PLSC 731 – Plant Molecular Genetics

Linkage Disequilibrium

Linkage disequilibrium (LD) is a measure of the degree to which alleles at two loci are associated. LD is also considered the non-random associations between alleles at two loci. Let's define our statistical variables that will allow us to determine if two loci are in linkage disequilibrium or linkage equilibrium. First, here are the frequencies of each haplotype.

Table 1. Definition of haplotype frequencies for two loci with two alleles.

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

From this table, the frequency of each allele at each locus can be calculated. We will use traditional population genetic nomenclature (p and q for allele frequencies at loci A and B) to denote the frequency of each allele.

Table 2. Definition of allele frequencies based on haplotype frequencies.

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} = x_{22}$

To derive a measure for linkage disequilibrium (LD), we can compare the observed and expected frequency of one haplotype. The standard measure of LD is typically

$$D = x_{11} - p_1q_1$$

If two loci are in linkage equilibrium, then

$$D = 0$$

If the two loci are in linkage disequilibrium, then

$$D \neq 0$$

From the definition of D above, we can determine the relationship of haplotype frequencies (Table 1) and D and allelic frequencies (Table 2).

Table 3. Relationships among haplotype and allelic frequencies.

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
$B2$	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	

Since **D** depends on allele frequencies, researchers have suggested that the value should be normalized based on the theoretical maximum and minimum relative to the value of *D*.

So for the case when $D \geq 0$

$$D' = \frac{D}{D_{max}}$$

Where D_{max} is the smaller of p_1q_2 and p_2q_1.

And for the case when $D < 0$

$$D' = \frac{D}{D_{min}}$$

Where D_{min} is the larger of $-p_1q_1$ and $-p_2q_2$.

Finally, another value, the correlation between a pair of loci is calculated using the following formula. This value is *r* or frequently r^2.

$$r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$$

OR

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

The r^2 value is a useful formula because the value ranges from 0 when the loci are in complete linkage equilibrium to 1 when the loci are in complete linkage disequilibrium.

In papers, authors often provide a graph showing the relationship of r^2 to either genetic or physical distance. Typically the r^2 vs. distance is calculated using non-linear regression. Here

are two examples from unpublished work of Chakari et al. It shows the r^2 relationship between SNP loci on the same chromosomes of soybean

2005

Genetic distance (cM)

Year 2005

Physical distance ((Mbp)
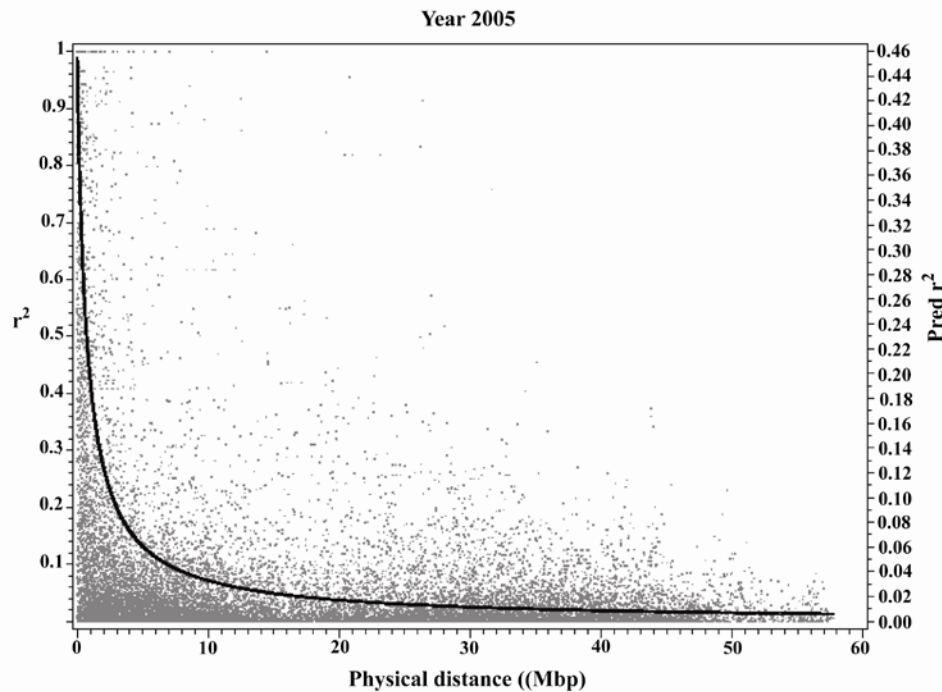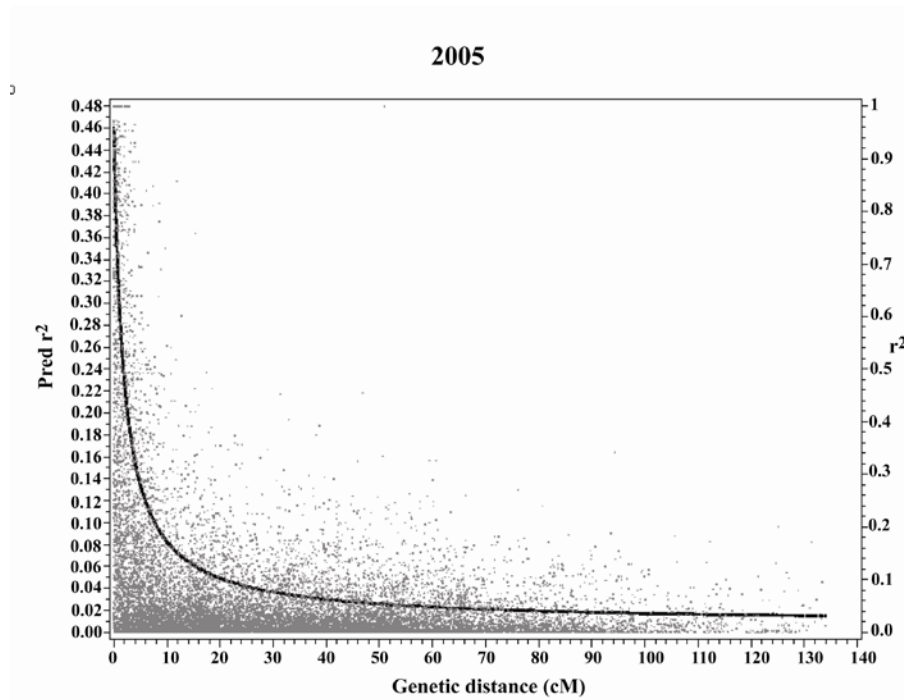
What is not entirely agreed up is the point at which distance, over the full genome, loci are in linkage equilibrium. Various r^2 have been reported as the value where the loci are in linkage equilibrium for example 0.5, 0.2, 0.1, and 0.05. Since there is no clear statistical measure, the graph should be shown and the authors should simply state what r^2 value they are using as a measure of linkage equilibrium.

For a more detailed discussion, see:

Hartl and Clark (1997) Principles of Population Genetics, 3rd edition; pages 95-105.

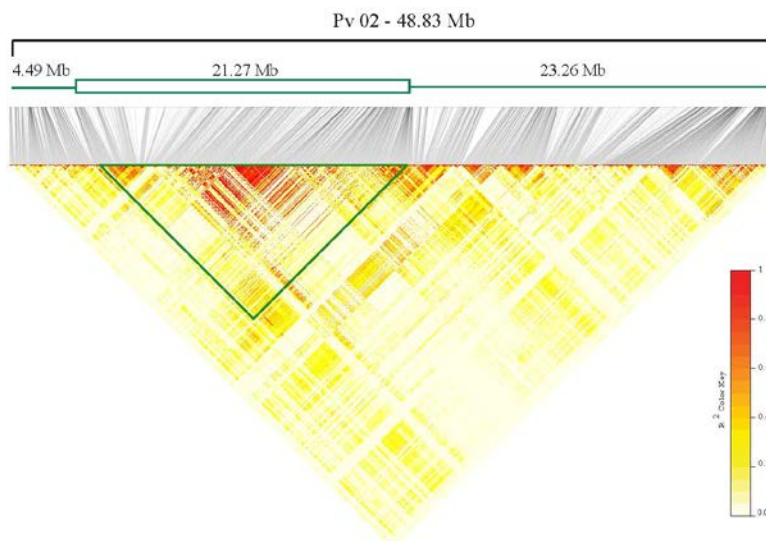**Linkage disequilibrium differs between chromosomes**

Heatmaps
- Better indicator of LD across a chromosome or region
  - Shows the relationship between regions of the genome you are targeting
    - Not an overall average
- The higher the LD (= r^2) the redder the color; all pairwise r^2 values are shown
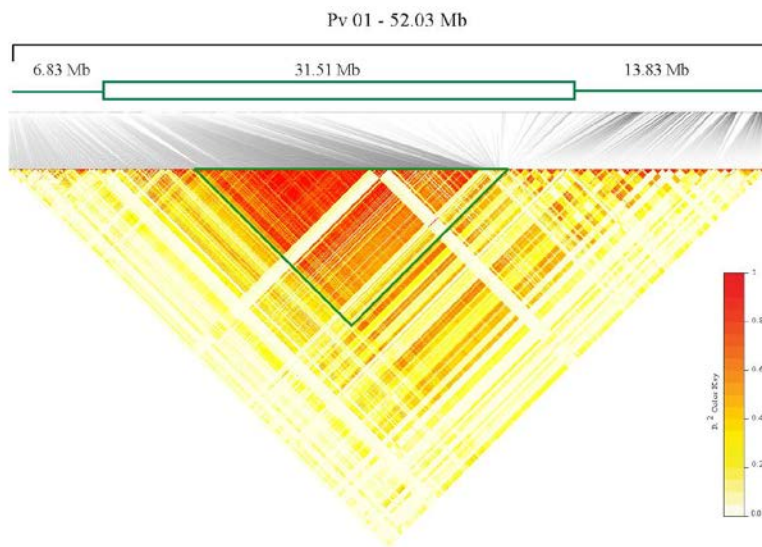
Common Bean chromosome Pv01

Common Bean chromosome Pv02
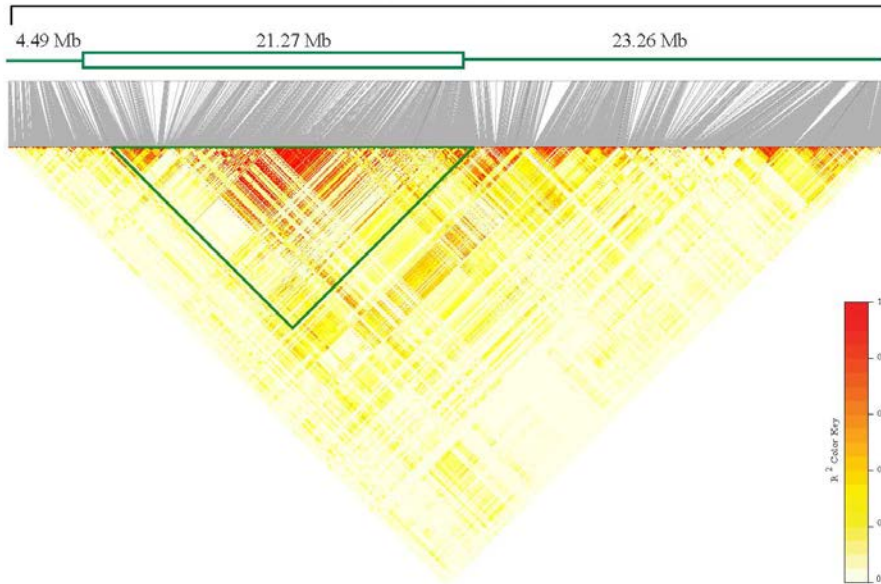
Linkage disequilibrium differs among populations
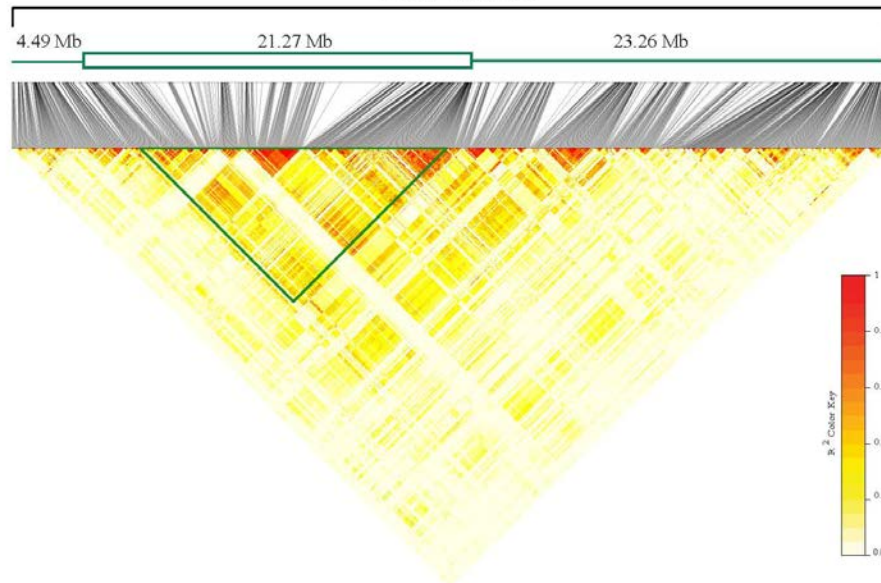
Race Durango and race Mesoamerica of common bean

Race Durango

Race Mesoamerica

Linkage Disequilibrium

Linkage disequilibrium (LD)
- Measures the degree to which alleles at two loci are associated
- The non-random associations between alleles at two loci

Define the statistical variables that will allow us
- To determine if two loci are in
  - Linkage disequilibrium or
  - Linkage equilibrium
- Frequencies of each haplotype are used.

Table 1. Definition of haplotype frequencies for two loci with two alleles.

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

From this table
- The frequency of each allele at each locus can be calculated
  - Using traditional population genetic nomenclature
    - p and q for
      - Allele frequencies at loci A and B.

Table 2. Definition of allele frequencies based on haplotype frequencies.

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} = x_{22}$

To measure linkage disequilibrium (LD)
- Compare the observed and expected frequency of one haplotype
- Standard measure of LD is typically

$$D = x_{11} - p_1 q_1$$

- If two loci are in linkage equilibrium, then

$$D = 0$$

- If the two loci are in linkage disequilibrium, then

$$D \neq 0$$

From the definition of D
- o We can determine
- o The relationship of haplotype frequencies (Table 1) and D and allelic frequencies (Table 2).

Table 3. Relationships among haplotype and allelic frequencies.

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
$B2$	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	

D depends on allele frequencies
- Researchers suggested the value should be normalized
  - o Based on the theoretical maximum and minimum relative to the value of D

- When $D \geq 0$

$$D' = \frac{D}{D_{max}}$$

D_{max} is the smaller of p_1q_2 and p_2q_1.

- When $D < 0$

$$D' = \frac{D}{D_{min}}$$

D_{min} is the larger of $-p_1q_1$ and $-p_2q_2$.

Another LD measure

●   Correlation between a pair of loci is calculated using the following formula
- o Value is r
- o Or frequently r^2.

$$r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$$
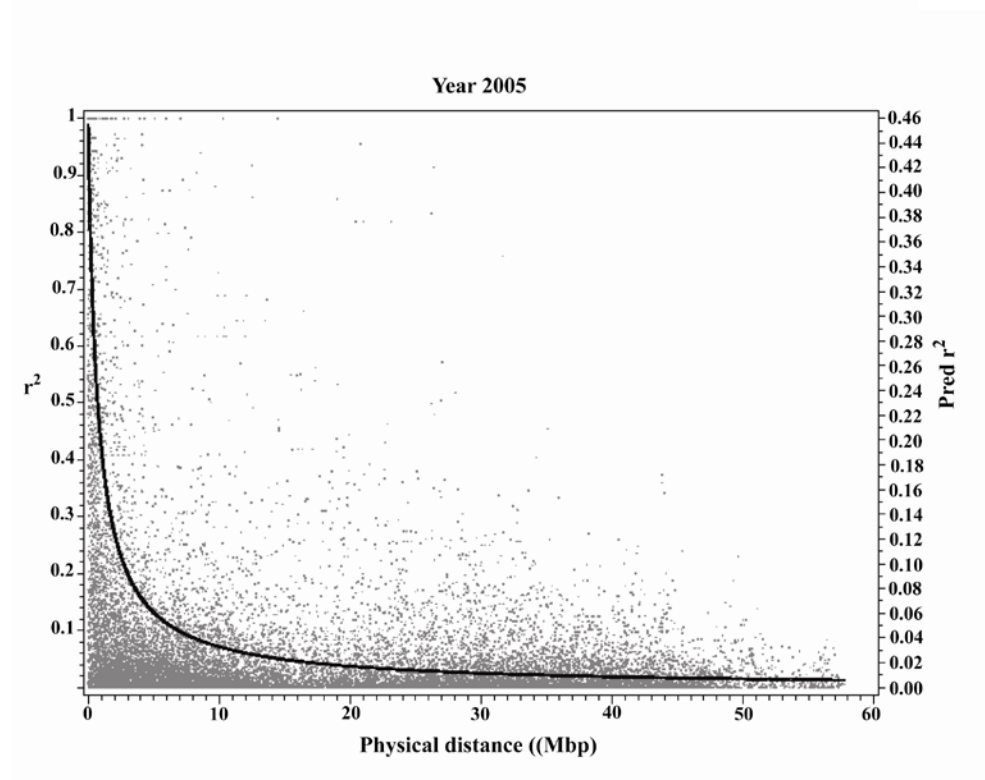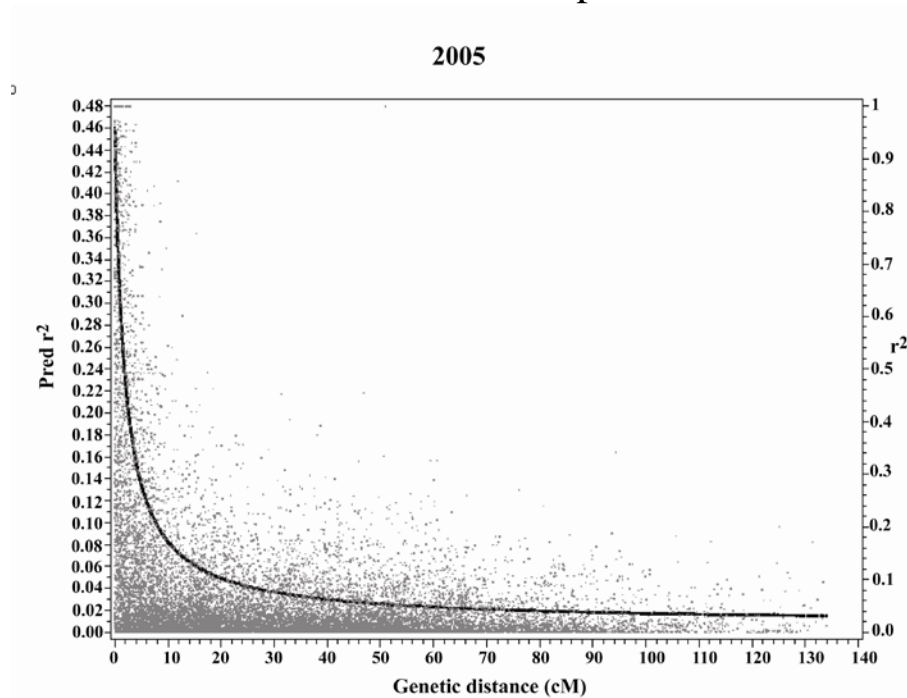
OR

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

r^2 is useful because it ranges from 0 to 1

Ranges from
- o $r^2 = 0$
  - ▪ Loci are in complete linkage equilibrium
- o $r^2 = 1$
  - ▪ Loci are in complete linkage disequilibrium.

Graphical relationship of
- r^2 to either genetic or physical distance
- r^2 vs. distance is calculated
  - Non-linear regression
    - Two examples

**When are loci in linkage equilibrium?**
- Examples
  - 0.5, 0.2, 0.1, and 0.05
    - N clear statistical measure
  - Show graph
  - State r^2 value
    - Use as measure of linkage equilibrium.

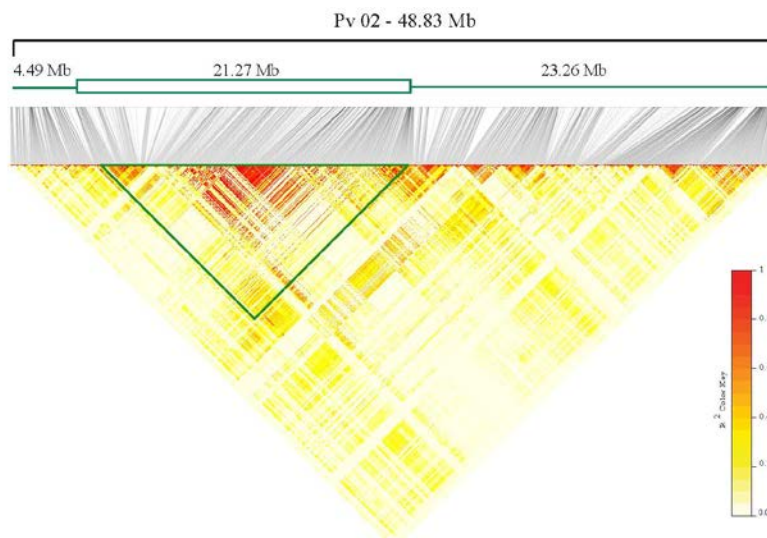# Linkage Disequilibrium Differs Between Chromosomes

Heatmaps
- Better indicator of LD across a chromosome or region
  - Shows the relationship between regions of the genome you are targeting
    - Not an overall average
- The higher the LD (= r^2) the redder the color; all pairwise r^2 values are shown
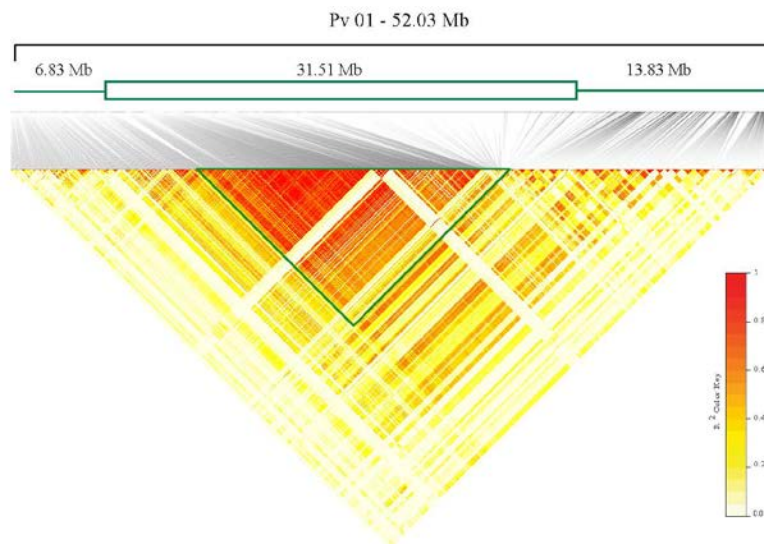
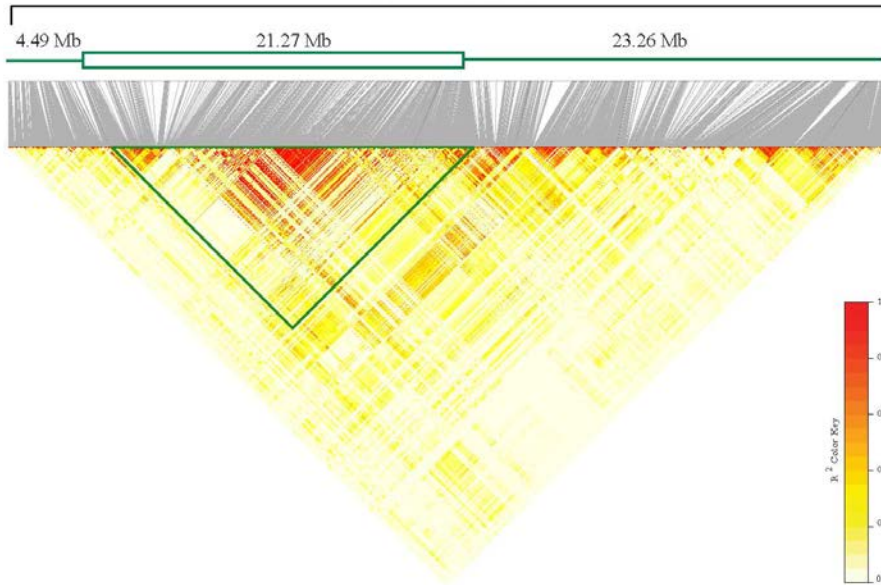## Common Bean chromosome Pv01

Common Bean chromosome Pv02

Linkage Disequilibrium Differs Among Populations

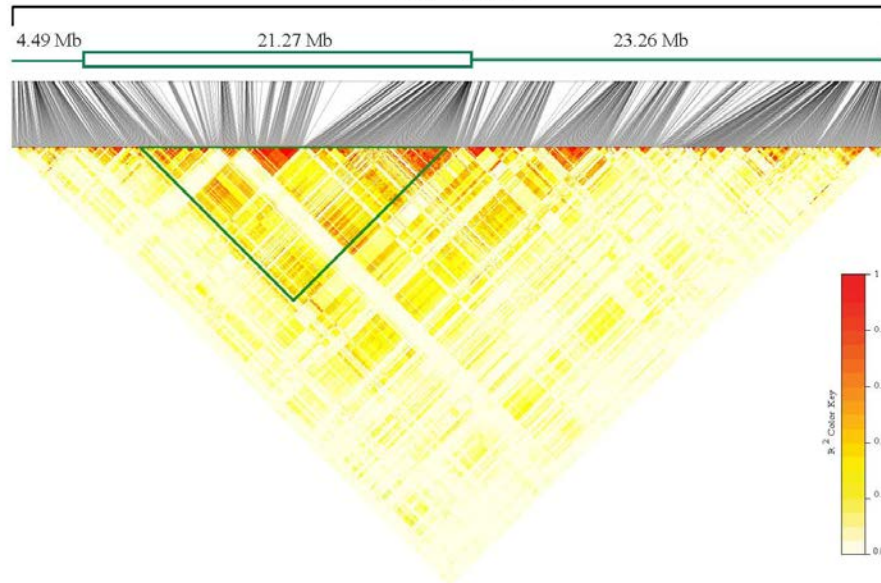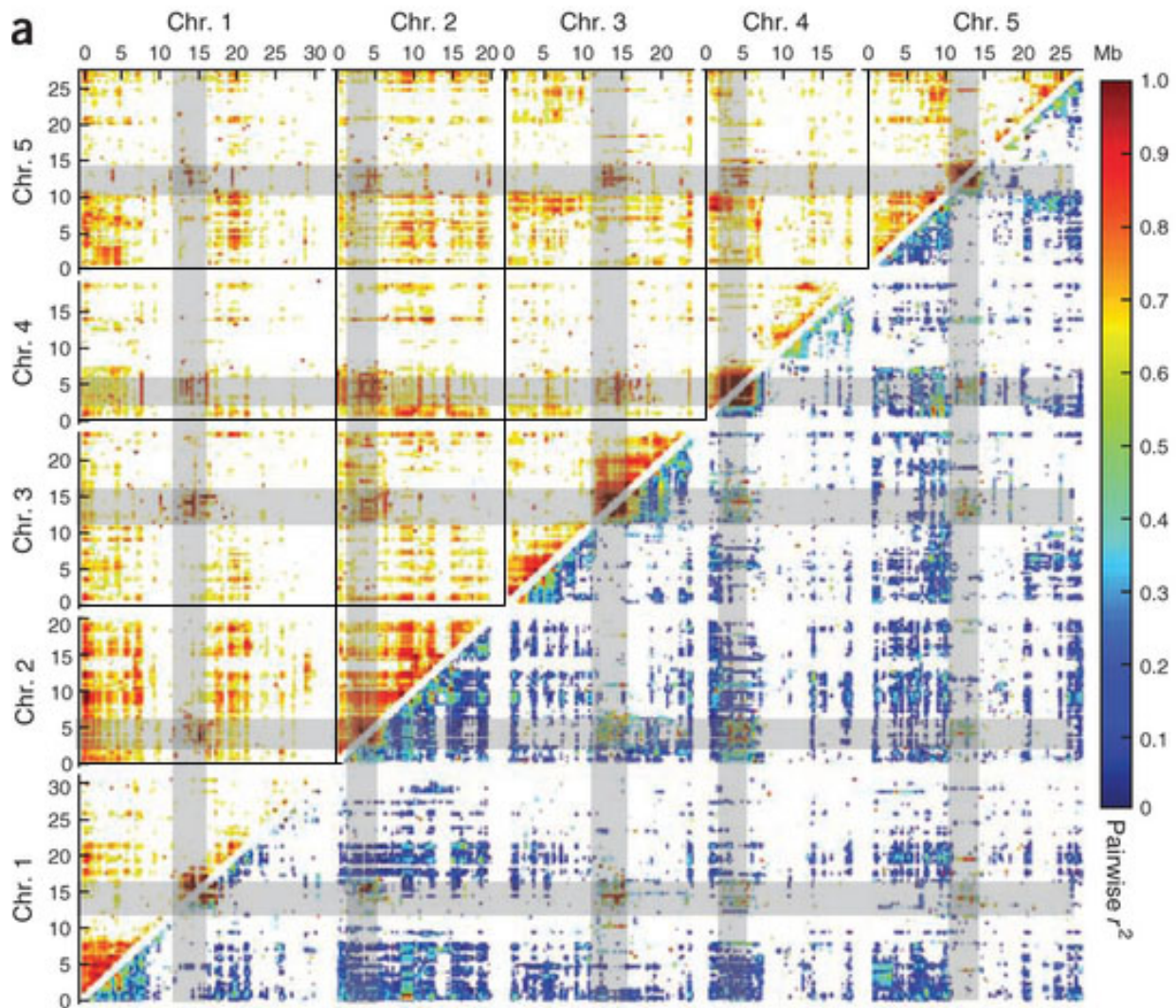# Race Durango and race Mesoamerica of common bean

Race Durango

Race Mesoamerica

(a) Genome-wide pairwise LD. Values before correcting for population structure are shown above the diagonal; for clarity, only values above 0.6 are shown. Values after applying a transformation to reduce theeffects of population structure (related to the correction used in genome-wide association mapping; (Supplementary Note) are shown below the diagonal.