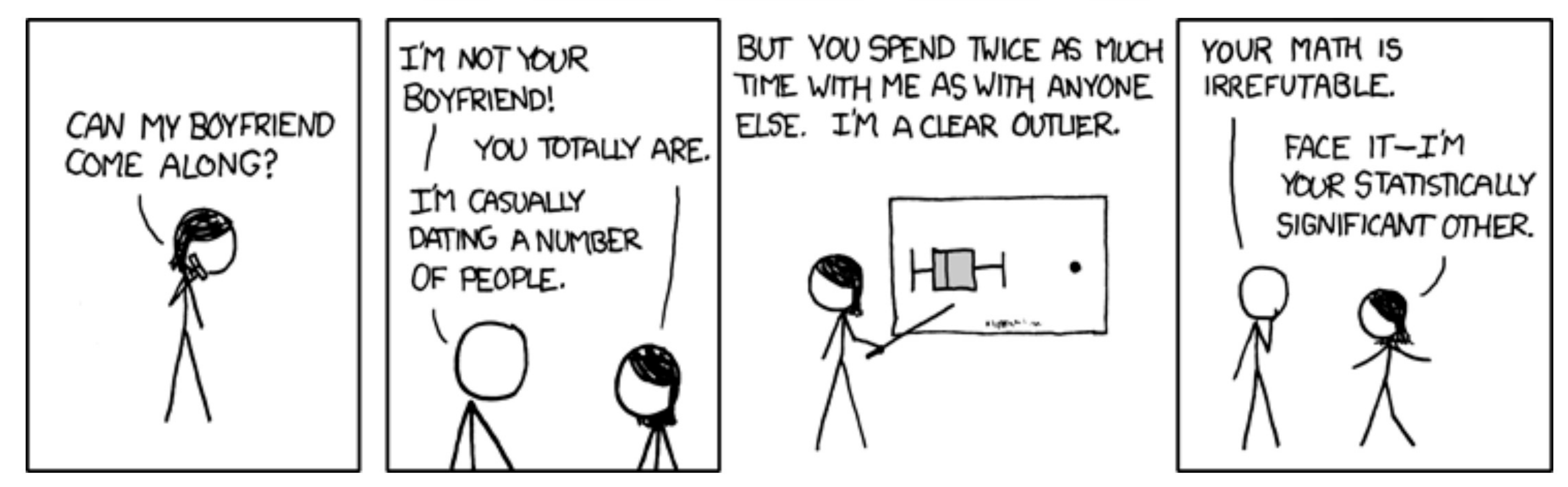
# Binary Response Variables, Random vs Fixed Effects, and Outliers

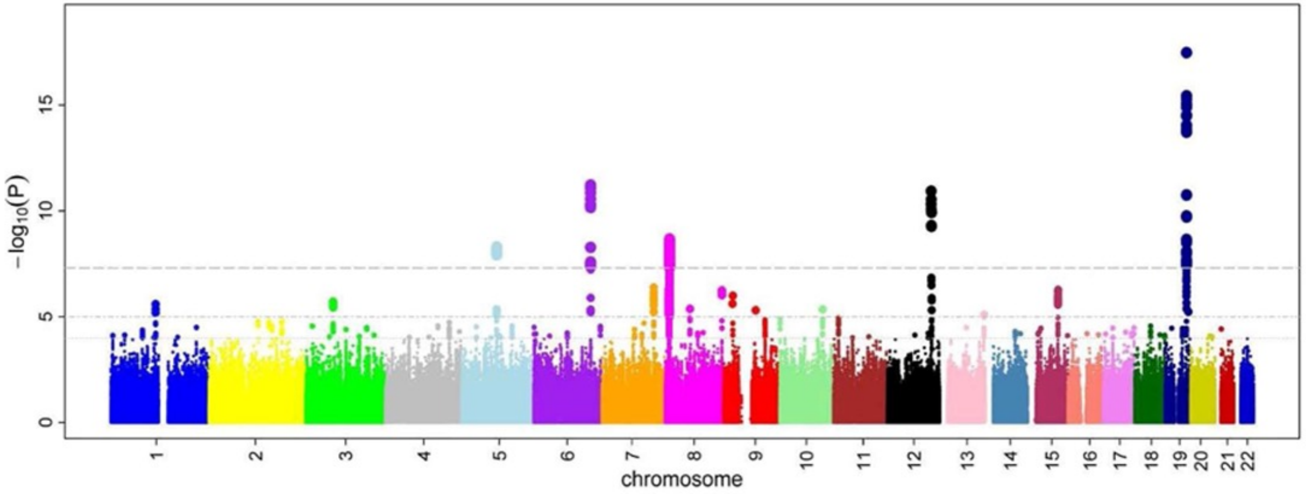
Biology 683

Lecture GLM2

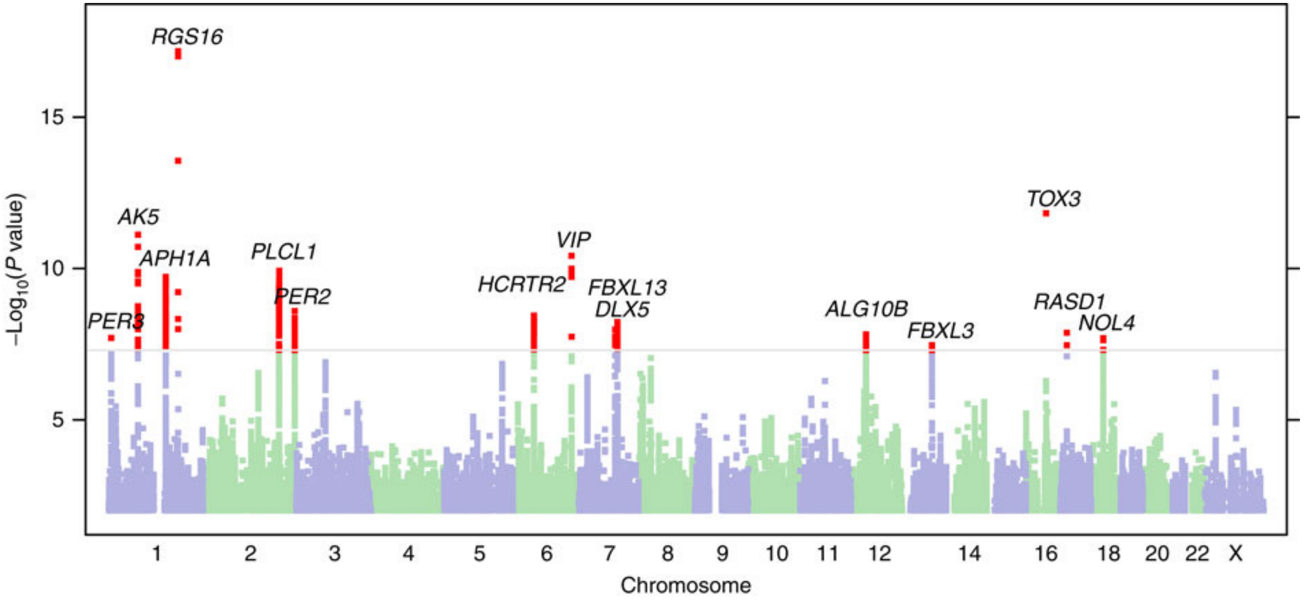
Heath Blackmon


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

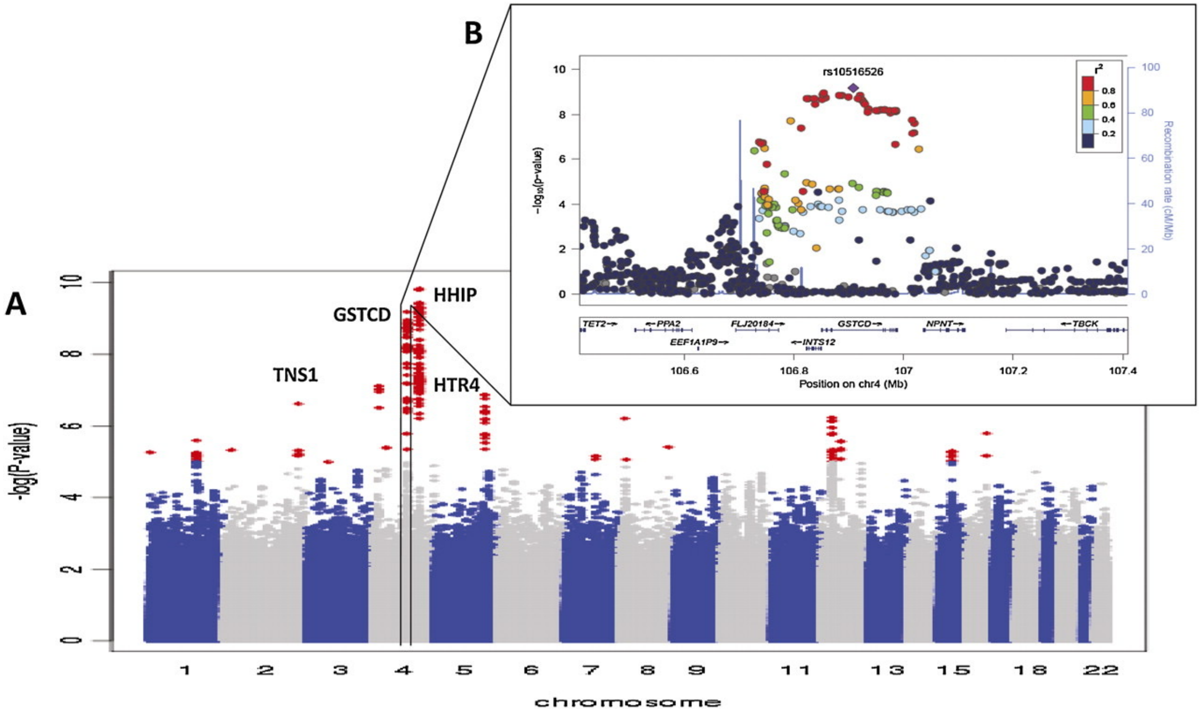
# GWAS



−log10(P)
0
5
10
15
chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



−Log10(P value)
5
10
15
RGS16
AK5
APH1A
PER3
PLCL1
PER2
HCRTR2
VIP
FBXL13
DLX5
ALG10B
FBXL3
TOX3
RASD1
NOL4
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
20
22
X



A
B
-log(P-value)
0
2
4
6
8
10
TNS1
GSTCD
HHIP
HTR4
chromosome
1
2
3
4
5
6
7
8
9
11
13
15
18
22
rs10516526
-log10(p-value)
Recombination rate (cM/Mb)
r²
0.8
0.6
0.4
0.2
0
20
40
60
80
100
TET2→
←PPA2
FLJ20184→
GSTCD→
NPNT→
←TBCK
EEF1A1P9→
←INTS12
106.6
106.8
107
107.2
107.4
Position on chr4 (Mb)

# GWAS

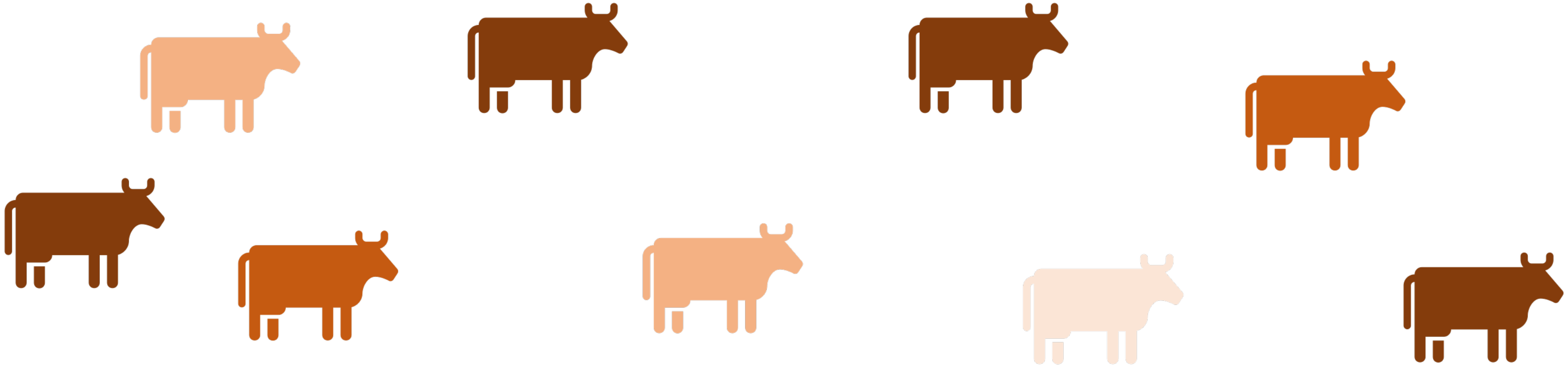
**GWAS: Genome wide association study.**

The goal of GWAS is to determine what genes have alleles that are responsible for a trait of interest. The trait can be any measurable trait in any organism that you wish to study. For instance, a disease in humans, an economically important trait of a crop or domestic animal, an adaptation like a certain color pattern in birds, etc.

# GWAS – Discrete condition (often disease)

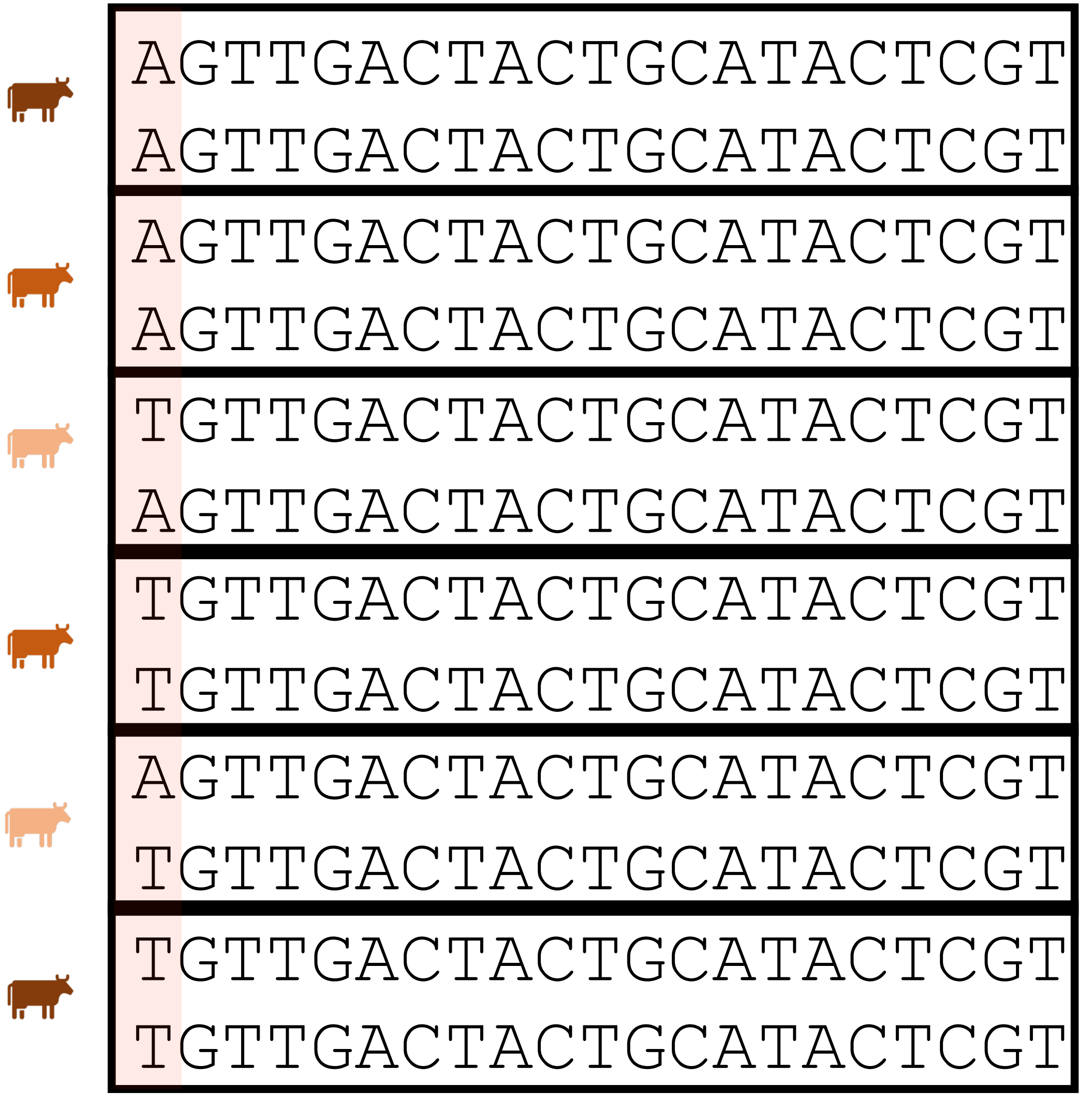
```
>case 1
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>case 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>case 3
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>control 1
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>control 2
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>control 3
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

# GWAS – Continuous trait



```
>sample 1
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>sample 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 3
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>sample 4
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 5
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>sample 6
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

# GWAS – Continuous trait



AGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

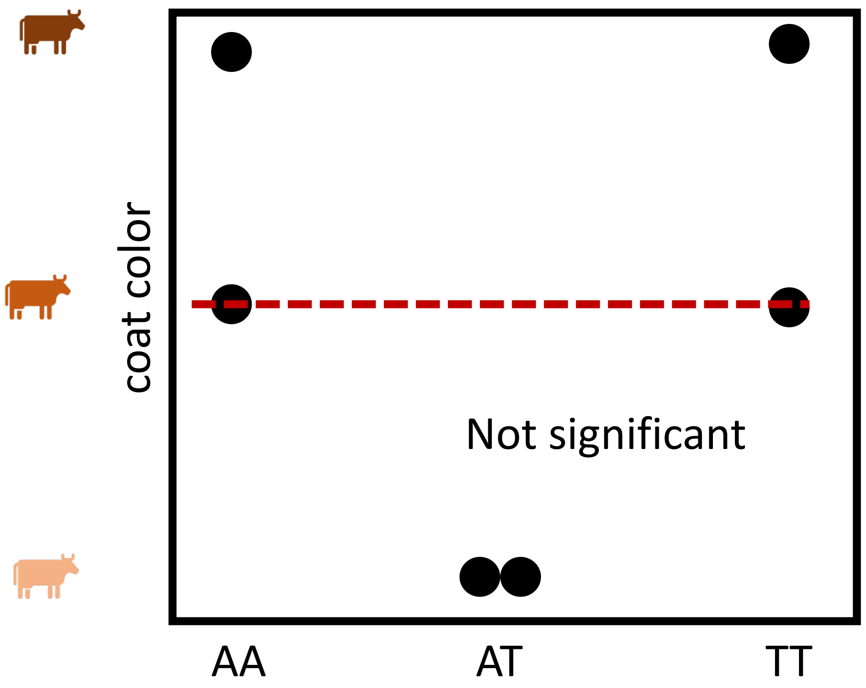
AGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

TGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

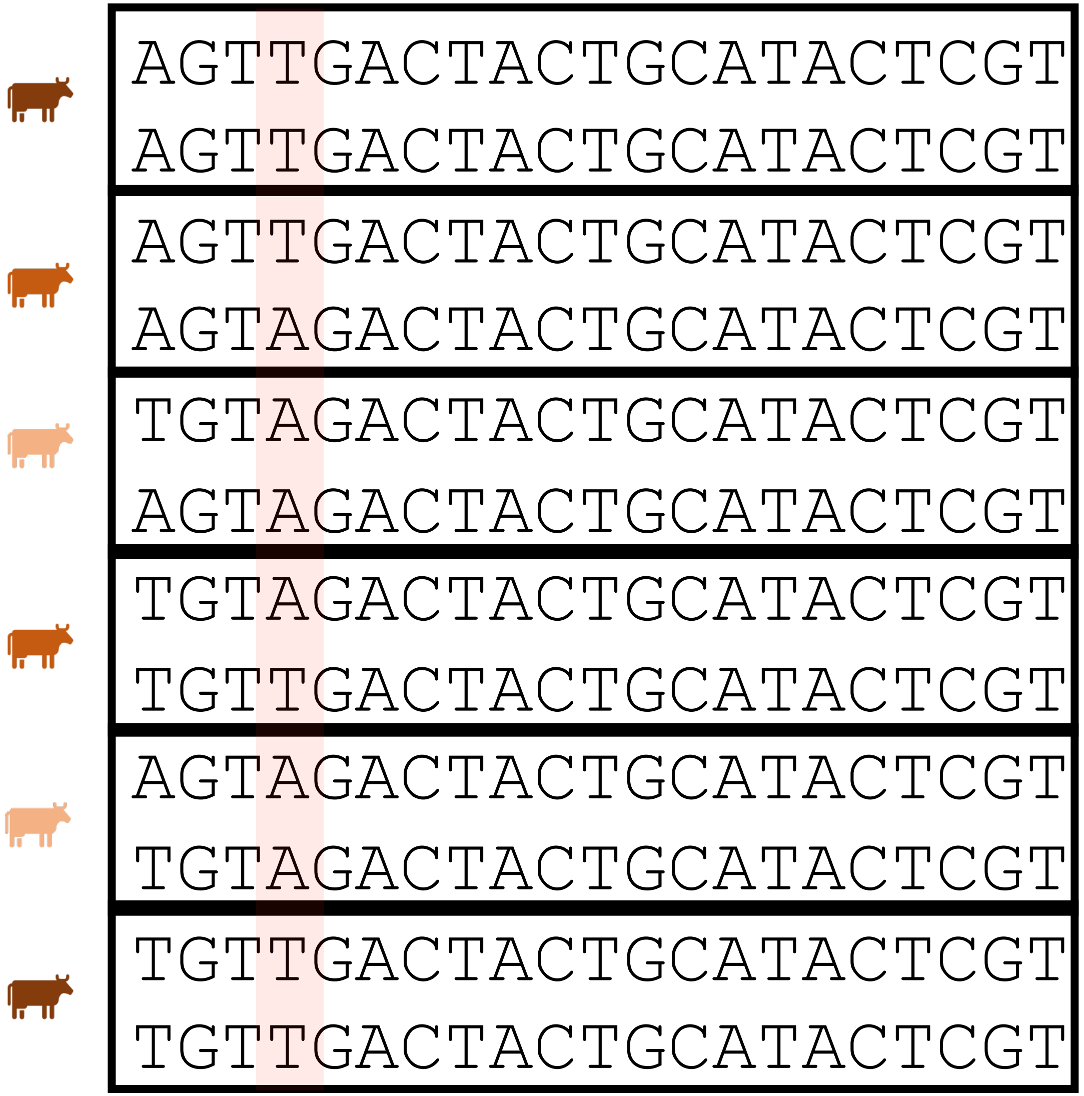
TGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

AGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

TGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT



# GWAS – Continuous trait



AGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

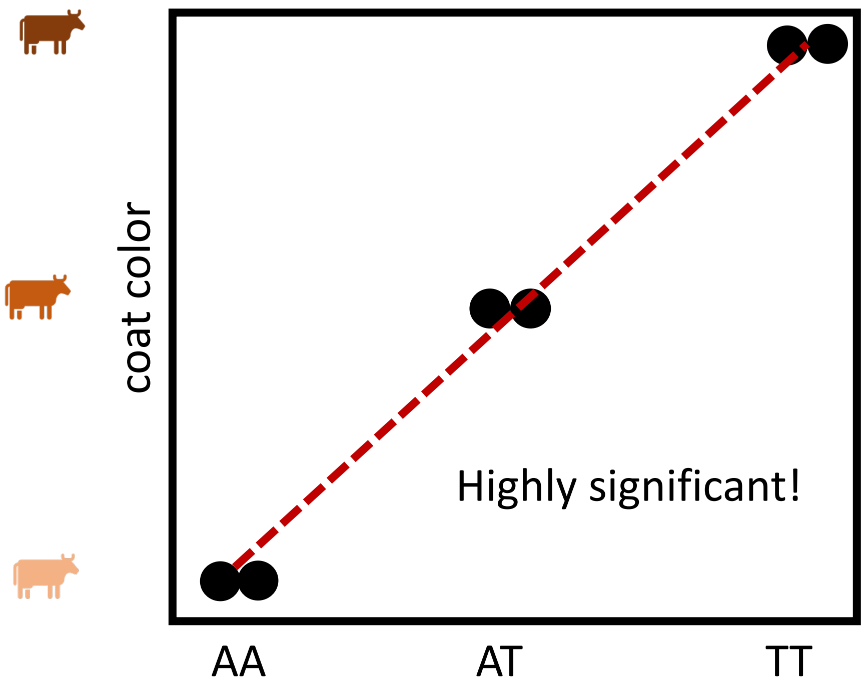
AGTTGACTACTGCATACTCGT
AGTAGACTACTGCATACTCGT

TGTAGACTACTGCATACTCGT
AGTAGACTACTGCATACTCGT

TGTAGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

AGTAGACTACTGCATACTCGT
TGTAGACTACTGCATACTCGT

TGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT



# GWAS - Problems

**What is the problem with doing this across the whole genome?**

**Multiple tests lead to more false positives!**

1) require a higher level of significance $5x10^{-8}$
2) only look at the very most significant
3) lots of more complicated approaches too!

**What is one of the most basic requirements of almost all statistical tests?**

**Tests normally assume independence of the data points!**

1) Samples from a population will be related to each other due to ancestry (trees!)
2) DNA sequencing is not done equally in all groups of people (western samples are usually over represented

# GWAS - Problems

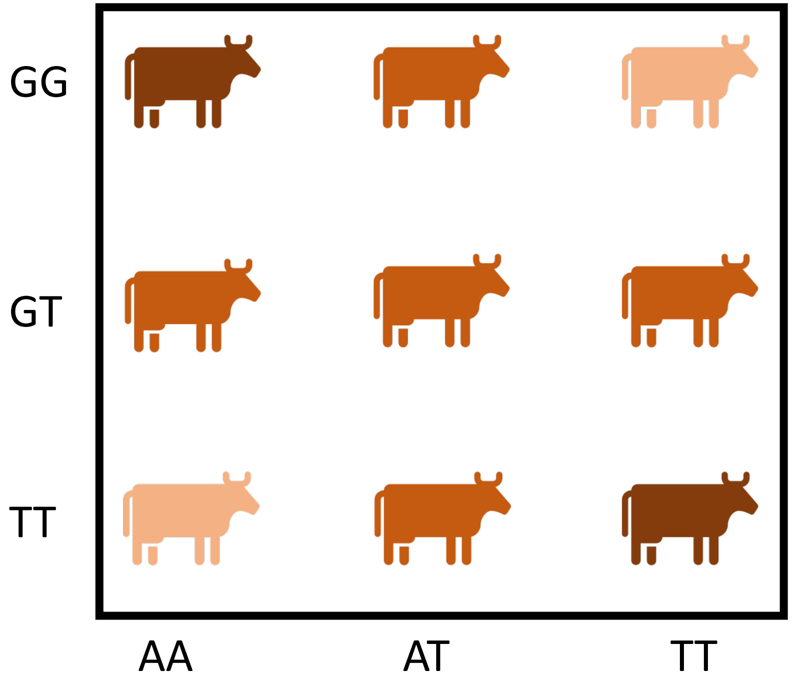
**What is epistasis?**

**It is the case where the impact of a genotype at one locus depend on the genotype at another locus!**

To find this type of effect you would need to look at all pairs of genotypes. How many tests would we need to do then?

About 6 orders of magnitude more tests would be required.

Humans have around 4,000,000 sites that are variable like this which equates to 7,999,998,000,000



# GWAS - Problems

**What about the environment?**

**Many diseases have a strong environmental component (heart disease, diabetes, cancer, etc.)**

If these are left out of the study often what is discovered is actually genetic variation that happens to coincide with environmental factors?

If a disease is more common in Europeans than Africans or Asians but it is because of a lifestyle characteristic any genetic variation that is common in Europeans but rare in Africans and Asians could appear associated with the disease.

# Applications and benefits of GWAS

Widely used in agriculture and domestication. For instance, you can do a GWAS on wild strains of rice (which have lots of variation in things we care about like grain size, growing time, etc.) This GWAS can tell you what variants at what locations in the genome should be introgressed into domestic varieties in hopes of introducing favorable traits.

Widely used in medicine to identify the genes responsible for disease. This is often the first step necessary in being able to create an animal model for a disease that will then allow researchers to study the disease and develop pharmaceutical interventions.

Many diseases have multiple different underlying genetic causes and treatments that are available may only work on some forms of the diseases. Thanks to GWAS studies we now know what these different causal genes are. Now patients can be genotyped at these causative loci and medication can be tailored to their version of the disease. (precision or personalized medicine)

Applied in a more limited fashion to inform patients of risk that they will develop more severe versions of a disease. For instance depending on your genotype at a gene you may choose a more aggressive form of treatment.

# Complex Disease / Complex Phenotypes

Many diseases are what we call complex diseases. There is no one gene responsible for the disease. Instead the disease can manifest due to variations in 10 or 100s of different genes in the genome acting in concert with the environment. GWAS is less insightful (though still important) for diseases like this.

Schizophrenia: Not really a clearly delineated disease like say COVID, type 1 diabetes, or cystic fibrosis. Instead it is a constellation of symptoms that individuals exhibit to varying degrees.

From studies of multiple generations of families we know that 20-30% of risk is inherited (genetic). The other 70-80% of your risk of developing the disease is environmental and is poorly understood. Massive GWAS studies with 1000s of individuals have studied the genetic component. These studies have identified more than 100 different loci in the genome that seem to have some predictive power. However, these variations are associated with an increase in risk and there are many people who carry many alleles that increase risk but that never develop the disease.

# Mixed models

Mixed models are models that include fixed and and random effects.

Fixed effects can be repeated by other researchers. These are the variables that you are interested in studying.

Random effects are usually nuisance parameters. These are variables that other researchers cannot replicate and you are not interested in inferring anything about them.

# Mixed models

Fixed effects are the variables whose impact we wish to determine

- Characteristics of the media or habitat
- Experimental treatments
- Age groups
- Time points
- Mutant genotypes

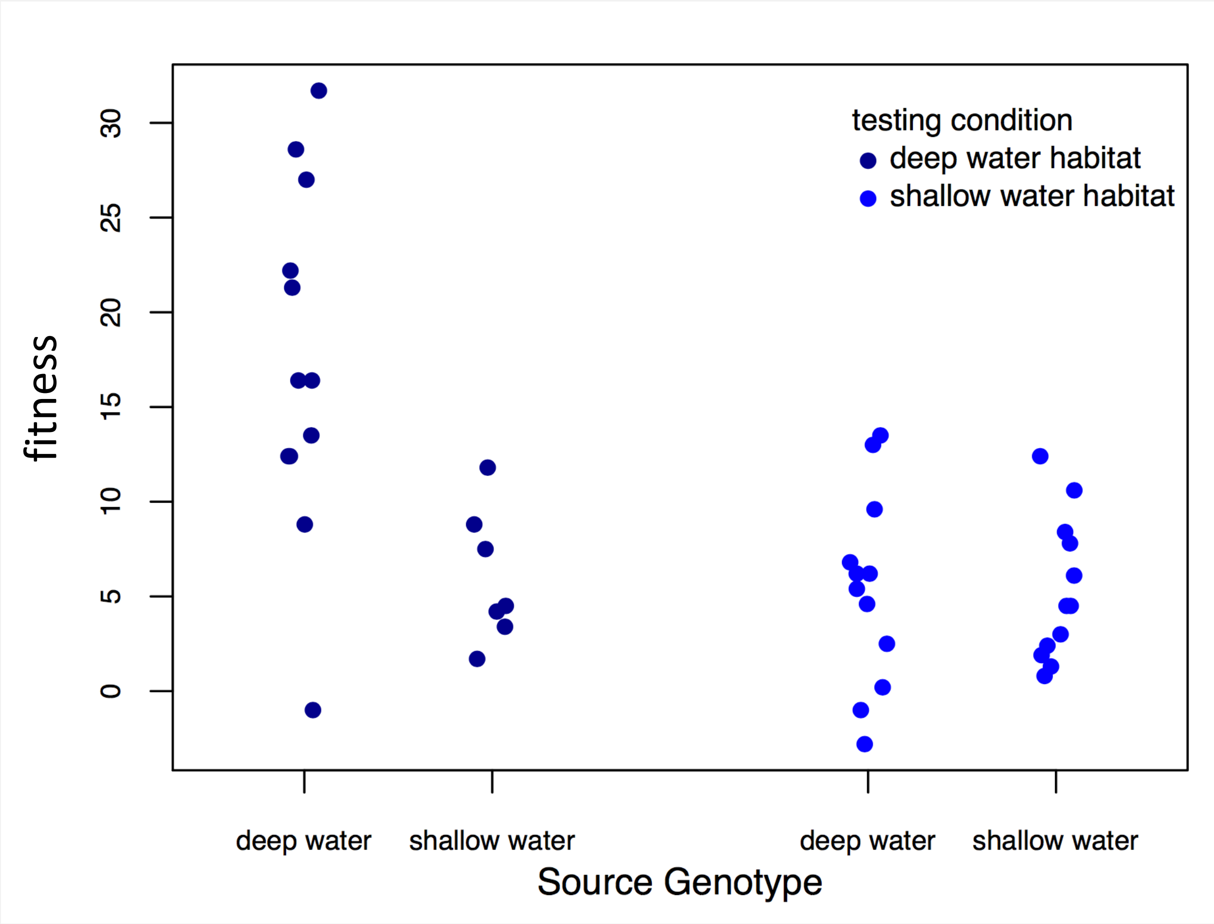
Conclusions that you reach are only applicable to the groups or treatments you include in the study

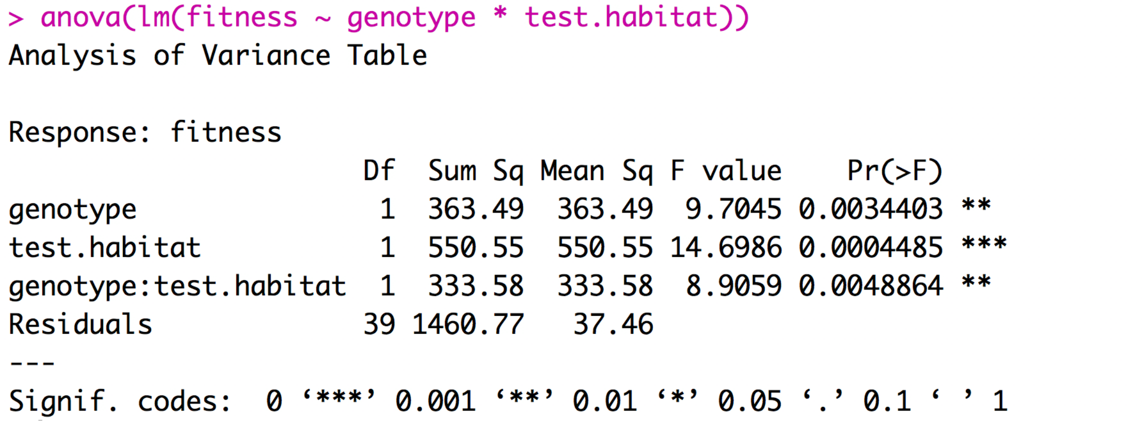
# Example of fixed effects

Reciprocal relocation experiment to investigate how genotype and habitat interact to determine the fitness of stickleback fish (Rundle 2002).

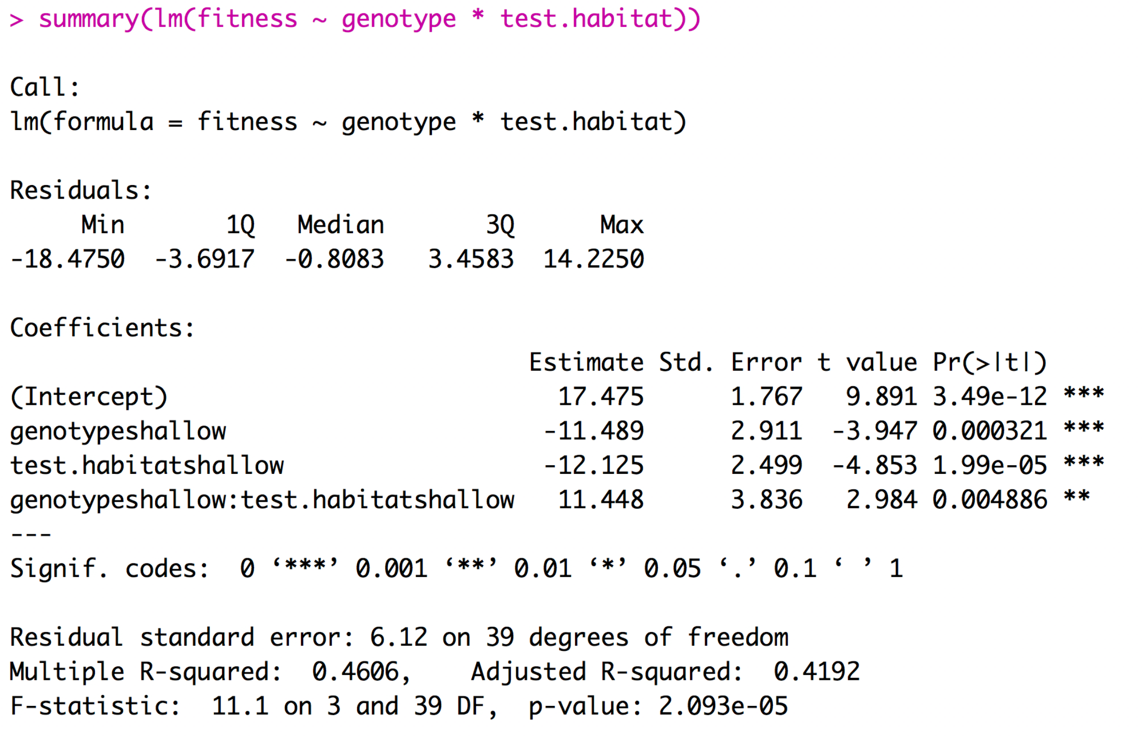
|  |  | Source habitat |  |
|---|---|---|---|
|  |  | Shallow | Deep |
| Test habitat | Shallow | 12 fish | 11 fish |
|  | Deep | 7 fish | 11 fish |

# Example of fixed effects (two factor ANOVA)



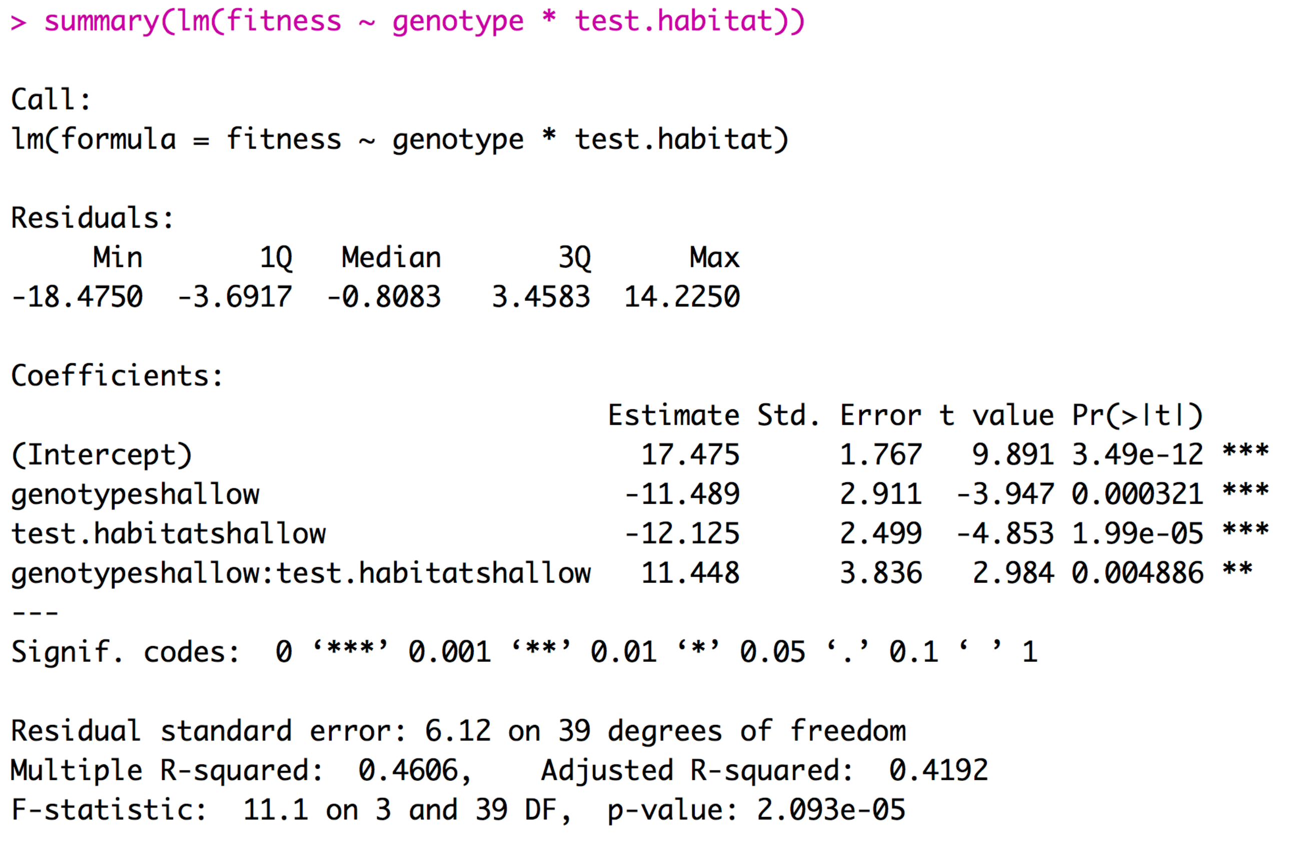


```
> anova(lm(fitness ~ genotype * test.habitat))
Analysis of Variance Table

Response: fitness
                      Df  Sum Sq Mean Sq F value    Pr(>F)    
genotype               1  363.49  363.49  9.7045 0.0034403 ** 
test.habitat           1  550.55  550.55 14.6986 0.0004485 ***
genotype:test.habitat  1  333.58  333.58  8.9059 0.0048864 ** 
Residuals             39 1460.77   37.46                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

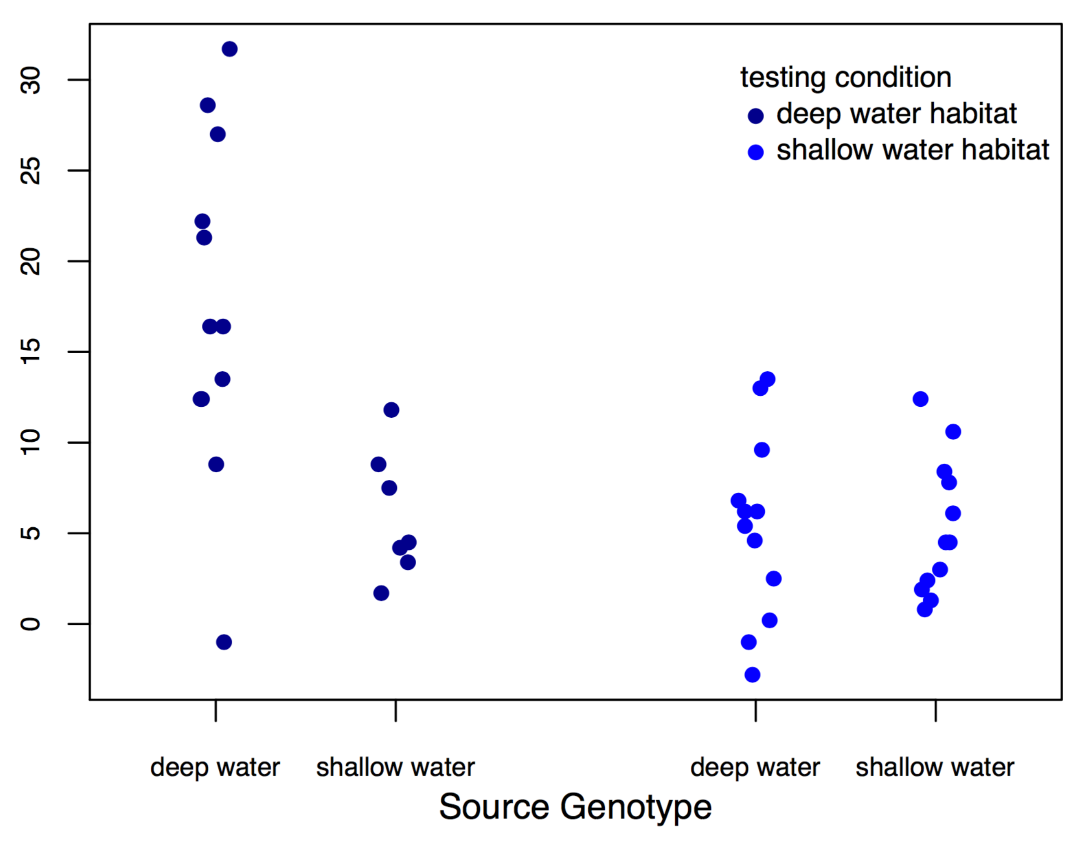


```
> summary(lm(fitness ~ genotype * test.habitat))

Call:
lm(formula = fitness ~ genotype * test.habitat)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4750  -3.6917  -0.8083   3.4583  14.2250 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        17.475      1.767   9.891 3.49e-12 ***
genotypeshallow                   -11.489      2.911  -3.947 0.000321 ***
test.habitatshallow               -12.125      2.499  -4.853 1.99e-05 ***
genotypeshallow:test.habitatshallow  11.448      3.836   2.984 0.004886 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.12 on 39 degrees of freedom
Multiple R-squared:  0.4606,	Adjusted R-squared:  0.4192 
F-statistic:  11.1 on 3 and 39 DF,  p-value: 2.093e-05
```

# Interpreting Coefficients



```
> summary(lm(fitness ~ genotype * test.habitat))

Call:
lm(formula = fitness ~ genotype * test.habitat)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4750  -3.6917  -0.8083   3.4583  14.2250 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         17.475      1.767   9.891 3.49e-12 ***
genotypeshallow                    -11.489      2.911  -3.947 0.000321 ***
test.habitatshallow                -12.125      2.499  -4.853 1.99e-05 ***
genotypeshallow:test.habitatshallow  11.448      3.836   2.984 0.004886 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.12 on 39 degrees of freedom
Multiple R-squared:  0.4606,	Adjusted R-squared:  0.4192 
F-statistic:  11.1 on 3 and 39 DF,  p-value: 2.093e-05
```



# What is a random effect

These are randomly sampled categories of a variable that represent groups of individual measurements.  Usually random effects are not repeatable.

- Study sites
- Environmental chambers
- Families made up of siblings
- Measurements within individuals

Conclusions that you reach are applicable only to the sample being studied.

# What is a random effect

Sometimes random effects are a nuisance

- Field sites
- Environmental chambers
- Field plots
- Repeated measures

Occasionally random effects are of great interest

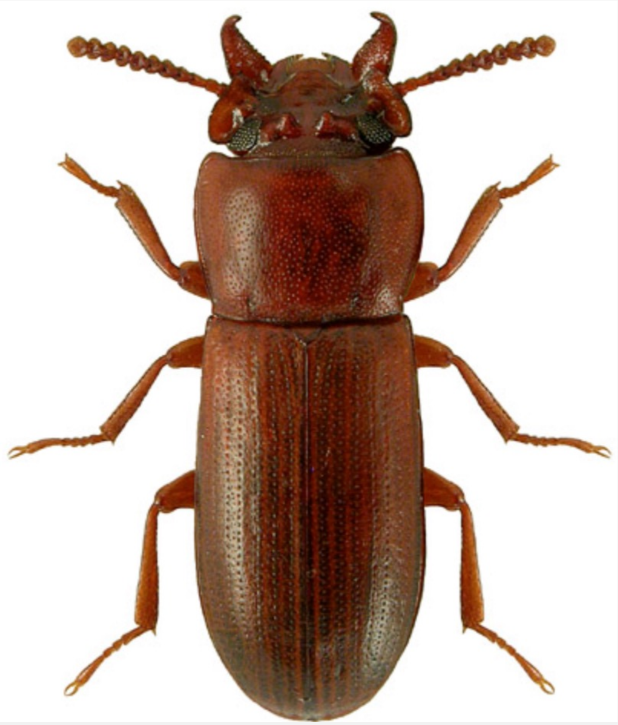
- Families - Heritability
- Individuals – Breeding value

# Example of random effects

Impact of selective regime on horn size.

Measure both the left and right horn in 25 beetles from two different selective regimes.

| Horn size | Beetle | Selective regime |
|---|---|---|
| 256 | 1 | High |
| 276 | 1 | High |
| 321 | 2 | High |
| 321 | 2 | High |
| 423 | 3 | Low |
| 401 | 3 | Low |
| 381 | 4 | Low |
| 409 | 4 | Low |



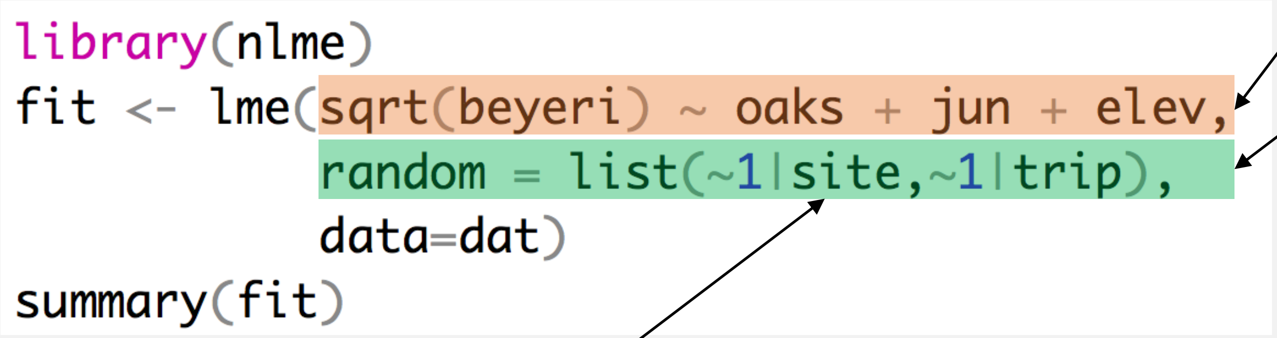
# Example of random effects

Identifying the predictors for the presence or absence of Chrysina beetles.

| Number collected | oak | juniper | site | date | trip |
|---|---|---|---|---|---|
| 8 | 1 | 0 | 21 | 210 | A |
| 2 | 1 | 1 | 13 | 210 | A |
| 1 | 0 | 1 | 31 | 211 | A |
| 5 | 0 | 1 | 15 | 212 | A |
| 4 | 1 | 1 | 21 | 242 | B |
| 6 | 1 | 0 | 13 | 242 | B |
| 0 | 1 | 1 | 31 | 245 | B |
| 7 | 1 | 1 | 15 | 245 | B |

# Implementing a mixed effects model

Mixed effect models can be fit using the LME function from the package nlme.



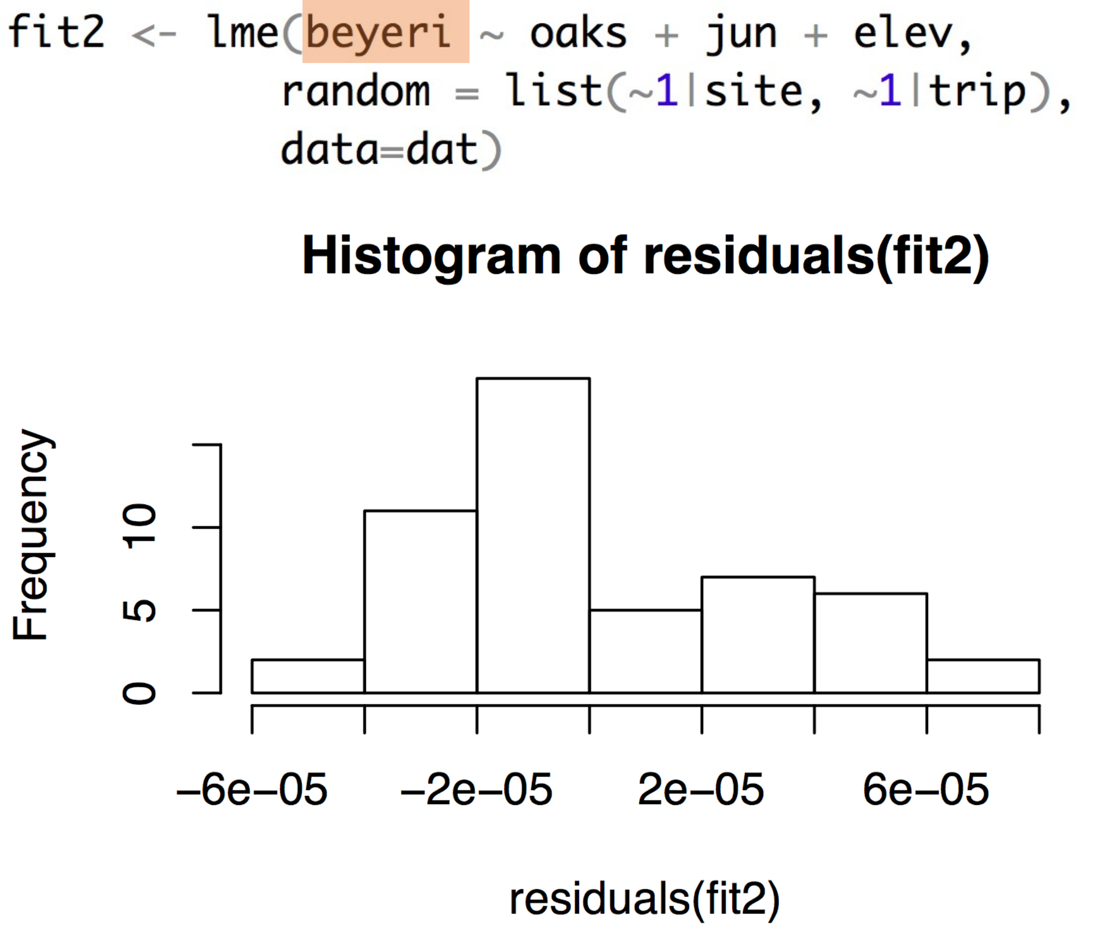
```
library(nlme)
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,
           random = list(~1|site,~1|trip),
           data=dat)
summary(fit)
```

Fixed effects

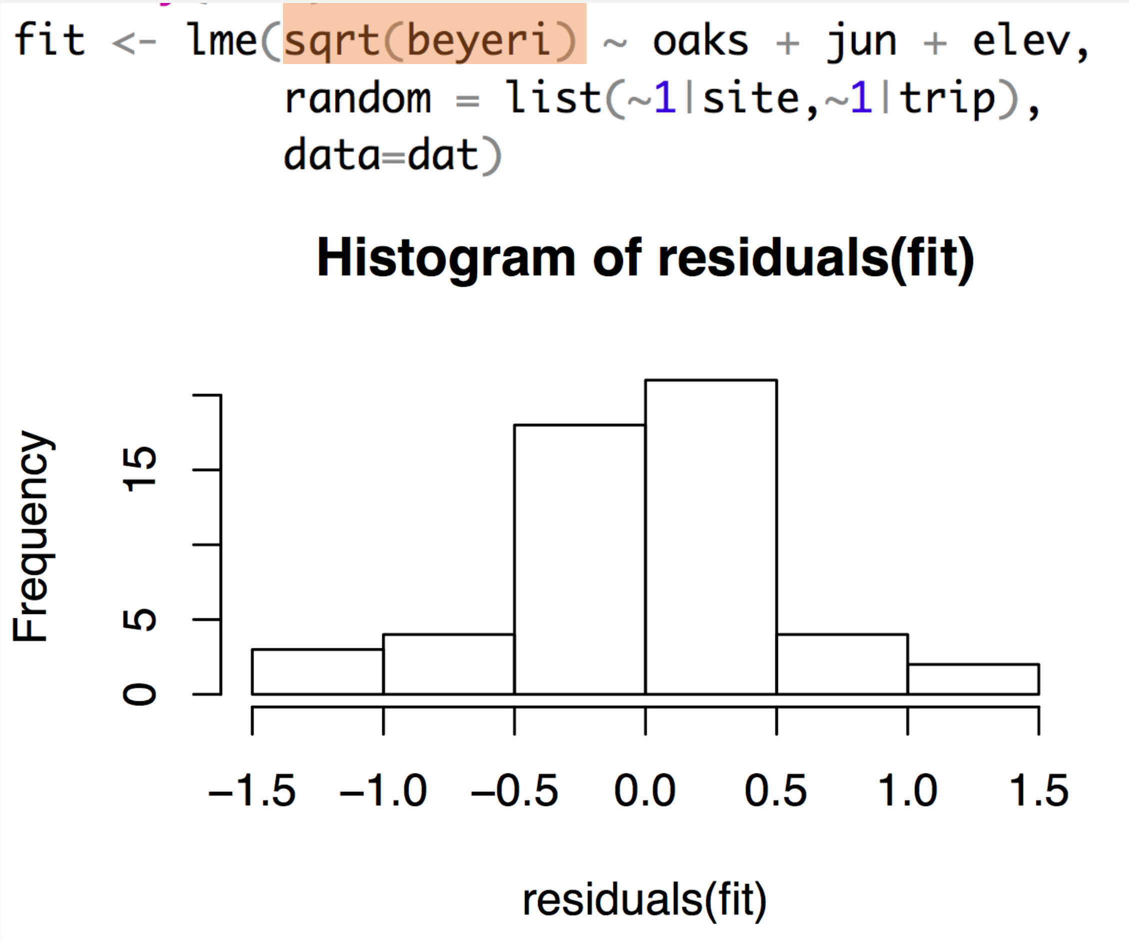
Random effects

Repeated measures at sites can’t
be treated as independent

# Implementing a mixed effects model



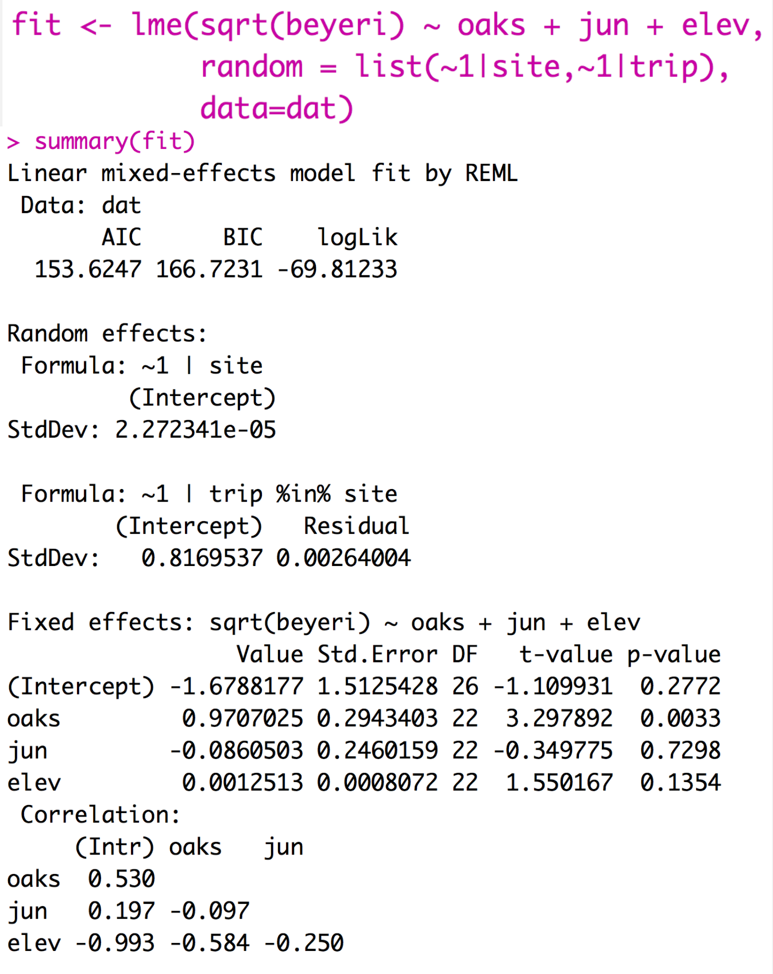
```
fit2 <- lme(beyeri ~ oaks + jun + elev,
            random = list(~1|site, ~1|trip),
            data=dat)
```



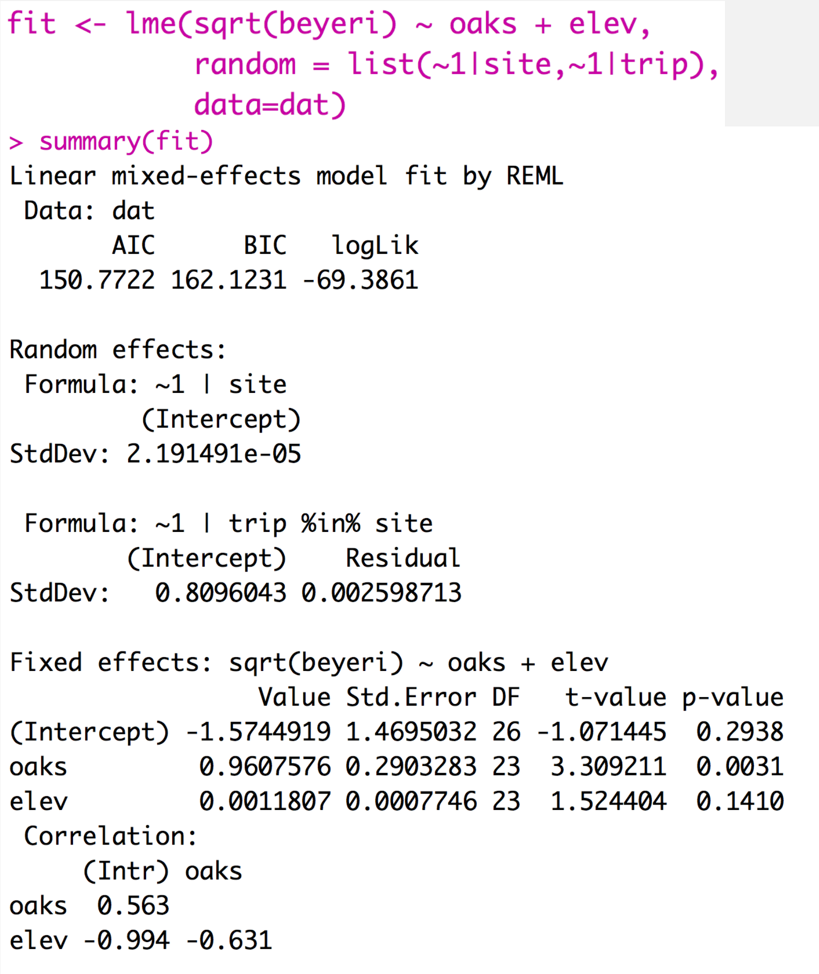
```
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,
           random = list(~1|site,~1|trip),
           data=dat)
```

# Implementing a mixed effects model

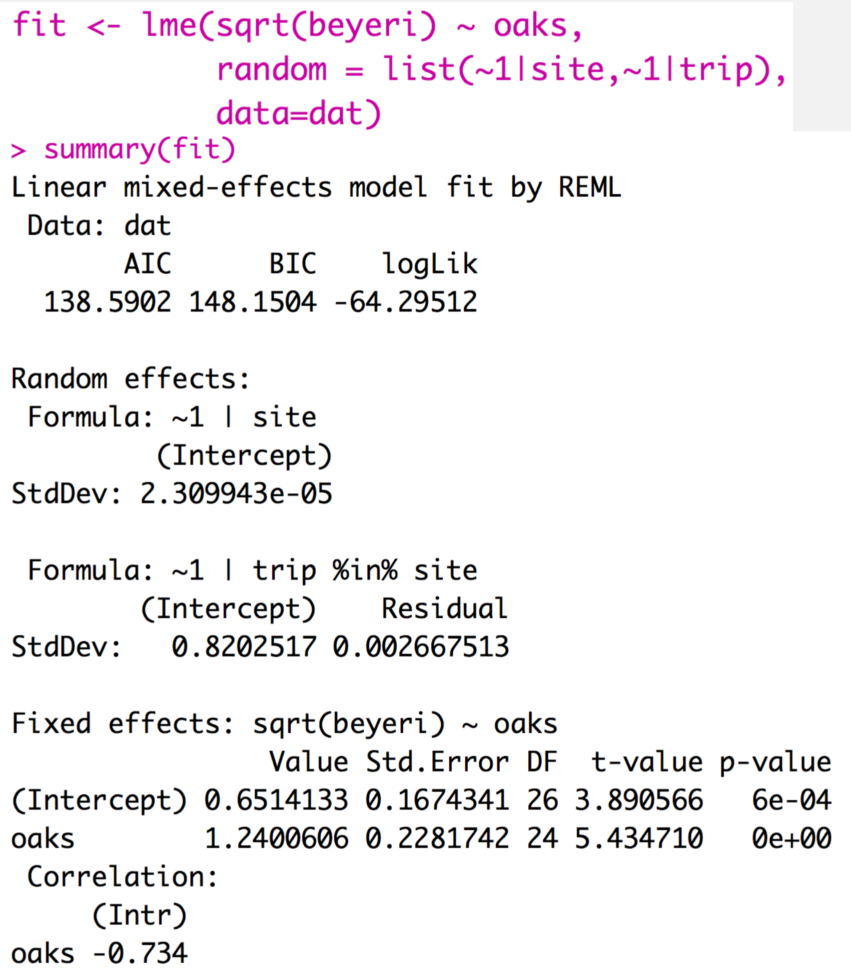
Mixed effect models can be fit using the LME function from the package nlme.



```
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,
             random = list(~1|site,~1|trip),
             data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
 Data: dat
       AIC      BIC    logLik
  153.6247 166.7231 -69.81233

Random effects:
 Formula: ~1 | site
        (Intercept)
StdDev: 2.272341e-05

 Formula: ~1 | trip %in% site
        (Intercept)   Residual
StdDev:   0.8169537 0.00264004

Fixed effects: sqrt(beyeri) ~ oaks + jun + elev
                 Value Std.Error DF   t-value p-value
(Intercept) -1.6788177 1.5125428 26 -1.109931  0.2772
oaks         0.9707025 0.2943403 22  3.297892  0.0033
jun         -0.0860503 0.2460159 22 -0.349775  0.7298
elev         0.0012513 0.0008072 22  1.550167  0.1354
 Correlation:
     (Intr) oaks   jun
oaks  0.530
jun   0.197 -0.097
elev -0.993 -0.584 -0.250
```



```
fit <- lme(sqrt(beyeri) ~ oaks + elev,
             random = list(~1|site,~1|trip),
             data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
 Data: dat
       AIC      BIC    logLik
  150.7722 162.1231 -69.3861

Random effects:
 Formula: ~1 | site
        (Intercept)
StdDev: 2.191491e-05

 Formula: ~1 | trip %in% site
        (Intercept)    Residual
StdDev:   0.8096043 0.002598713

Fixed effects: sqrt(beyeri) ~ oaks + elev
                 Value Std.Error DF   t-value p-value
(Intercept) -1.5744919 1.4695032 26 -1.071445  0.2938
oaks         0.9607576 0.2903283 23  3.309211  0.0031
elev         0.0011807 0.0007746 23  1.524404  0.1410
 Correlation:
     (Intr) oaks
oaks  0.563
elev -0.994 -0.631
```



```
fit <- lme(sqrt(beyeri) ~ oaks,
             random = list(~1|site,~1|trip),
             data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
 Data: dat
       AIC      BIC    logLik
  138.5902 148.1504 -64.29512

Random effects:
 Formula: ~1 | site
        (Intercept)
StdDev: 2.309943e-05

 Formula: ~1 | trip %in% site
        (Intercept)    Residual
StdDev:   0.8202517 0.002667513

Fixed effects: sqrt(beyeri) ~ oaks
                Value Std.Error DF  t-value p-value
(Intercept) 0.6514133 0.1674341 26 3.890566   6e-04
oaks        1.2400606 0.2281742 24 5.434710   0e+00
 Correlation:
     (Intr)
oaks -0.734
```

If you report your df with your F-statistic the reviewer will know if you did the right type of model

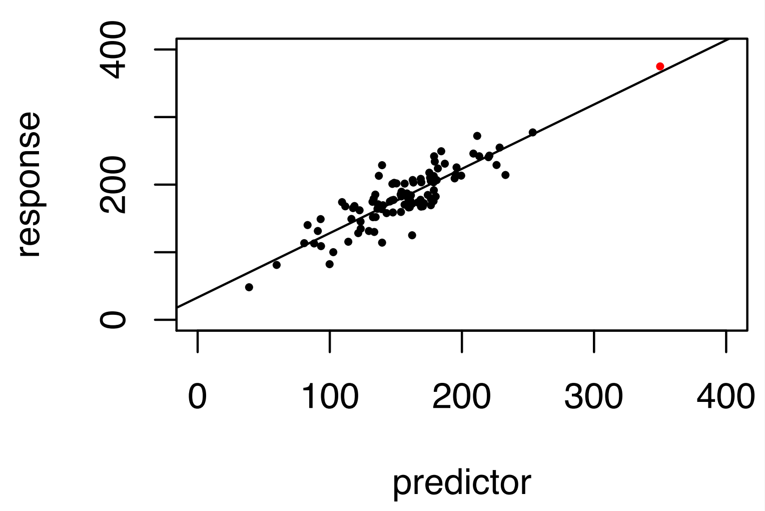
# Considerations for models with random effects

- Most software will assume that all factors are fixed unless you specify them as mixed.
- Designating factors as random effects takes extra work.
- The lm function treats all predictors as fixed effects.
- Treating random effects as fixed effects is fundamentally wrong.

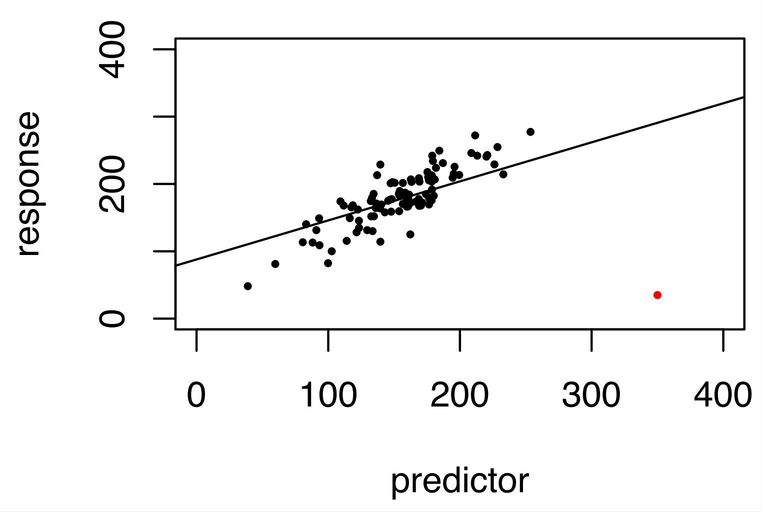
# 4 Types of outliers

1) Obviously erroneously recorded or measured data

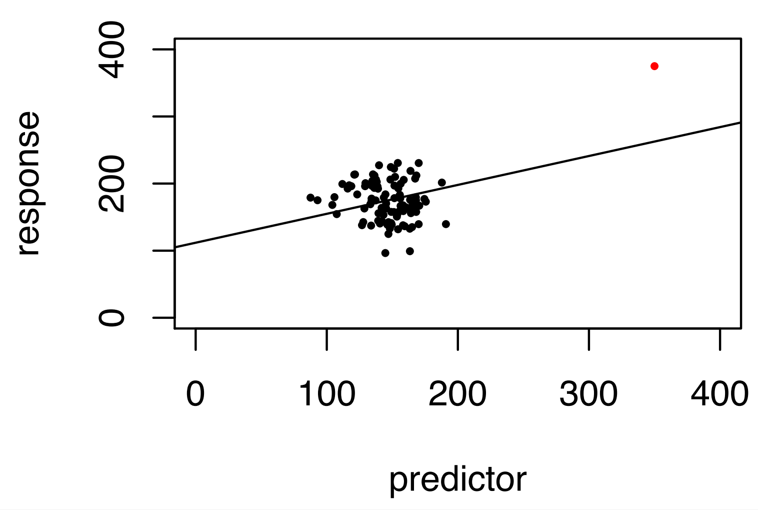
2) Extreme data point that doesn't impact statistic of interest but does impact p-value.



3) Extreme data point that impacts statistic of interest.



4) Extreme data point that creates significance.



# Links

MCMCMglmm: Fit mixed models with phylogenetic or pedigree information in a Bayesian framework.

Outlier Package: Apply outlier tests to identify possible outlier datapoints - I don’t recommend this.

**If time permits: repeated median regression (Theil-Sen Estimator)**

- **simulate x and y with medium strength relationship and add one outlier**
- **standard regression record Beta**
- **record slope between all pairs and find median slope**
- **repeat standard regression without outlier**