# GBIO0009

# Effects of confounding factors on linkage disequilibrium and GWAS

17.10.2018

Sandra Negro

PLAN

I. Introduction

- Reminder on genetic terms
→ genotype, haplotype, homozygote, etc.

- Reminder on the principal of GWAS and linkage disequilibrium

II. Confounding factors

- Effects on LD
- Effects on GWAS

III. Existing approaches to resolve the confounder effects

# Introduction

## Reminder on genetic terms

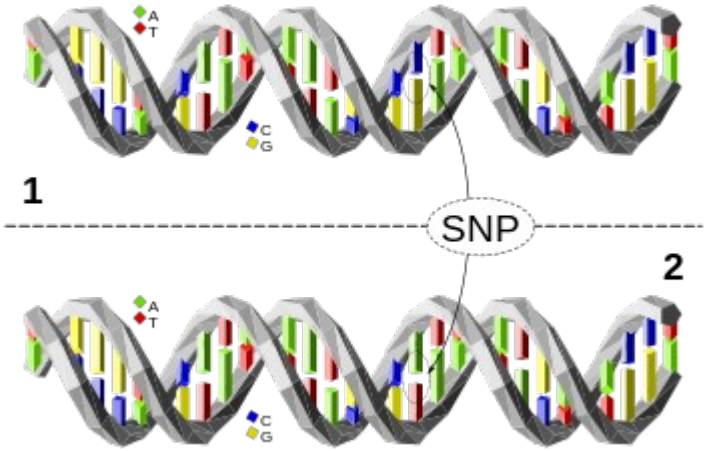
Pair of chromosomes

| | | | | | | |
|---|---|---|---|---|---|---|
| Forward strand | A | A | A | T | - | C |
| Reverse strand | A | T | A | T | - | G |
| Forward strand | A | G | A | T | - | G |
| Reverse strand | A | C | A | T | - | G |

**Genotype** (forward) of a diploïde organism:
AA AG AA TT – CG

homozygote heterozygote

An **haplotype** is a sequence of alleles:
A A A T – C



# Introduction

## Reminder on genetic terms

Example of the type of data in plants :

| | SNP 1 | SNP 2 | SNP 3 | SNP 4 | SNP 5 | SNP 6 | SNP 7 | SNP 8 | SNP 9 |
|---|---|---|---|---|---|---|---|---|---|
| Ind1 | GG | GG | AG | TT | TT | TT | CC | AA | AA |
| Ind2 | AA | GG | GG | TT | -- | -- | CC | AA | CC |
| Ind3 | GG | CC | -- | TT | CC | CC | CC | TT | AA |
| Ind4 | -- | CG | AA | GG | TT | CC | TT | -- | AA |

Alleles of this bi-allelic marker is « G » and « A »

NB :
indels (=insertion-deletion) can be found at certain SNP postions, but it remains a minority
Rarely, SNPs can be tri-allelic or more

In general SNPs are bi-allelic and co-dominant

Rate of heterozygosity in human ~= 95 %
Rate of heterozygosity in inbred lines (plants) ~= 0.3 %

# Introduction

## Reminder on the principal of GWAS

Quantitative Trait Loci (QTL) = genomic regions involved in quantitative effects on the phenotype

Genome-wide **molecular information** can be used together with **phenotypes** to identify the genetic architecture of traits

Two main approaches for the detection of QTL

1) Linkage analyses (QTL approach)
2) **Genome-Wide Association Studies** (**GWAS): highly diverse panels**

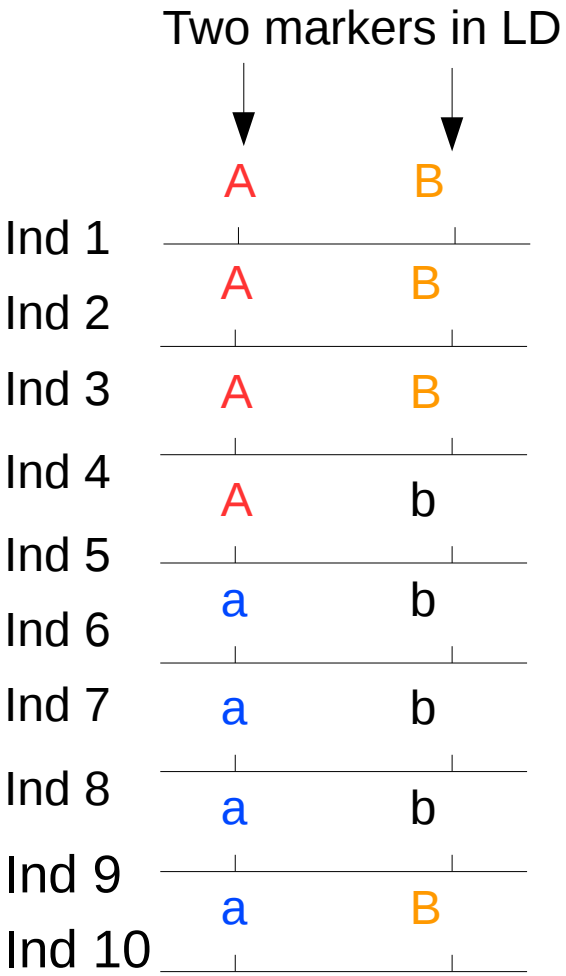
Identified QTLs can be used for e.g. in Marker Assisted Selection (MAS) in plants: Combine favorable alleles

The GWAS approach rely on the fact that molecular markers can capture QTL effect thanks to linkage disequilibrium

# Introduction

## Reminder on Linkage disequilirium

### Linkage Disequilibrium (LD) = non random association between alleles of different loci




Considering these 2 SNPs, the possible haplotypes are: AB, Ab, aB and ab

- Linkage equilibrium :

Haplotype frequency = product of the corresponding allelic frequency

- Linkage disequilibrium refers to the deviation from this equilibrium :

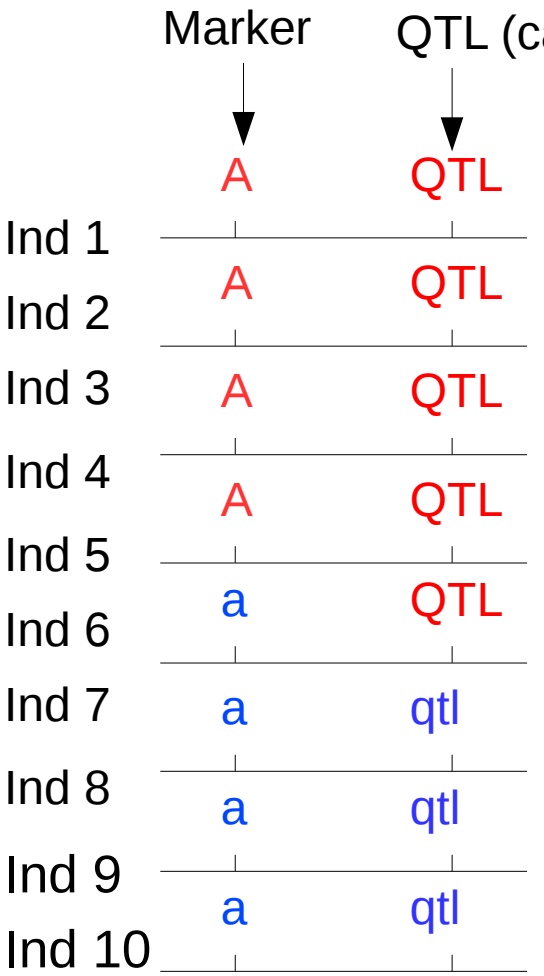
→ $D_{AB} = p_{AB} - p_A p_B$

→ $r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$ One of the LD estimator

# Introduction

## Reminder on Linkage disequilirium

### The GWAS approach rely on the fact that SNPs can capture QTL effect thanks to linkage disequilibrium



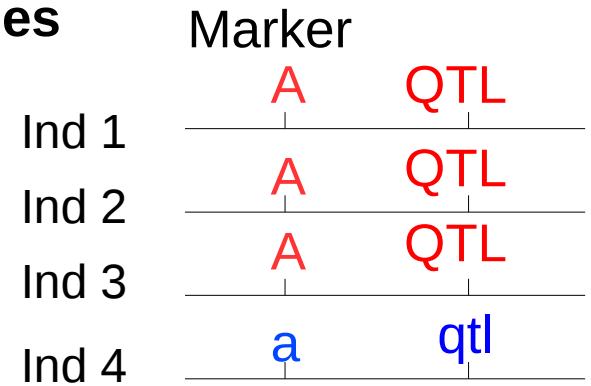

LD can be due to:
1) physical linkage
2) genetic structure in the population
3) kinship between individuals

In genome-wide association studies (GWAS), we are only interested by physical linkage, Population structure and kinship have to be controlled to prevent false positives

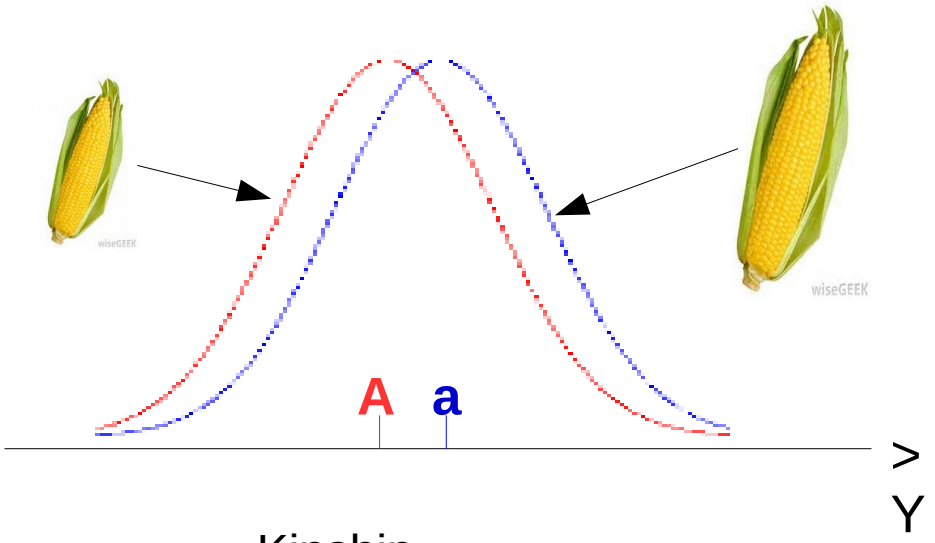
# Introduction

## Reminder on the principal of GWAS and linkage disequilibrium

**Genotypes ($X_l$) :**




**Phenotypes (Y) :**




**Ex. of statistical model: (Yu et al. 2006)**

$$Y = 1\mu + X_l\beta_l + U + E \qquad U \sim N(0, K\sigma^2_{gl}) \qquad E \sim N(0, I\sigma^2_e)$$

SNP effect

Residuals

Kinship

Genetic background

**Is there an association?:** If there is an association between the marker and the phenotype, then the marker is correlated (in LD) to a QTL

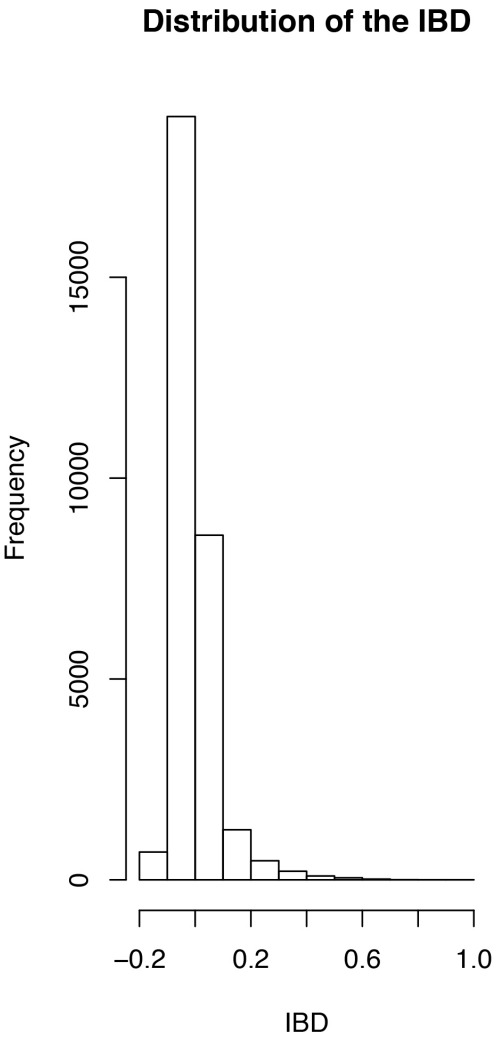
# Confounding factors
## Relatedness

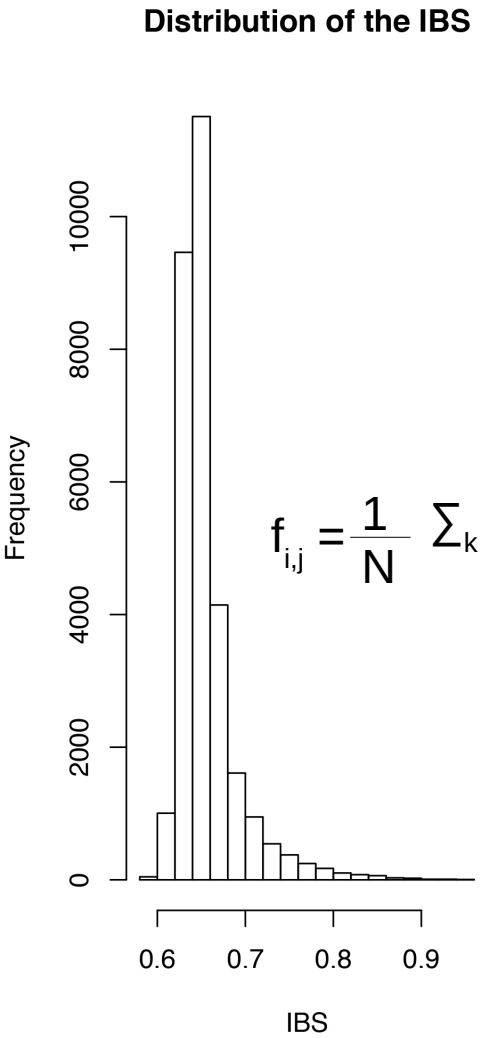

Two estimators (Identity-By-Descent, IBD and Identity-By-State, IBS) can be used to estimate relatedness amongst several individuals.

Individual i, individual j
Locus k, k={1,…, N}
$x_{i,k}$ : genotype of $i^{th}$ ind. at the $k^{th}$ locus, coded as 0 for the 1st allele and as 1 for the second allele
$p_k$ : frequency of the « + » allele
$f_{i,j}$ : kinship coefficient

$$f_{i,j} = \frac{1}{N} \sum_k \frac{(x_{i,k} - p_k)*(x_{j,k} - p_k)}{(p_k * (1 - p_k))}$$

(Astle and Balding 2009)




$$f_{i,j} = \frac{1}{N} \sum_k x_{i,k} * x_{j,k}$$

# Confounding factors
## Population structure

Example of population struture in maize using ADMIXTURE

6 groups identified by Admixture (6 ancestral fractions per ind.)
It is important to understand the organization of genetic diversity within a panel used in association genetics to define statistical models, to analyze the relationship between genetic polymorphism and the variation of traits, and to define the density of markers that is suitable to localize causal polymorphisms.

Lancaster Oh43

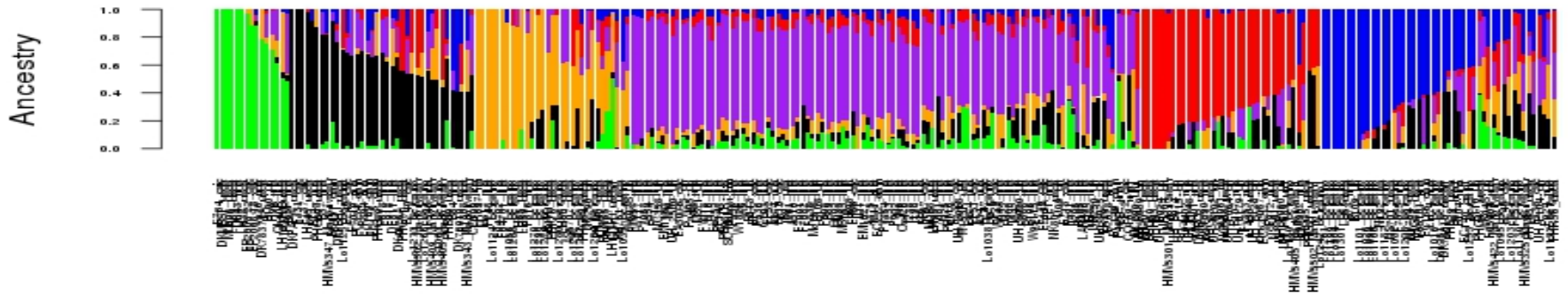
Stiff Stalk B14a

Lancaster Mo17

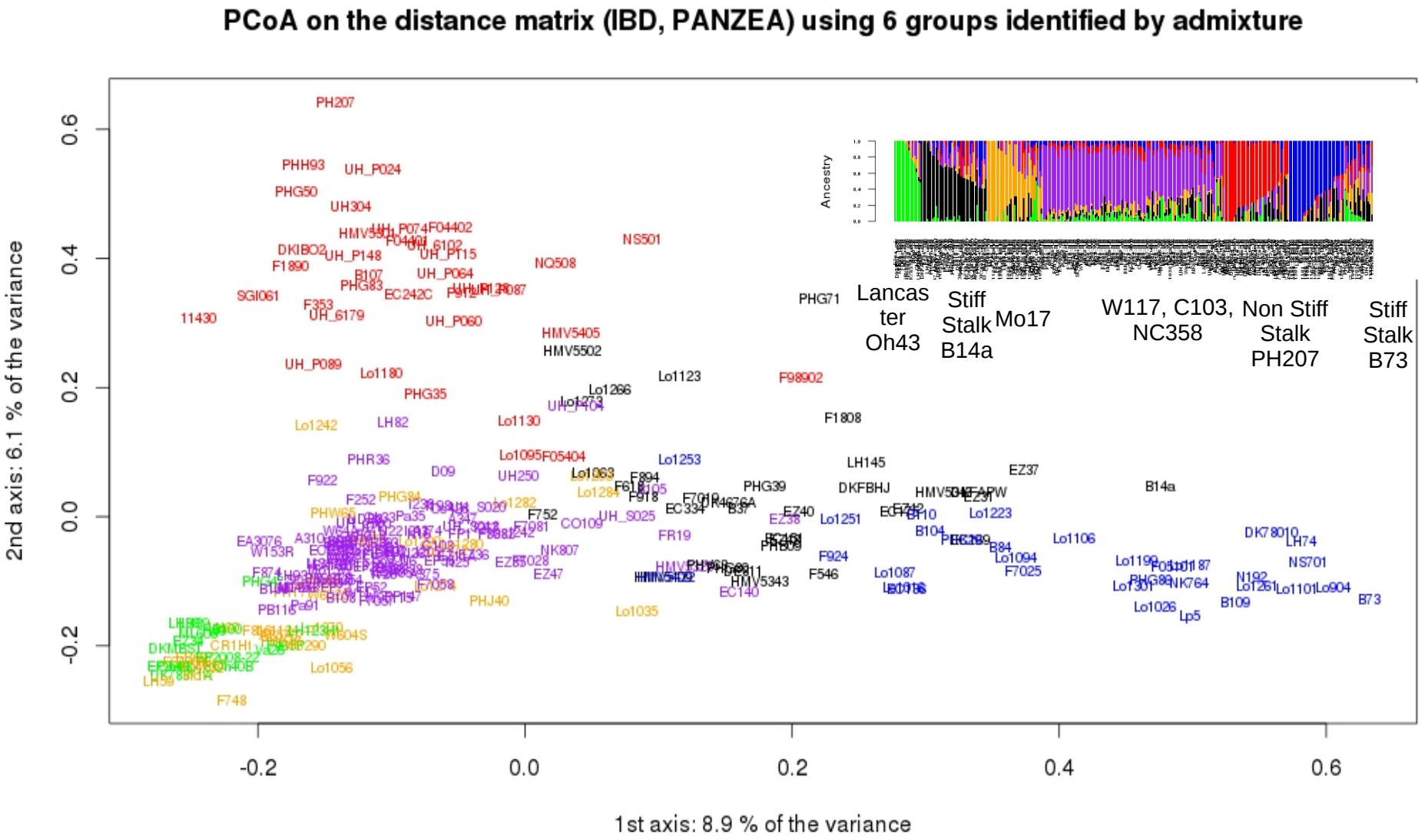
W117, C103, NC358

Non Stiff Stalk PH207

Stiff Stalk B73



# Confounding factors




- The two first PCoA axes explain an important part of the variance (15%)
- Major genetic groups are visualized
- Genetic groups and PCoA estimates are in accordance with breeder's knowledge

# Confounding factors
## Effects on the LD and GWAS

The presence of ind. from different pop with different genetic origins within a sample can produce **LD between unlinked** loci, simply because of differences of allelic frequencies
→ Hence, such structured sample can lead to a bias estimate of LD
⇒ which may increase the rate of false positive in GWAS
⇒ which could lead to inappropriate choice of marker density and therefore to a decreased power

A biased estimate of LD is also obtenained when genotyped ind. are not independent

!!! All the analyses using $r^2$ rely on the assumption that the extent of $r^2$ around the causal polymporphism depends only a drift-recombination process in a random mating pop without selection !!! → But it is not always the case in real life data

# Confounding factors

## Existing approaches to resolve the confounder effects

Structure and relatedness can be corrected for in diversity analyses and association mapping (GWAS) to take into account the non-indepence of loci due to both population differenciation and uneven levels of relatedness.

⇒ Read paper provided (Mangin et al. 2012) to have details on the new adjusted $r^2$

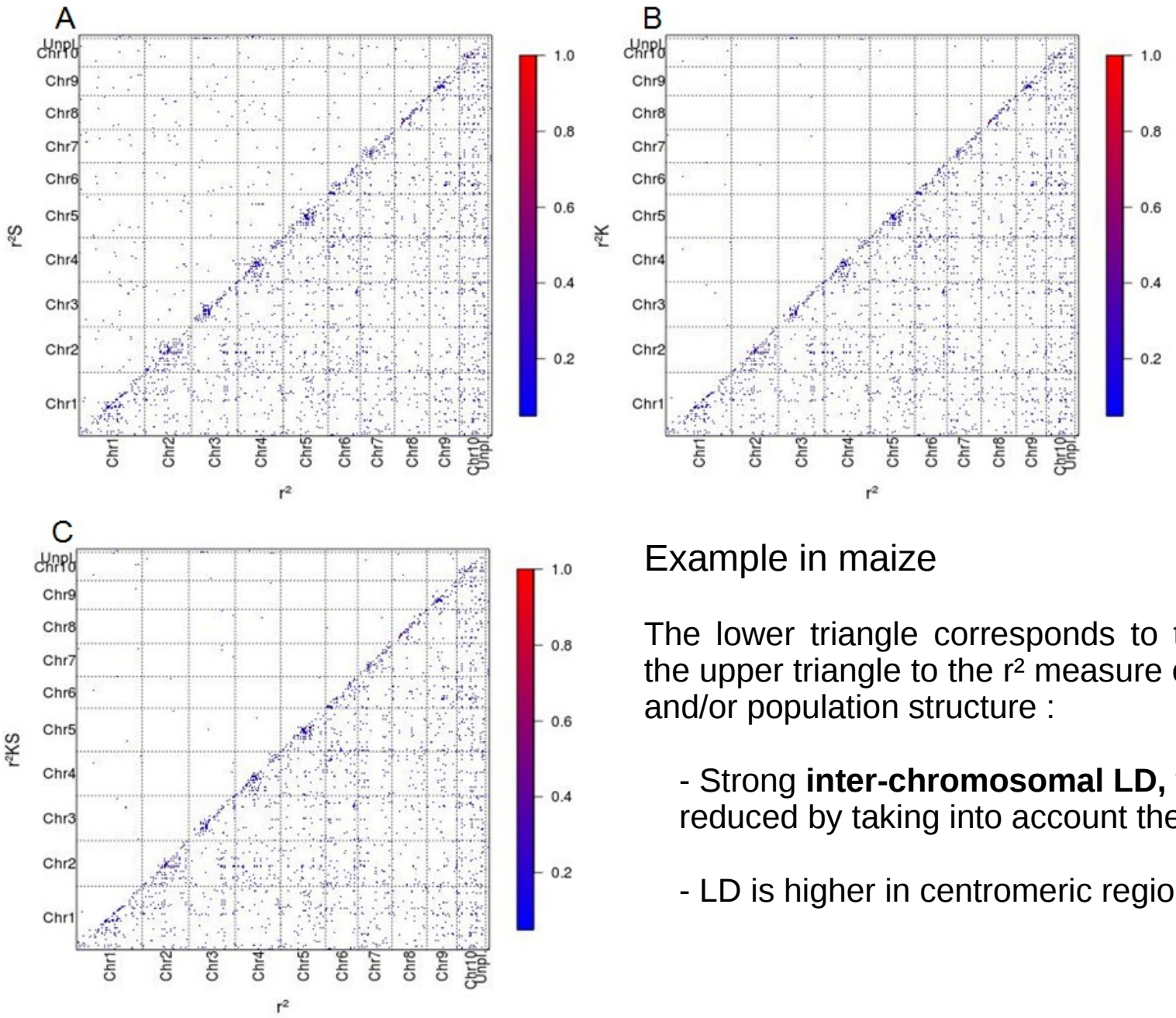
Real life data and simulation analyses highlighted:

1) in a two-population structured sample;
$r^2$ bias increases with the differentiation of loci and with the decrease of LD

2) in a highly related sample;
$r^2$ overestimate the true LD value

# Confounding factors

## genome-wide linkage desequilibrium between all loci within and between chromosomes : $r^2$ *versus* adjusted $r^2$




Example in maize

The lower triangle corresponds to the $r^2$ measure and the upper triangle to the $r^2$ measure corrected for kinship and/or population structure :

- Strong **inter-chromosomal LD,** which is much reduced by taking into account the **kinship**

- LD is higher in centromeric regions

# Confounding factors

## Existing approaches to resolve the confounder effects

### Statistical approaches to resolve the confounder effects :

Four statistical models can be tested to limit false positive

Model 1: without correction; $Y \sim X\beta + E$

Model 2: takes into account group structure; $Y \sim Qs + X\beta + E$

Model 3: takes into account kinship between individuals; $Y \sim X\beta + Ku + E$

Model 4: takes into account both group structure and kinship; $Y \sim Qs + X\beta + Ku + E$

Home-made scripts or different informatic tools (EMMAX, ASReml, FASTLMM, etc...) can incorporate these models.