK_Line  8    98791       132     686698  2.3738   0.02021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
intresults$lohi <- factor(ifelse(intresults$first.fitdev<=0,"low","high"))
fitlowhi <- glm(fitdev~treattype*lohi,subset=(treattype!="FIRST.EXP"),data=intresults)
(anova(fitlowhi,test="F"))

## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: fitdev
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
## NULL                             383    3103317
## treattype       2   647029       381    2456288 52.4749 < 2.2e-16 ***
## lohi            1    68562       380    2387726 11.1209 0.0009381 ***
## treattype:lohi  2    57305       378    2330421  4.6475 0.0101397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova(lm(fitdev~treattype*lohi,subset=(treattype!="FIRST.EXP"),data=intresults),contrast

## Anova Table (Type III tests)
##
## Response: fitdev
##                 Sum Sq  Df F value    Pr(>F)
## (Intercept)      23242   1  3.7699   0.05293 .
## treattype       428976   2 34.7905 1.353e-14 ***
## lohi             15911   1  2.5808   0.10900
## treattype:lohi   57305   2  4.6475   0.01014 *
## Residuals      2330421 378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fiteco <-  glm(fitdev~treattype*SALK_Line,data=intresults.ecotypes)
(anova(fiteco,test="F"))

## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: fitdev
```

```
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev      F Pr(>F)
## NULL                                    29     154737
## treattype            2    18581        27     136156 1.5574 0.2625
## SALK_Line            9    23496        18     112660 0.4376 0.8829
## treattype:SALK_Line  9    58971         9      53689 1.0984 0.4456
```

6.3 Tables for the MS

6.3.1 Gene families

```
famtable <- unique(fitmerg[!is.na(fitmerg$Regulatory),c("GeneFamilyName","FamilySize","R
linesfromfams <- with(unique(fitmerg[fitmerg$GeneFamilyName!="Control",c("SALK_Line","Ge
famtable <- merge(famtable,linesfromfams,all.x=T)
famtable <- famtable[order(famtable$Regulatory,-famtable$FamilySize),]
famtable$Function <- ifelse(famtable$Regulatory=="yes","Regulatory","Metabolic")
famtable <- famtable[,-which(names(famtable)=="Regulatory")]

require(xtable)

## Loading required package:  xtable
##
## Attaching package:  'xtable'
##
## The following objects are masked from 'package:Hmisc':
##
##    label, label<-

print(file="redundancy-tables-figs/tbl1-genefams.html",xtable(famtable),type="html",
      include.rownames=F)

## Warning in file(file, ifelse(append, "a", "w")):  cannot open file 'redundancy-tables
No such file or directory
## Error in file(file, ifelse(append, "a", "w")):  cannot open the connection
```

6.3.2 SALK Line list

```
salktbl <- unique(fitmerg[,c("SALK_Line","Gene","Gene_Family")])
salktbl <- salktbl[complete.cases(salktbl),]
salktbl <- salktbl[order(salktbl$Gene_Family,salktbl$SALK_Line),]
print(file="redundancy-tables-figs/tbls1-salk-lines.html",xtable(salktbl),type="html",
      include.rownames=F)

## Warning in file(file, ifelse(append, "a", "w")):  cannot open file 'redundancy-tables
No such file or directory
## Error in file(file, ifelse(append, "a", "w")):  cannot open the connection
```

6.4 Test for effect of line on fitness for first exp.

```
summary(aov(log(fitness+1)~SALK_Line,data=fitmerg[grep("SALK.[0-9]+C",fitmerg$SALK_Line)

##              Df Sum Sq Mean Sq F value Pr(>F)
## SALK_Line   115  604.2   5.254   2.991 <2e-16 ***
## Residuals   928 1630.4   1.757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7 Naturally occuring variants

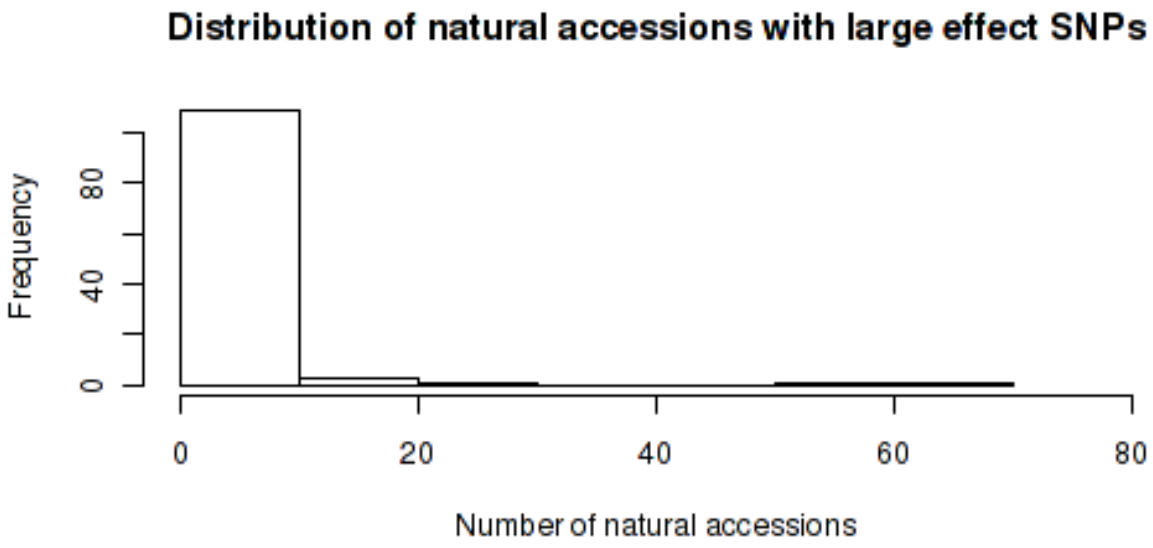
We took the data from Cao et al 2010 and determined how many of the lines in this experiment also showed some sort of natural variation in gene function

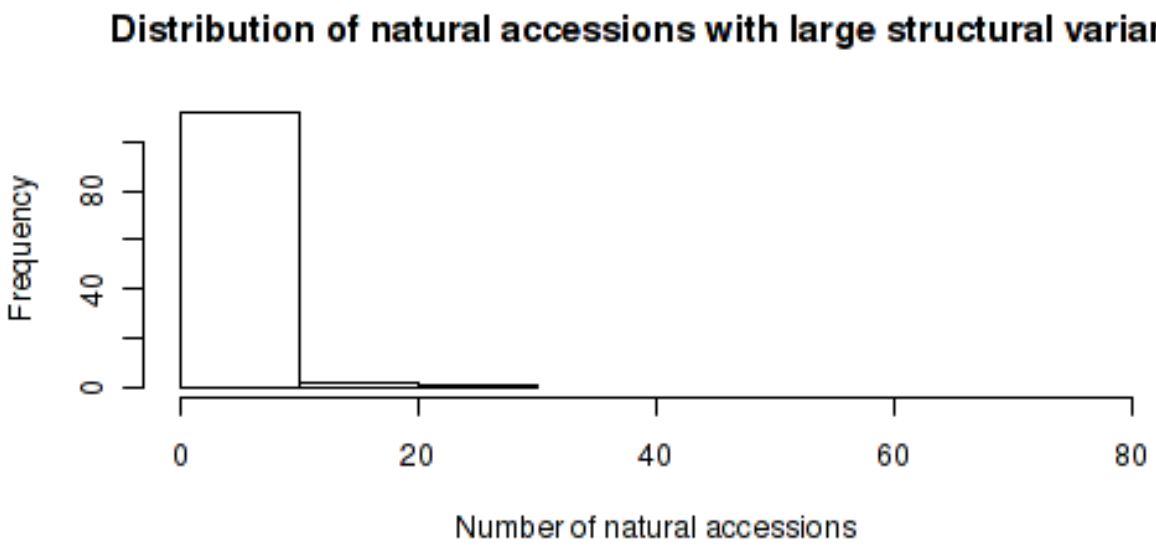
```
suppressMessages(require(dplyr))
snp <- unique(read.csv(paste0(csvdir,"/phen-snp.csv"))[,1:2])
names(snp)[2] <- "snp.strains"
snp <- snp %>% group_by(Accession) %>% summarise(snp.strains.mn=sum(snp.strains))

## Note: no visible binding for global variable 'Accession'
## Note: no visible binding for global variable 'snp.strains'

sv <- unique(read.csv(paste0(csvdir,"/phen-sv.csv"))[,1:2])
names(sv)[2] <- "sv.strains"
sv <- sv %>% group_by(Accession) %>% summarise(sv.strains.mn=sum(sv.strains))

## Note: no visible binding for global variable 'Accession'
## Note: no visible binding for global variable 'sv.strains'

write.table(file="cao-digested.csv",sep=",",row.names=F,unique(merge(snp,sv)))
```

There are definitely lines that are knocked out in nature. The first table is the frequency of lines with no natural variants (false) versus variants for SNPs that should drastically alter gene function. The second is for the distribution of lines with large structural variants

```
with(unique(snp),table(snp.strains.mn>0))

##
## FALSE  TRUE
##    99    16

with(unique(sv),table(sv.strains.mn>0))

##
## FALSE  TRUE
##   106     9
```

The following figure illustrates the distribution of the naturally occurring variants in our





collection of lines.