D and D' statistics to measure LD

When we talk about **linkage** we are always considering **two loci**. It is important we keep this in mind. We can take as example positions 82 and 83 from the figure 6.1 from Nielsen and Slatkin book (Figure 1). From now on, the yellow allele in position 82 is going to be denoted as "a" and blue allele as "A"; for position 83 I'm going to call them "B" and "b" respectively.

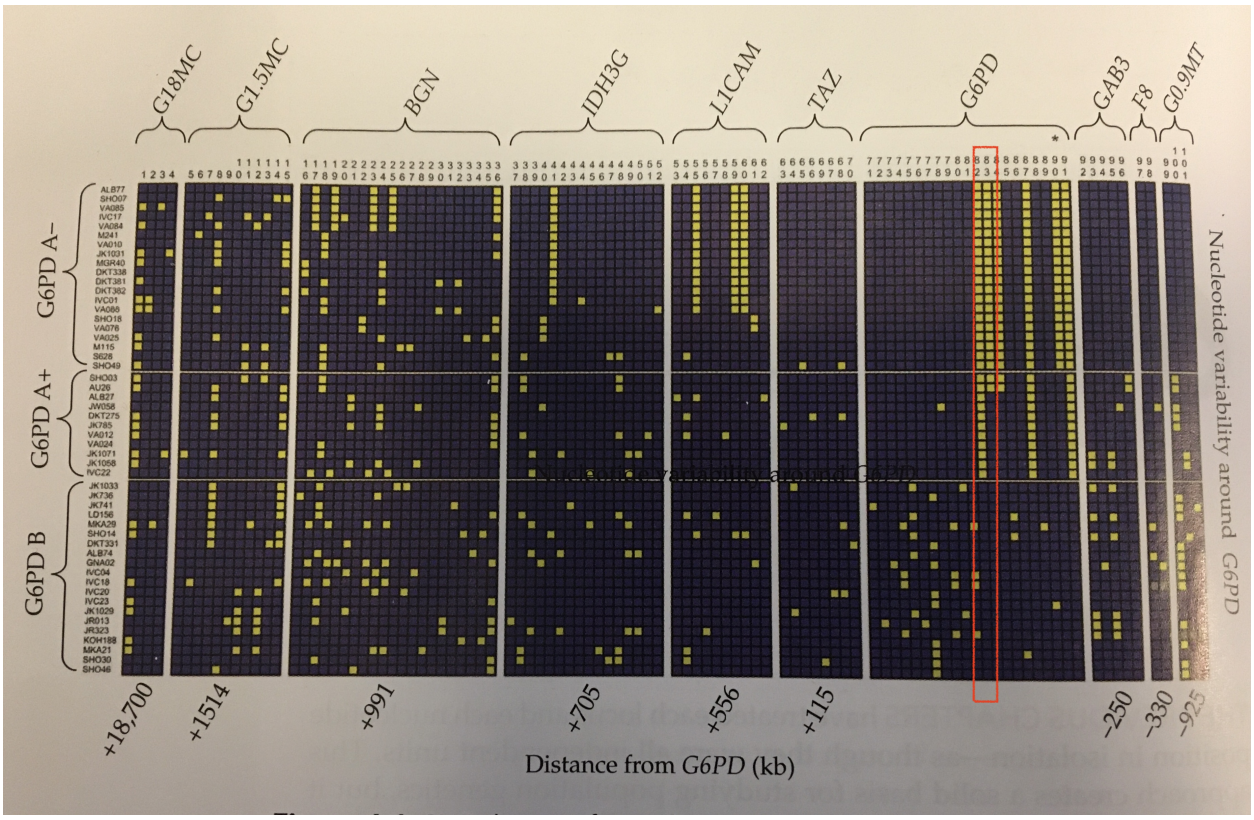


Figure 1. Figure 6.1 from Nielsen and Slatkin book. The two loci highlighted in a red rectangle are the two taken as an example.

Let's start by counting the haplotypes and calculate their frequency:

Haplotypes	Counts	Frequency
AB	22	$f_{AB} = 22/51 = 0.43$
Ab	9	$f_{Ab} = 9/51 = 0.18$
aB	0	$f_{aB} = 0/51 = 0$
ab	20	$f_{ab} = 20/51 = 0.39$

and the allele frequency is:

Allele	Counts	Frequency
A	22+9 = 31	$f_A = 31/51 = 0.61$
a	0 + 20 = 20	$f_a = 20/51 = 0.39$
B	22 + 0 = 22	$f_B = 22/51 = 0.43$
b	9 + 20 = 29	$f_b = 29/51 = 0.57$

If we say that these two loci are linked, it means that when we observe one allele we have a great chance to predict what allele is in the other position. For example, we can see from the haplotype table that A comes together with B, since AB haplotypes have a frequency of 0.43 while Ab has only a frequency of 0.18. Thus, in a relative way, A goes $0.43/0.18 = 2.38$ times more with B than with b. But how can we quantify this linkage?

D to measure LD

This statistic is based on **comparing the observed haplotype frequencies f with the expected allele frequencies f^E under the null hypothesis of independent segregation**. From basic probability, we know that this is calculated by multiplying the frequencies of each allele (similar to Hardy-Weinberg):

$$f^E_{AB} = f_A f_B$$
$$f^E_{Ab} = f_A f_b$$
$$f^E_{aB} = f_a f_B$$
$$f^E_{ab} = f_a f_b$$

Thus, we can compare the observed by just **subtracting the expected frequency from the observed frequency for each haplotype**. This quantity is known as D. For example, we can compute D_{AB} as:

$$D_{AB} = f_{AB} - f^E_{AB}$$
$$D_{AB} = f_{AB} - f_A f_B$$
$$D_{AB} = 0.43 - 0.61(0.43) = 0.168$$

When we only have two alleles, the absolute quantity of D is the same for all haplotypes. You can calculate yourself D for the other haplotypes and you will see that:

$$D_{AB} = 0.168$$
$$D_{Ab} = -0.168$$
$$D_{aB} = -0.168$$
$$D_{ab} = 0.168$$

And the fact that $D_{AB} = D_{ab}$ and $D_{aB} = D_{Ab}$ and $D_{AB} = -D_{Ab}$ can be shown mathematically as:

By definition:

$$D_{AB} = f_{AB} - f_A f_B \ (1)$$
$$D_{Ab} = f_{Ab} - f_A f_b \ (2)$$
$$f_A = f_{Ab} + f_{AB} \ (3)$$
$$f_{AB} = f_A - f_{Ab} \ (4)$$
$$f_b = f_{Ab} + f_{ab} \ (5)$$
$$f_{Ab} = f_b - f_{ab} \ (6)$$
$$f_b = 1 - f_B \ (7)$$
$$f_A = 1 - f_a \ (8)$$

We can show that $D_{AB} = -D_{Ab}$ by:

$$D_{AB} = f_{AB} - f_A f_B \ (starting\ at\ 1)$$
$$f_A - f_{Ab} - f_A f_B \ (using\ 4)$$
$$-f_{Ab} + f_A - f_A f_B$$
$$-f_{Ab} + f_A(1 - f_B)$$
$$-f_{Ab} + f_A f_b \ (using\ 7)$$
$$-(f_{Ab} - f_A f_b)$$
$$D_{AB} = -D_{Ab} \ (using\ 2)$$

If we carry on a bit more, we can show that $D_{AB} = D_{ab}$ by:

$$D_{AB} = -D_{Ab}$$
$$-f_{Ab} + f_A f_b$$
$$-(f_b - f_{ab}) + f_A f_b \ (using 6)$$
$$-(f_b - f_{ab}) + (1 - f_a) f_b \ (using 8)$$
$$-f_b + f_{ab} + f_b - f_b f_a$$
$$f_{ab} - f_a f_b = D_{ab}$$
$$D_{AB} = D_{ab}$$

So, D is going to be a quantity that can only get values between 1 and -1 (theoretically, but we will see soon that this is not true) and which $D_{AB} = D_{ab}$ and $D_{aB} = D_{Ab}$ and $D_{AB} = -D_{Ab}$. In the example above, we found that for this example we get $D = 0.168$, but is this D value high or low?

Intuition tells us that D is going to be constrained by the allele frequencies. For example, if we have that $f_a = 0.01$ and $f_b = 0.01$, D can't be very large since the observed and expected haplotype frequencies can't be very big and the difference is going to be small. However, when $f_a = 0.5$ and $f_b = 0.5$ then D can get larger absolute values.

To illustrate that, I compute in the next R chunk all the possible observed haplotype frequency scenarios and from that I calculate D_{ab} and keep track of f_a and f_b.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
fAB <- c()
faB <- c()
fAb <- c()
fab <- c()
fa  <- c()
fb  <- c()
D   <- c()

for(vfAB in c(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)){
  for(vfaB in c(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)){
    for(vfAb in c(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)){
      for(vfab in c(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)){
        if(round(vfAB+vfaB+vfAb+vfab, digits = 1) == 1){
          fAB <- c(fAB, vfAB)
          faB <- c(faB, vfaB)
          fAb <- c(fAb, vfAb)
          fab <- c(fab, vfab)
          vfa <- vfaB+vfab
          vfb <- vfAb+vfab
          fa  <- c(fa, vfa)
          fb  <- c(fb, vfb)
          D   <- c(D, vfab - (vfa*vfb))
        }
      }
    }
  }
}

data.frame(fAB, faB,fAb,fab,fa,fb,D) %>%
  head(20)
```

```
##    fAB faB fAb fab  fa  fb     D
## 1    0 0.0 0.0 1.0 1.0 1.0  0.00
## 2    0 0.0 0.1 0.9 0.9 1.0  0.00
## 3    0 0.0 0.2 0.8 0.8 1.0  0.00
## 4    0 0.0 0.3 0.7 0.7 1.0  0.00
## 5    0 0.0 0.4 0.6 0.6 1.0  0.00
## 6    0 0.0 0.5 0.5 0.5 1.0  0.00
## 7    0 0.0 0.6 0.4 0.4 1.0  0.00
## 8    0 0.0 0.7 0.3 0.3 1.0  0.00
## 9    0 0.0 0.8 0.2 0.2 1.0  0.00
## 10   0 0.0 0.9 0.1 0.1 1.0  0.00
## 11   0 0.0 1.0 0.0 0.0 1.0  0.00
## 12   0 0.1 0.0 0.9 1.0 0.9  0.00
## 13   0 0.1 0.1 0.8 0.9 0.9 -0.01
## 14   0 0.1 0.2 0.7 0.8 0.9 -0.02
## 15   0 0.1 0.3 0.6 0.7 0.9 -0.03
## 16   0 0.1 0.4 0.5 0.6 0.9 -0.04
## 17   0 0.1 0.5 0.4 0.5 0.9 -0.05
## 18   0 0.1 0.6 0.3 0.4 0.9 -0.06
## 19   0 0.1 0.7 0.2 0.3 0.9 -0.07
## 20   0 0.1 0.8 0.1 0.2 0.9 -0.08
```

I know that for the same fa and fb values I can have different combinations of the haplotype frequencies.

How large can D get for a given fa and fb values? For example, we take the case in which $f_a = 0.2$ and $f_b = 0.4$.

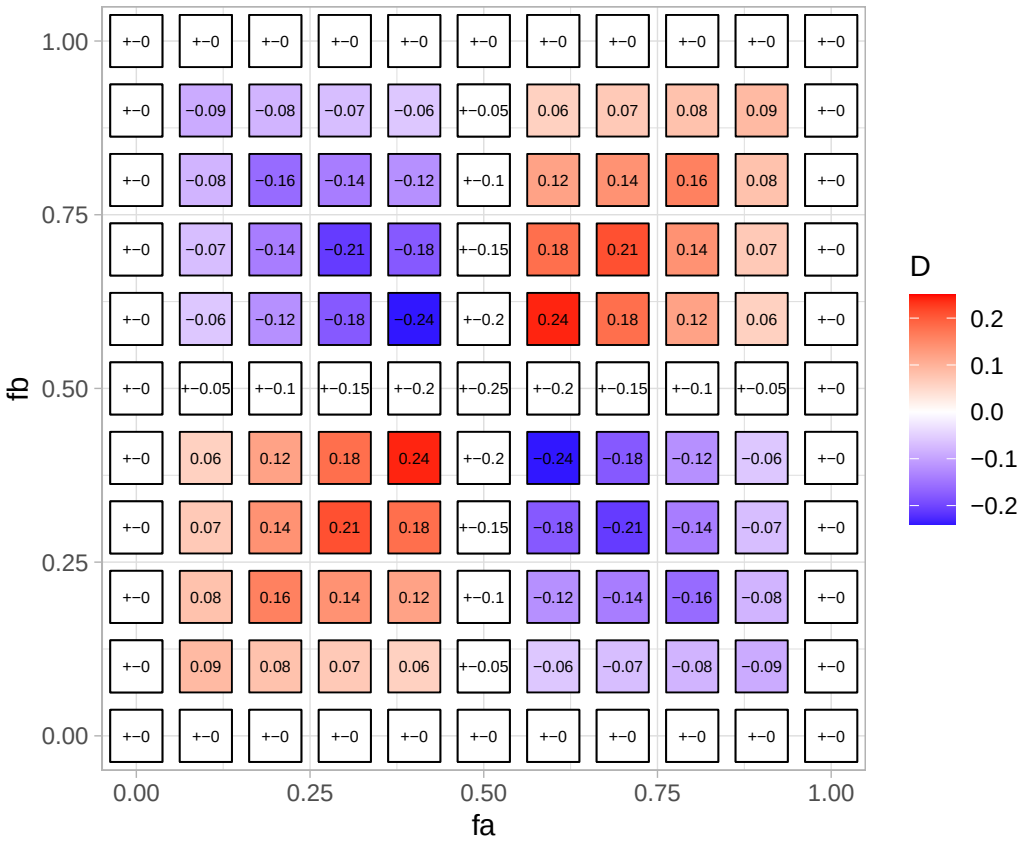
```
data.frame(fAB, faB,fAb,fab,fa,fb,D) %>%
  filter(fa == 0.2, fb == 0.4)
```

```
##   fAB faB fAb fab  fa  fb     D
## 1 0.4 0.2 0.4 0.0 0.2 0.4 -0.08
## 2 0.5 0.1 0.3 0.1 0.2 0.4  0.02
## 3 0.6 0.0 0.2 0.2 0.2 0.4  0.12
```

From that table we've calculated that the range of values for D is [-0.08, 0.12]. So, in this case, the most extreme value that D can take (calculated as $max(max(D_{ab}), abs(min(D_{ab})))$ is 0.12. How is the landscape for all combinations of f_a and f_b?

This is plotted in the next chunk in which I show for a given pair of f_a (x-axis) and f_b (y-axis).

```
data.frame(fAB, faB,fAb,fab,fa,fb,D) %>%
  mutate(fa = round(fa, digits = 2), fb = round(fb, digits = 2)) %>%
  group_by(fa, fb) %>%
  summarize(maxD = round(max(D), digits = 2), minD = round(min(D), digits = 2)) %>%
  mutate(D = ifelse(minD*-1 > maxD, minD, ifelse(minD*-1 < maxD, maxD, maxD)),
         both = ifelse(minD*-1 == maxD, 0, 1)) %>%
  ggplot() +
  geom_point(aes(x = fa, y = fb, fill = D), shape = 22, size = 10) +
  geom_text(aes(x = fa, y = fb, label = round(D, digits = 2)), size = 2) +
  geom_point(data = . %>% filter(both == 0), aes(x = fa, y = fb), fill = "white", shape = 22, size = 10
  geom_text(data = . %>% filter(both == 0), aes(x = fa, y = fb, label = paste("+-", round(D, digits = 2
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_light() +
  theme(aspect.ratio = 1)
```



We can see that there are some constrains. For example, when $f_a = 0.1$ and $f_b = 0.1$ we see that most extreme value of D_{ab} is 0.09, but when $f_a = 0.5$ and $f_b = 0.5$ then D_{ab} can get either positive or negative 0.25. This tells us that **depending of the allele frequency, D is going to be constrained**. We also see that the greatest value D can get is 0.25 or -0.25 (instead of 1 that we though before).

So, if we say that we cheeked two alleles and they are in linkage with $D = 0.168$ are they in high or low LD?

D' to measure LD

What we can do is to compare D to the most extreme value that D can take according to the allele frequencies. But... is there a way to compute the range of values that D can take in an easy way?

The answer is yes! We can put upper and lower boundaries to D with the following reasoning:

We know from before that $D_{AB} = D_{ab}$ and $D_{aB} = D_{Ab}$ and $D_{AB} = -D_{Ab}$. If we assume that $D_{AB} = D$ and that D is positive

$$D = f_{AB} - f_A f_B$$
$$f_{AB} = f_A f_B + D$$

Then:

$$f_{Ab} = f_A f_b - D$$
$$f_{aB} = f_a f_B - D$$
$$f_{ab} = f_a f_b + D$$

Imagine now we don't know D, how positive D can be?Well, if we look at these two formulas we just derived:

$$f_{Ab} = f_A f_b - D$$
$$f_{aB} = f_a f_B - D$$

we see that D can't be bigger than $f_A f_b$ or $f_a f_B$, because otherwise f_{Ab} or f_{aB} are going to be < 0, which can't happen (the frequency of a haplotype can't be negative, right?). Thus, we just defined an upper limit: D can't be bigger than any of $f_A f_b$ or $f_a f_B$. so:

$$D \leq min(f_A f_b, f_a f_B)$$

Now, let's assume that D is negative. Similarly to what we've just done, if we look at:

$$f_{AB} = f_A f_B + D$$
$$f_{ab} = f_a f_b + D$$

and knowing that $D < 0$, let's say $D = -x$ (x>0):

$$f_{AB} = f_A f_B - x$$
$$f_{ab} = f_a f_b - x$$

How negative D can be?

Well, since D can't be bigger than $f_A f_B$ or $f_a f_b$, because otherwise f_{AB} or f_{ab} are going to be < 0, which can't happen. Thus, we just defined an lower limit: $abs(-D)$ can't be bigger than any of $f_A f_B$ or $f_a f_b$. so:

$$D \geq -min(f_A f_B, f_a f_b)$$

You might see that this is different from the formula you have in the book, but you can get the one on the book by:

$$-D \leq min(f_A f_B, f_a f_b)$$

So, from the example we started with, we get that

$$-min(f_A f_B, f_a f_b) \leq D \leq min(f_A f_b, f_a f_B)$$

$$-min(0.26, 0.22) \leq D \leq min(0.168, 0.348)$$

$$-0.22 \leq D \leq 0.168$$

An here is where things get a bit complicated and confused.

On the book it is said:

To describe the extent of LD in a way that takes the range of possible values into account, we define D' to be the ratio of $|D|$ to its **maximum possible value**:

$$D' = \frac{D}{min(f_A f_b, f_a f_B)} \; if \; D > 0$$
$$D' = \frac{-D}{min(f_A f_B, f_a f_b)} \; if \; D < 0$$

Then, I firstly interpret that the maximum possible value in the previous calculated range of values for D is 0.22. But then:

$$D' = \frac{0.168}{0.22} = 0.763$$

which is different from the solution provided by the book $D' = 1$.

Then, I tried to calculate D' using the two formulas provided:

$$D' = \frac{0.168}{0.168} = 1 \;\; if \; D > 0$$
$$D' = \frac{0.168}{0.22} = 0.763 \;\; if \; D < 0$$

So... depending on which formula we use (arbitrarily choosing $D > 0$ or $D < 0$) we get different results for D'; one being correct and the other being incorrect.

To explain why this is I have two explanations:

1 If you read again what the book says, you will see that it says that $|D|$ has to be contrasted to the "maximum POSSIBLE value". Then, I asked myself, is possible that $D = 0.22$? From previous formulas in which we obtained the boundaries for D we know that:

$$f_{aB} = f_a f_B - D$$

So, from the data provided we get that:

$$0 = 0.168 - D$$

Is it possible that $D = 0.22$ NO! because then $0 = -0,052$. But is it possible that $D = 0.168$? Yes! Because there is no other constrain preventing it and $0 = 0$. Therefore, the maximum POSSIBLE value of D is 0.168.

2 There are some details in the derivation of the formulas that we don't get from the book because it is too complicated. One aspect I've found here is that D has an actual definition, that is not just one $D_{AB}, D_{aB}, D_{Ab} or D_{ab}$.

Basically, $D = D_{AB} = D_{ab}$ in the condition that the allele A is the one that has more counts for that particular loci (so A > a) and the same with loci B/b (B>b). Then, the formulas we obtained before about $D > 0$ or $D < 0$ fit our example and everything is perfect!

Overall, what you should know is that:

1. D is the difference between the observed haplotype frequency and the expected (under the null hypotesis of independence of segregation among loci):

$$D_{AB} = f_{AB} - f_A f_B$$

2. D is constrained depending on the allele frequencies (f_a and f_b) and a value of D might mean a lot of linkage for a certain frequency values and might also mean very few linkage for other frequency values.

3. A way to deal with this problem is to correct for the maximum possible value that D can have depending on the frequency values. This is known as D' and can get values from [0-1]. In a way, is a proportion that tell us how much linkage you observe out of the maximum possible linkage you can get.

$$D' = \frac{D}{max(D)}$$

Hope this was helpful!