

STAT588/BIOL588: Genomic Data Science
Lecture 7: Simple Marker Association Test

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

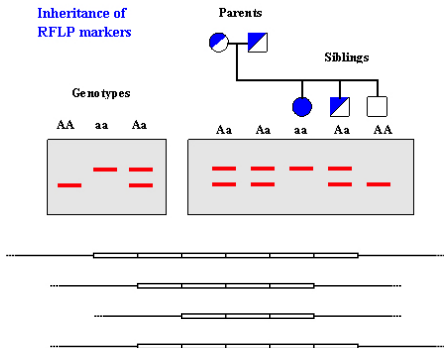
Genetic Markers



A genetic marker is a DNA sequence with a known physical location on a chromosome.

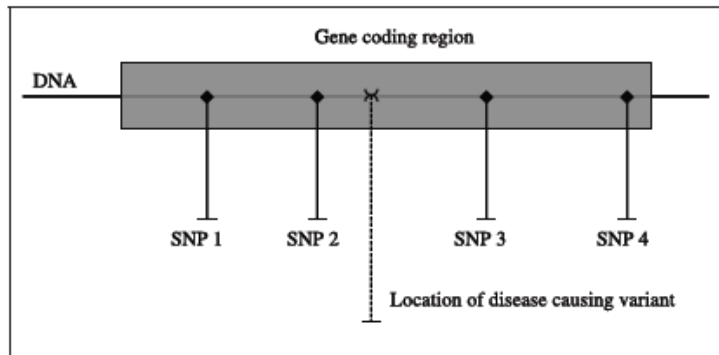
Commonly Used Genetic Markers

- ▶ Restriction fragment length polymorphism (RFLP)
- ▶ SNPs (single nucleotide polymorphism)
- ▶ Variable number tandem repeat (VNTR)
- ▶ Microsatellites: length of the repeating unit is less than five base pairs.



Gene Association Analysis

SNP markers



Recombination Fraction

The recombination fraction (r) between two loci is the probability that a recombination occurs between the two loci.

Kosambi

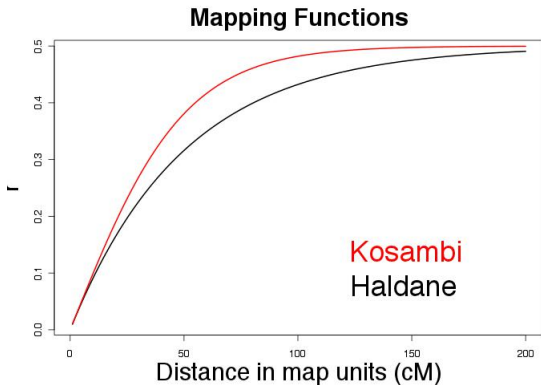
$$r = 1/2 \times \frac{e^{d/25} - 1}{1 + e^{d/25}}$$

Haldane

$$r = 1/2 \times (1 - e^{-d/50})$$

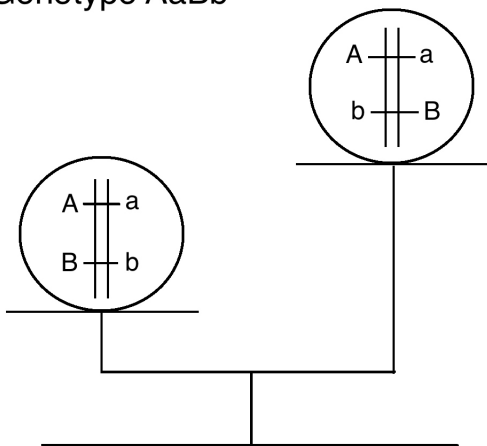
d : map units (cM)

1cM = 1% $\approx 10^6$ base pairs.

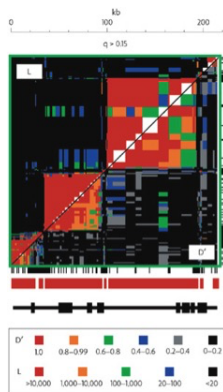


Linkage Disequilibrium

Genotype AaBb



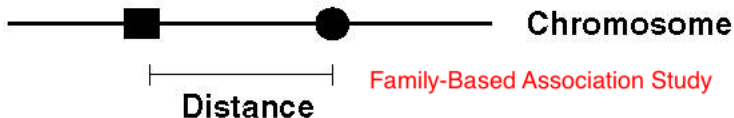
Linkage Disequilibrium



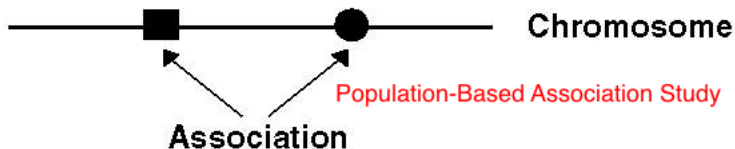
Linkage disequilibrium declined quickly to the background level at an average distance of 17 kb.

Linkage versus Linkage Disequilibrium

RECOMBINATION RATE:



LINKAGE DISEQUILIBRIUM:



The Cattle Data

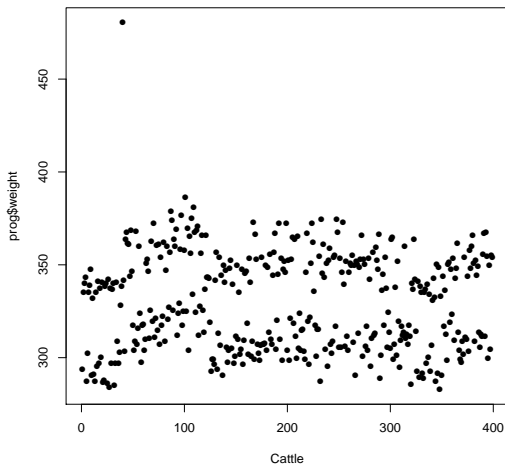
Identifying genetic loci associated with cattle's weight Quantitative Trait Locus (QTL) analysis

```
>sireurl<-"http://people.stat.sc.edu/hoyen/BIOL599/Data/siredata.txt"
>sires<-read.table(file=sireurl, header=T, sep="\t")
>progurl<-"http://people.stat.sc.edu/hoyen/BIOL599/Data/progdata.txt"
>prog<-read.table(file=progurl, header=T, sep="\t")
>str(sires)
>str(prog)
> str(sires)
'data.frame': 10 obs. of 12 variables:
 $ id      : Factor w/ 10 levels "sire1","sire10",...: 1 3 4 5 6 7 8 9 10 2
 $ weight: num  334 365 384 350 358 ...
 $ m11    : Factor w/ 4 levels "M1","M2","M3",...: 2 3 2 2 1 3 4 3 1 2
 $ m12    : Factor w/ 4 levels "M1","M2","M3",...: 1 2 4 1 3 2 1 4 2 3
> str(prog)
'data.frame': 400 obs. of 14 variables:
 $ id      : Factor w/ 400 levels "id1","id10","id100",...: 1 112 223 334 346 357
 $ sire    : Factor w/ 10 levels "sire1","sire10",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ sex     : Factor w/ 2 levels "F","M": 1 2 2 2 1 1 2 2 2 1 ...
 $ weight  : Factor w/ 386 levels "-", "282.83", "283.83",...: 26 205 225 241 8
```

10 Sires



40 offersprings / per sire

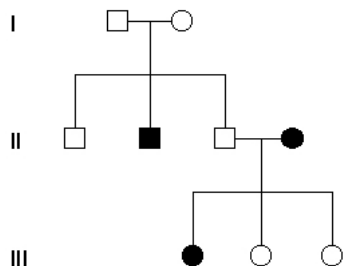


10 Sires



Pedigree format

Recessive Pedigree

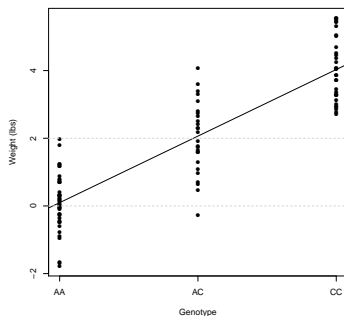


	faid	moid	iid	sex	trait
1	0	0	1	1	0
2	0	0	2	0	0
3	1	2	3	1	0
4	1	2	4	1	1
5	1	2	5	1	0
6	0	0	6	0	1
7	5	6	7	0	1
8	5	6	8	0	0
9	5	6	9	0	0

Building Regression Model: Linear

$$\hat{Y} = \beta_0 + \beta_1 X$$

	y	Genotype	x
1	2.85	CC	2
2	0.40	AA	0
3	3.28	CC	2
4	1.80	AA	0
5	2.19	CA	1
6	1.97	AA	0
7	0.64	CA	1



```
>fit<-lm(y ~ x)
```

```
>summary(fit)
```

```
Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.1713      0.1821    0.941  0.349
x               1.9490      0.2313   8.427 3.09e-13 ***
```

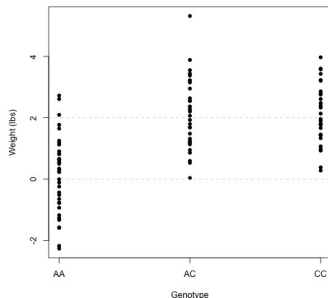
```
----
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Building Regression Model: Dominant

$$\hat{Y} = \beta_0 + \beta_1 X$$

	y	Genotype	x
1	2.85	CC	1
2	0.40	AA	0
3	3.28	CC	1
4	1.80	AA	0
5	2.19	CA	1
6	1.97	AA	0
7	0.64	CA	1



```
>fit<-lm(y ~ x)
```

```
>summary(fit)
```

```
Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2154      0.1603  -1.344   0.182
x              2.0322      0.1044  19.474  <2e-16 ***
```

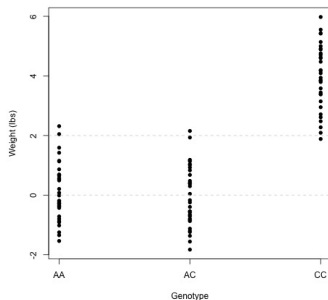
```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Building Regression Model: Recessive

$$\hat{Y} = \beta_0 + \beta_1 X$$

	y	Genotype	x
1	2.85	CC	1
2	0.40	AA	0
3	3.28	CC	1
4	1.80	AA	0
5	2.19	CA	?
6	1.97	AA	0
7	0.64	CA	?



```
>fit<-lm(y ~ x)
```

```
>summary(fit)
```

Coefficients:

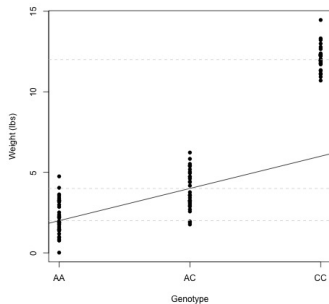
```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2154      0.1603   -1.344   0.182
x              2.0322      0.1044  19.474 <2e-16 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Building Regression Model: Codominant

$$\hat{Y} = \beta_0 + \beta_1 X$$

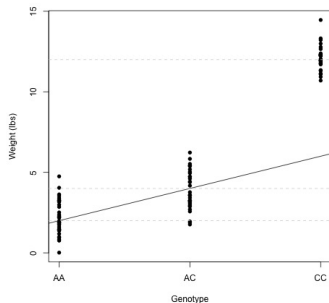
	y	Genotype	x
1	2.85	CC	?
2	0.40	AA	?
3	3.28	CC	?
4	1.80	AA	?
5	2.19	CA	?
6	1.97	AA	?
7	0.64	CA	?



Building Regression Model: Dummy Variables

$$\hat{Y} = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

	y	Genotype	x1	x2
1	2.85	CC	0	1
2	0.40	AA	0	0
3	3.28	CC	0	1
4	1.80	AA	0	0
5	2.19	CA	1	0
6	1.97	AA	0	0
7	0.64	CA	1	0



```
>fit<-lm(y ~ x1 + x2)
```

```
>summary(fit)
```

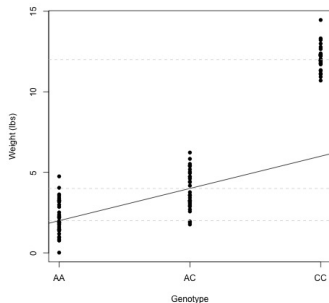
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.2065	0.1694	13.024	< 2e-16 ***
x1	1.6969	0.2448	6.931	4.62e-10 ***
x2	9.9155	0.2556	38.796	< 2e-16 ***

Building Regression Model: Dummy Variables

$$\hat{Y} = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

	y	Genotype	x1	x2
1	2.85	CC	0	1
2	0.40	AA	0	0
3	3.28	CC	0	1
4	1.80	AA	0	0
5	2.19	CA	1	0
6	1.97	AA	0	0
7	0.64	CA	1	0



```
>fit<-lm(y ~ x1 + x2)
```

```
>summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.2065	0.1694	13.024	< 2e-16 ***
x1	1.6969	0.2448	6.931	4.62e-10 ***
x2	9.9155	0.2556	38.796	< 2e-16 ***

Based on the regression output, what is the difference of average weight between CA and AA? CC and AA? CC and CA?

Treat X as a factor

	y	x
1	3.74	CA
2	1.66	CA
3	1.62	CC
4	3.89	AA
5	2.06	AA
6	2.68	CC

```
> str(dat3)
'data.frame': 100 obs. of 2 variables:
 $ y: num 3.74 1.66 1.62 3.89 2.06 ...
 $ x: chr "CA" "CA" "CC" "AA" ...
> dat3$x<-as.factor(x)
> str(dat3)
'data.frame': 100 obs. of 2 variables:
 $ y: num 3.74 1.66 1.62 3.89 2.06 ...
 $ x: Factor w/ 3 levels "AA","CA","CC": 2 2 3 1 1 3 1 2 3 2 ...
```

Treat X as a factor

$$\hat{Y} = \beta_0 + \beta_1 \times XCA + \beta_2 \times XCC$$

```
> dat3$x <- as.factor(x)
```

```
> fit <- lm(y ~ x, data = dat3)
```

```
> summary(fit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.7384	0.1977	8.793	5.41e-14	***
xCA	2.2884	0.2680	8.538	1.91e-13	***
xCC	10.1082	0.2774	36.438	< 2e-16	***

Treat X as a factor

$$\hat{Y} = \beta_0 + \beta_1 \times XCA + \beta_2 \times XCC$$

```
> dat3$x<-as.factor(x)
> fit<-lm(y~ x, data=dat3)
> summary(fit)
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.7384	0.1977	8.793	5.41e-14	***
xCA	2.2884	0.2680	8.538	1.91e-13	***
xCC	10.1082	0.2774	36.438	< 2e-16	***

Based on the regression output, what is the difference of average weight between CA and AA? CC and AA? CC and CA? Who is the “reference group”?

The Sire Data: Clean

	id	sire	sex	weight	M1	M2	M3	M4	M5
1	id2	sire1	M	335.43	M1_M4	M1_M3	M4_M5	M3_M4	M2_M2
2	id3	sire1	M	340.09	M2_M3	M2_M6	M1_M3	M3_M4	M3_M4
3	id4	sire1	M	343.08	M2_M3	M1_M2	M4_M6	M2_M3	M4_M5
4	id5	sire1	F	287.08	M1_M3	M2_M4	M4_M6	M4_M5	M2_M3
5	id6	sire1	F	302.17	M2_M2	M2_M5	M3_M5	M2_M4	M1_M4

```
> cleandata<-read.table("cleandata.txt", header=T, sep="\t")
> str(cleandata)
'data.frame': 396 obs. of 9 variables:
 $ id : Factor w/ 396 levels "id10","id100",...: 111 222 333 344 355 366 376 387 1 12 ...
 $ sire : Factor w/ 10 levels "sire1","sire10",...: 1 1 1 1 1 1 1 1 1 ...
 $ sex : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 2 2 1 2 ...
 $ weight: num 335 340 343 287 302 ...
 $ M1 : Factor w/ 18 levels "M1_M1","M1_M2",...: 4 8 8 3 7 2 1 5 5 2 ...
 $ M2 : Factor w/ 18 levels "M1_M1","M1_M2",...: 3 11 2 9 10 2 15 12 7 7 ...
 $ M3 : Factor w/ 18 levels "M1_M1","M1_M2",...: 17 3 18 18 14 12 13 12 16 13 ...
 $ M4 : Factor w/ 18 levels "M1_M1","M1_M2",...: 13 13 8 17 9 9 9 9 11 ...
 $ M5 : Factor w/ 18 levels "M1_M1","M1_M2",...: 7 13 17 8 4 10 11 18 9 11 ...
```

Association Test: Marker M1

$$\hat{Y} = \beta_0 + \beta_1 \times M1_M2 + \beta_2 \times M1_M3 + \beta_3 \times M1_M4 + \dots$$

```
> table(cleandata$M1)
M1_M1 M1_M2 M1_M3 M1_M4 M1_M5 M1_M6 M2_M2 M2_M3 M2_M4 M2_M5 M2_M6
  14    42    34    22    26    14    28    35    38    17    19
M3_M3 M3_M4 M3_M5 M3_M6 M4_M4 M4_M5 M4_M6
  21    24    19    13     7    12    11
```

```
> fit1 <- lm(weight ~ M1, data=cleandata)
```

```
> summary(fit1)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.253e+02	6.763e+00	48.095	<2e-16	***
M1M1_M2	1.771e+00	7.809e+00	0.227	0.8207	
M1M1_M3	2.256e+00	8.035e+00	0.281	0.7790	
M1M1_M4	8.919e+00	8.651e+00	1.031	0.3032	
M1M1_M5	-1.372e+00	8.388e+00	-0.164	0.8702	
M1M1_M6	7.143e-04	9.564e+00	0.000	0.9999	
M1M2_M2	-1.049e+00	8.283e+00	-0.127	0.8993	
M1M2_M3	9.872e+00	8.002e+00	1.234	0.2181	
M1M2_M4	-3.144e-01	7.911e+00	-0.040	0.9683	
M1M2_M5	-2.640e+00	9.132e+00	-0.289	0.7727	
M1M2_M6	1.145e+01	8.913e+00	1.284	0.1998	
M1M3_M3	4.787e+00	8.731e+00	0.548	0.5838	
M1M3_M4	7.156e+00	8.510e+00	0.841	0.4009	
M1M3_M5	1.295e+01	8.913e+00	1.453	0.1472	
M1M3_M6	8.932e+00	9.746e+00	0.916	0.3600	
M1M4_M4	1.955e+01	1.171e+01	1.669	0.0959	.
M1M4_M5	1.065e+01	9.954e+00	1.069	0.2856	
M1M4_M6	6.623e-02	1.020e+01	0.006	0.9948	

Association Test: Marker M1

$$\hat{Y} = \beta_0 + \beta_1 \times M1_M2 + \beta_2 \times M1_M3 + \beta_3 \times M1_M4 + \dots$$

```
> table(cleandata$M1)
M1_M1 M1_M2 M1_M3 M1_M4 M1_M5 M1_M6 M2_M2 M2_M3 M2_M4 M2_M5 M2_M6
  14    42    34    22    26    14    28    35    38    17    19
M3_M3 M3_M4 M3_M5 M3_M6 M4_M4 M4_M5 M4_M6
  21    24    19    13     7    12    11
```

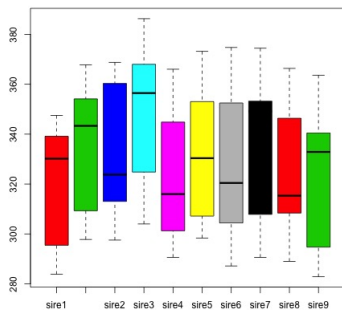
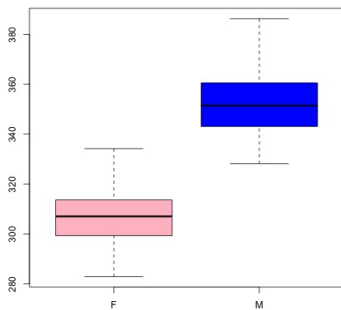
```
> fit1 <- lm(weight ~ M1, data=cleandata)
```

```
> summary(fit1)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.253e+02	6.763e+00	48.095	<2e-16 ***
M1M1_M2	1.771e+00	7.809e+00	0.227	0.8207
M1M1_M3	2.256e+00	8.035e+00	0.281	0.7790
M1M1_M4	8.919e+00	8.651e+00	1.031	0.3032
M1M1_M5	-1.372e+00	8.388e+00	-0.164	0.8702
M1M1_M6	7.143e-04	9.564e+00	0.000	0.9999
M1M2_M2	-1.049e+00	8.283e+00	-0.127	0.8993
M1M2_M3	9.872e+00	8.002e+00	1.234	0.2181
M1M2_M4	-3.144e-01	7.911e+00	-0.040	0.9683
M1M2_M5	-2.640e+00	9.132e+00	-0.289	0.7727
M1M2_M6	1.145e+01	8.913e+00	1.284	0.1998
M1M3_M3	4.787e+00	8.731e+00	0.548	0.5838
M1M3_M4	7.156e+00	8.510e+00	0.841	0.4009
M1M3_M5	1.295e+01	8.913e+00	1.453	0.1472
M1M3_M6	8.932e+00	9.746e+00	0.916	0.3600
M1M4_M4	1.955e+01	1.171e+01	1.669	0.0959
M1M4_M5	1.065e+01	9.954e+00	1.069	0.2856
M1M4_M6	6.623e-02	1.020e+01	0.006	0.9948

What is the average weight for cattles with genotype M1_M1?
M1_M2?

Gender & Sire effect



Which gene is associated with weight?

```
>fit0<-lm(weight ~ sex + sire, data=cleandata)
>fit1<-lm(weight ~ sex + sire + M1, data=cleandata)
>summary(fit3)
>anova(fit0, fit2)
Analysis of Variance Table
Model 1: weight ~ sex + sire
Model 2: weight ~ sex + M1 + sire
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     385 26328
2     368 25370 17    958.11 0.8175 0.6729
```

The p value is $0.673 > 0.05$, hence model 2 is not significantly different from model 1. We conclude marker M1 is not significant.

Which gene is associated with weight?

```
>fit0<-lm(weight ~ sex + sire, data=cleandata)
>fit1<-lm(weight ~ sex + sire + M1, data=cleandata)
>summary(fit3)
>anova(fit0, fit2)
Analysis of Variance Table
Model 1: weight ~ sex + sire
Model 2: weight ~ sex + M1 + sire
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     385 26328
2     368 25370 17    958.11 0.8175 0.6729
```

The p value is $0.673 > 0.05$, hence model 2 is not significantly different from model 1. We conclude marker M1 is not significant.

Exercise: Create the clean data, fit models for all the markers and identify the alleles that are associated with weight.

Which gene is associated with weight?

```
> fit0<-lm(weight ~ sex + sire, data=cleandata)
> fit5<-lm(weight ~ sex + sire + M5, data=cleandata)
> anova(fit0, fit5)
```

Analysis of Variance Table

Model 1: weight ~ sex + sire

Model 2: weight ~ sex + M5 + sire

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	385	26328.0				
2	368	9366.4	17	16962	39.2	< 2.2e-16 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
>fit5<-lm(weight ~ sex + M5 + sire, data=cleandata)
```

```
>summary(fit5)
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	305.860	1.813	168.731	< 2e-16	***
sexM	46.151	0.519	88.921	< 2e-16	***
M5M1_M2	1.505	1.804	0.834	0.4046	
M5M1_M3	2.329	1.699	1.371	0.1713	
M5M1_M4	3.405	1.794	1.898	0.0585	.
M5M1_M5	1.928	1.923	1.003	0.3166	
M5M1_M6	2.232	1.852	1.206	0.2287	
M5M2_M2	-16.271	2.123	-7.665	1.60e-13	***
M5M2_M3	-14.598	1.728	-8.448	6.98e-16	***
M5M2_M4	-14.061	1.795	-7.835	5.08e-14	***
M5M2_M5	-14.299	1.937	-7.382	1.05e-12	***
M5M2_M6	-17.000	2.091	-8.130	6.64e-15	***
M5M3_M3	-14.144	1.869	-7.566	3.12e-13	***
M5M3_M4	-13.574	1.685	-8.053	1.13e-14	***
M5M3_M5	-12.417	1.844	-6.735	6.33e-11	***
M5M3_M6	-14.058	1.788	-7.863	4.21e-14	***
M5M4_M4	-9.444	2.030	-4.652	4.59e-06	***
M5M4_M5	-12.355	1.995	-6.193	1.58e-09	***
M5M4_M6	-10.021	2.393	-4.188	3.53e-05	***
siresire10	14.829	1.181	12.553	< 2e-16	***
siresire2	10.591	1.299	8.151	5.72e-15	***
siresire3	18.050	1.276	14.144	< 2e-16	***
siresire4	6.169	1.200	5.139	4.48e-07	***
siresire5	13.321	1.185	11.237	< 2e-16	***
siresire6	9.559	1.283	7.449	6.77e-13	***
siresire7	12.990	1.193	10.889	< 2e-16	***
siresire8	5.610	1.334	4.205	3.29e-05	***
siresire9	1.308	1.188	1.101	0.2717	

```
>fit5<-lm(weight ~ sex + M5 + sire, data=cleandata)
>summary(fit5)
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	305.860	1.813	168.731	< 2e-16	***
sexM	46.151	0.519	88.921	< 2e-16	***
M5M1_M2	1.505	1.804	0.834	0.4046	
M5M1_M3	2.329	1.699	1.371	0.1713	
M5M1_M4	3.405	1.794	1.898	0.0585	.
M5M1_M5	1.928	1.923	1.003	0.3166	
M5M1_M6	2.232	1.852	1.206	0.2287	
M5M2_M2	-16.271	2.123	-7.665	1.60e-13	***
M5M2_M3	-14.598	1.728	-8.448	6.98e-16	***
M5M2_M4	-14.061	1.795	-7.835	5.08e-14	***
M5M2_M5	-14.299	1.937	-7.382	1.05e-12	***
M5M2_M6	-17.000	2.091	-8.130	6.64e-15	***
M5M3_M3	-14.144	1.869	-7.566	3.12e-13	***
M5M3_M4	-13.574	1.685	-8.053	1.13e-14	***
M5M3_M5	-12.417	1.844	-6.735	6.33e-11	***
M5M3_M6	-14.058	1.788	-7.863	4.21e-14	***
M5M4_M4	-9.444	2.030	-4.652	4.59e-06	***
M5M4_M5	-12.355	1.995	-6.193	1.58e-09	***
M5M4_M6	-10.021	2.393	-4.188	3.53e-05	***
siresire10	14.829	1.181	12.553	< 2e-16	***
siresire2	10.591	1.299	8.151	5.72e-15	***
siresire3	18.050	1.276	14.144	< 2e-16	***
siresire4	6.169	1.200	5.139	4.48e-07	***
siresire5	13.321	1.185	11.237	< 2e-16	***
siresire6	9.559	1.283	7.449	6.77e-13	***
siresire7	12.990	1.193	10.889	< 2e-16	***
siresire8	5.610	1.334	4.205	3.29e-05	***
siresire9	1.308	1.188	1.101	0.2717	

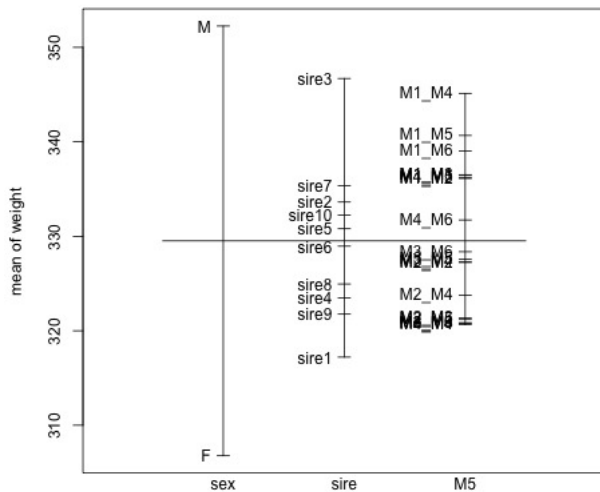
What is the allele strongly associated with weight?

Effect of M1 allele in Marker M5

```
>iM1<-grep("M1", cleandata$M5)
>cleandata$M5[iM1]
>M5M1<-rep(0, length=nrow(cleandata))
>M5M1[iM1]<-1
>fitM5M1<-lm(weight ~ sex + sire +M5M1, data=cleandata)
>summary(fitM5M1)
oefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	292.1434	0.8778	332.827	< 2e-16	***
sexM	46.0481	0.5207	88.443	< 2e-16	***
siresire10	14.3304	1.1634	12.318	< 2e-16	***
siresire2	10.5240	1.2238	8.599	< 2e-16	***
siresire3	18.8832	1.2241	15.426	< 2e-16	***
siresire4	7.1005	1.1649	6.096	2.66e-09	***
siresire5	12.8756	1.1656	11.046	< 2e-16	***
siresire6	9.7480	1.1943	8.162	4.77e-15	***
siresire7	13.6143	1.1646	11.690	< 2e-16	***
siresire8	5.2919	1.2251	4.320	1.99e-05	***
siresire9	0.7412	1.1682	0.634	0.526	
M5M1	15.7699	0.6359	24.798	< 2e-16	***

Effect of M1 allele in Marker M5



Calculating Allele Frequency

```
> M5allele<-strsplit(as.character(cleandata$M5), split="_")
> M5allele<-unlist(M5allele)
> tab<-table(M5allele)/sum(table(M5allele))
> tab
```

M5allele	M1	M2	M3	M4	M5	M6
	0.17550505	0.19949495	0.25126263	0.19823232	0.09595960	0.07954545

Calculating Minor Allele Frequency (MAF)

```
> getMAF<-function(genotype){
+   tmp<-strsplit(as.character(genotype), split="_")
+   tmp2<-unlist(tmp)
+   tabf<-table(tmp2)/sum(table(tmp2))
+   imin<-which(tabf==min(tabf))
+   ans<-c(names(tabf[imin]),tabf[imin])
+   names(ans)<-NULL
+   return(ans)
+ }
>
> getMAF(cleandata$M5)
[1] "M6"    "0.08"
> MAF<-apply(cleandata[,5:9], 2, getMAF)
> MAF
      M1      M2      M3      M4      M5
[1,] "M6"    "M5"    "M6"    "M6"    "M6"
[2,] "0.072" "0.067" "0.077" "0.067" "0.08"
```

The FAMuSS Data

The Functional SNPs Associated with Muscle Size and Strength (FAMuSS) study was conducted to identify the genetic determinants of skeletal muscle size and strength before and after exercise training. A total of $n=1397$ college student volunteers participated in the study and data on 225 SNPs across multiple genes were collected. The exercise training involved students training their non-dominant arms for 12 weeks. The primary aim of the study was to identify genes associated with muscle performance.

Thompson et al. (2004) Functional polymorphisms associated with human muscle size and strength. *Medicine and Science in Sports and Exercise*. 36(7), 1132; 1139.

Calculating MAF for FAMuSS Data

```
> fmsURL<-"http://people.stat.sc.edu/hoyen/STAT588/Data/FMS_data.txt"
> fms<-read.delim(file=fmsURL, header=TRUE, sep="\t")
>str(fms[,1:5])
'data.frame': 428 obs. of 5 variables:
 $ id          : Factor w/ 428 levels "FA-1801","FA-1802",...: 1 2 3 4 5 6 7 8
 $ acdc_rs1501299: Factor w/ 3 levels "AA","CA","CC": 2 2 2 3 2 3 3 3 3 3 ...
 $ ace_id       : Factor w/ 3 levels "DD","ID","II": 1 2 2 1 2 2 3 2 2 2 ...
 $ actn3_r577x  : Factor w/ 3 levels "CC","CT","TT": 1 2 2 2 1 2 3 2 2 1 ...
 $ actn3_rs540874: Factor w/ 3 levels "AA","GA","GG": 3 2 2 2 3 2 1 2 2 3 ...
```

Exercise: Calculate allele frequency and MAF for SNP markers `actn3_rs540874`.