Resemblance among relatives. Pedigree vs. Genomic based ...

Evangelina López de Maturana & Oscar González-Recio

Topics

Quantitative genetics

Resemblance among relatives

Genomic resemblance

GRM examples

Why assessing resemblance is important?

How to estimate it?

Kinship coefficient

Inbreeding coefficient

Additive relationship coefficient

Moving to genomics

SNPready R package

- VanRaden (2008)
- Yang (2010)
- Yang’s modified
- Gaussian kernel

Quantitative genetics

Linking genotypes and phenotypes through genetic similarity among individuals = covariance between relatives (Wright,1921) is a fundamental concept in quantitative genetics

Main focus nowadays is to statistically model variation in DNA sequences affecting phenotypic variation in quantitative traits

Less interest in understanding the biological pathways (molecular genetics domain)

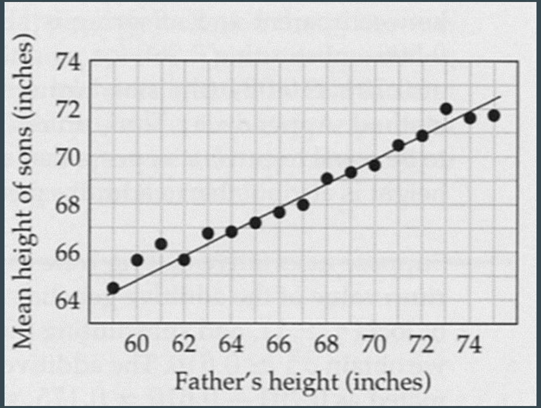
Genome-based prediction aims to predict unobserved values by regressing phenotypes on measures of genetic resemblance, based on DNA data

Resemblance among relatives

To calculate the resemblance among relatives *x* and *y*:

$$cov(P_x, P_y) = cov(G_x + E_x, G_y + E_y) = cov(G_x, G_y) + cov(E_x, E_y)$$

Heredity seems to act in a linear manner



Resemblance among relatives

Degree of relationship between two related individuals is:

- The probability that a gene in one subject is identical by descent to the corresponding gene (i.e., in the same locus) in the other individual
    - Identical by descent (IBD): both copies of the same ancestral gene
    - Identical by state (IBS): identical through separate mutations

Measurements

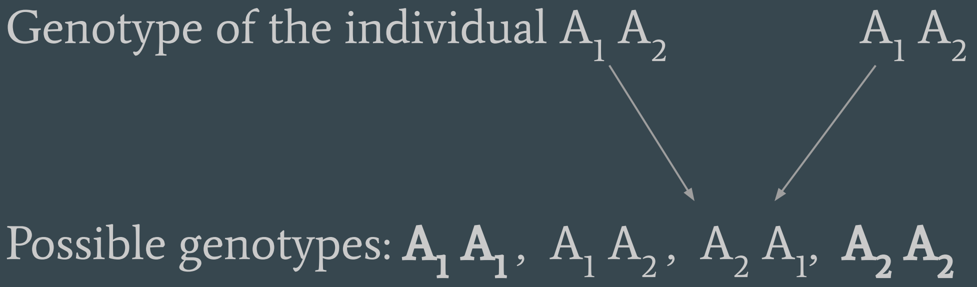
Kinship coefficient between two individuals ($f_{x,y}$) is:

- a simple measure of relatedness
- defined as the probability that a pair of randomly sampled homologous alleles are IBD
- indicates the probability that an individual receives the same allele from both parents because they are related (= **inbreeding coefficient**)
  - $f_{x,y}$ = F_{animal} = ½ $a_{between\ parents}$, where $a_{between\ parents}$ is the additive relationship coefficient between parents

$$F_X = \sum \left(\frac{1}{2}\right)^n (1 + F_A)$$

Kinship coefficient (autogamy)

$f_{AA} = \frac{1}{2}\left(1 + F_A\right)$, where F_A is the inbreeding coefficient of ind A



Genotype of the individual A_1 A_2 A_1 A_2

Possible genotypes: $\mathbf{A_1}$ $\mathbf{A_1}$, A_1 A_2, A_2 A_1, $\mathbf{A_2}$ $\mathbf{A_2}$

Relationship between the kinship coefficient and the relationship coefficient

	Kinship coefficient	Additive relationship
Sire-daughter	0.25	0.5
Grandsire-daughter	0.125	0.25
Full sibs	0.25	0.5
Half sibs	0.125	0.25

Resemblance among relatives

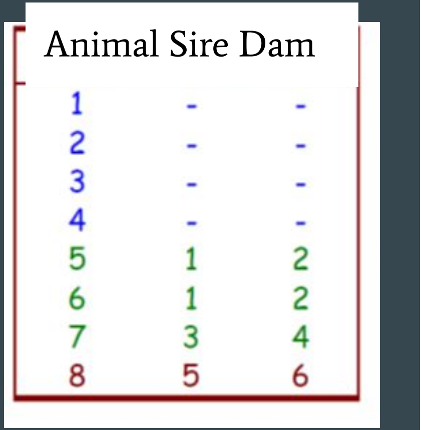
Considering the additive genetic variance:

$$cov\left(A_i, A_i\right) = a_{ii}\sigma_a^2 \;; \quad a_{ii} = 1 + F_A$$

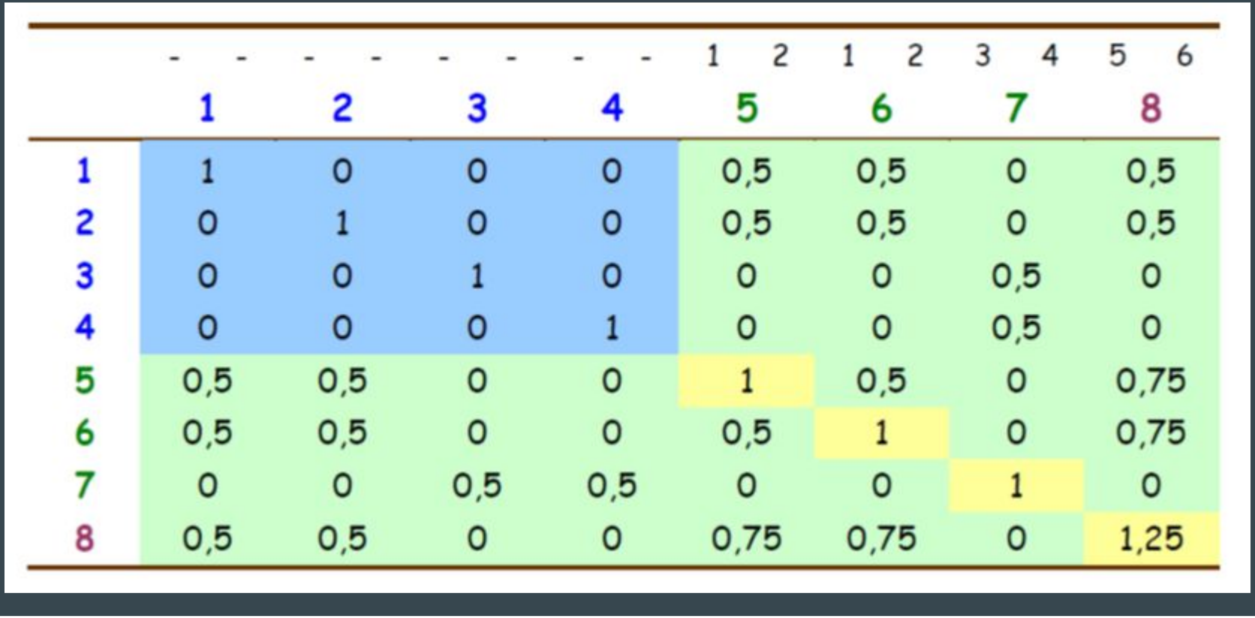
$$cov\left(A_i, \phi_i\right) = 0$$

$cov\left(A_i, A_j\right) = 0$, if i and j are not related

Numerator relationship matrix



Animal	Sire	Dam
1	-	-
2	-	-
3	-	-
4	-	-
5	1	2
6	1	2
7	3	4
8	5	6



	- -	- -	- -	- -	1 2	1 2	3 4	5 6
	1	2	3	4	5	6	7	8
1	1	0	0	0	0,5	0,5	0	0,5
2	0	1	0	0	0,5	0,5	0	0,5
3	0	0	1	0	0	0	0,5	0
4	0	0	0	1	0	0	0,5	0
5	0,5	0,5	0	0	1	0,5	0	0,75
6	0,5	0,5	0	0	0,5	1	0	0,75
7	0	0	0,5	0,5	0	0	1	0
8	0,5	0,5	0	0	0,75	0,75	0	1,25

BLUP

Model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{I}\alpha + \epsilon$$

MME:

$$\begin{bmatrix} X^T X & X^T \\ X & I + A^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} X^T y \\ y \end{bmatrix}$$

A: Pedigree-based relationship matrix

A^{-1}

We can compute it directly using the Henderson’s rules:

Animal model

- Both parents are known:
    - We add 2 in the position (i,i) of the matrix
    - We add 1 in the position (s,i), (i,s), (d,i), (i,d) of the matrix
    - We add ½ in the position (s,s), (s,d), (d,s), (d,d) of the matrix
- Only one parent is known:
    - We add 4/3 in the position (i,i) of the matrix
    - We add -2/3 in the position (s,i), (i,s), (d,i), (i,d) of the matrix
    - We add ⅓ in the position (s,s), (s,d), (d,s), (d,d) of the matrix
- Both parents are unknown:
    - We add 1 in the position (i,i) of the matrix

Moving to genomic resemblance

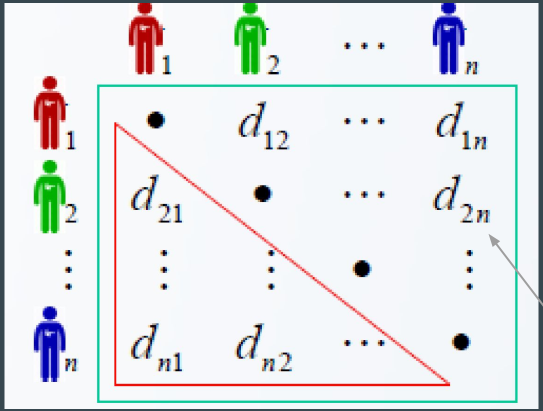
Genome-based prediction can be considered a field in the quantitative genetics area aiming to predict unobserved values by regressing phenotypes on measures of genetic resemblance obtained from germline DNA genotypes

Early attempts in the 80’s with few molecular information

Meuwissen et al (2001) paved the way in the joint use of whole-genome markers for genomic prediction

Gianola et al (2003) were pioneers in considering the resemblance of individuals at the genomic level

Genetic (genomic) resemblance



$$\begin{array}{c|cccc} & 1 & 2 & \cdots & n \\ \hline 1 & \bullet & d_{12} & \cdots & d_{1n} \\ 2 & d_{21} & \bullet & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \bullet & \vdots \\ n & d_{n1} & d_{n2} & \cdots & \bullet \end{array}$$

$$K_h(x_i, x_j) = f(h^{-1}dist(x_i, x_j))$$

GBLUP

Model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{I}\alpha + \epsilon$$

MME:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T \\ \mathbf{X} & \mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_\epsilon^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{y} \end{bmatrix}$$

Genome-based relationship matrix

GBLUP

G has a covariance structure for the genetic values of the i-th and j-th individuals

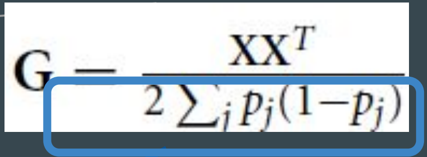
$$\mathrm{Cov}\left(u_i, u_j\right) = \sigma_a^2 K_h\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\ (\sigma_a^2 > 0)$$

GRM examples - VanRaden et al, 2008 (1)

Linear kernel

$x_{ij} - 2p_j$

X is an incidence matrix with the genotypes of each individual for each SNP
Dimension: # ind x # SNPs



$$\mathbf{G} = \frac{\mathbf{X}\mathbf{X}^T}{\boxed{2\sum_j p_j(1-p_j)}}$$

It scales **G** to be analogous to the numerator relationship matrix **A**

It is assumed that the marker variance is homogeneous

GRM examples - VanRaden et al, 2008 (1)

Genomic inbreeding coefficient for ind *j*: can be obtained as G_{jj}-1

Genomic relationships between individuals *j* and *k* (analogous to the relationship coefficients of Wright (1922)) can be obtained as G_{jk}

$$\frac{G_{jk}}{\sqrt{G_{jj}}\sqrt{G_{kk}}}$$

GRM examples - VanRaden et al 2008 (2)

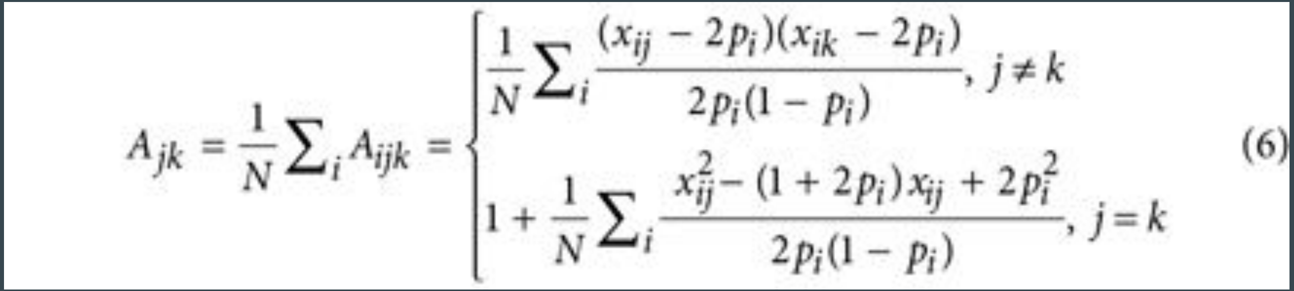
$x_{ij} - 2p_j$

$$\mathbf{G} = \mathbf{ZDZ'},$$

$$D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$$

GRM examples - Yang et al, 2010

The same as Vanraden (2)



$$A_{jk} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, & j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2 - (1 + 2p_i)x_{ij} + 2p_i^2}{2p_i(1 - p_i)}, & j = k \end{cases} \tag{6}$$

GRM examples: Gaussian kernel

Non-linear kernel

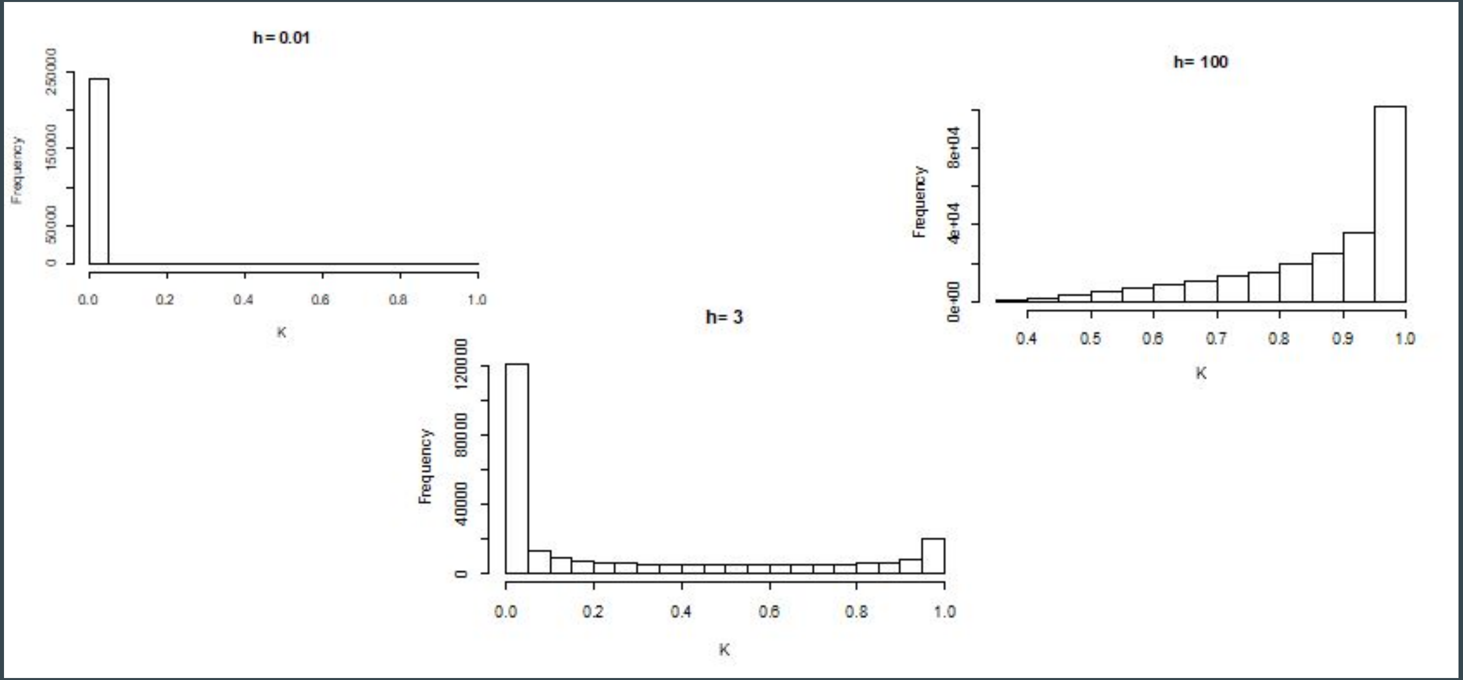
It can capture small complex interactions and non-additive variation (de los Campos, et al, 2010)

$$K_h(x_i, x_j) = \exp(-\frac{(x_i - x_j)^2}{h})$$

When the individuals are related, the value is close to 1. Otherwise, close to 0

h is the tunning parameter

Gaussian kernel



Gaussian kernel

Pérez-Elizalde et al (2015) shows how to select the Bandwidth Parameter (h or scale parameter) in a Bayesian Kernel Regression Model

This strategy was in general superior to the kernel averaging strategy proposed by de los Campos et al (2010), based on defining a set of kernels based on different values of h

GRM examples: Speed's GRM (LDAK)

A method for weighting markers to account for LD

It scales SNP genotypes according to local patterns of LD

It computes optimal SNP weights considering local SNP correlation caused by LD

$$\mathbf{G}_S = \frac{\mathbf{WW}'}{\sum_{j=1}^{m} k_j}$$

$$w_{ij} = \sqrt{k_j \bar{z}_{ij}}$$

$$\bar{z}_{ij} = \frac{z_{ij}}{\sqrt{2p_j\left(1-p_j\right)}}$$

k_j is the weighting factor of the j-th SNP (LDAK determines SNP weightings so that the sum of the values in row i times the SNP weightings equals (approximately) one)

Some considerations when building GRM

Matrix G may be singular, for example, if the number of markers does not exceed the number of individuals genotyped

A simple solution could be to add a small number (i.e., 0.00001) to diagonal elements of each GRM to avoid near singularity problems

Some considerations when building GRM

Sub-population or ancestry-related **positive assortative mating** (Risch et al., 2009; Sebro et al., 2010) results in **population stratification**, and is seen at all loci where the allele frequency differs between sub-populations.

Although there is no genetic correlation between spouses (random mating) within sub-populations, when the entire stratified population is considered, there is a significant positive genetic correlation between spouses, denoted by Wright's coefficient of inbreeding F.

There is **increased genetic covariance between relatives** in the **presence of population stratification**.

Some considerations when building GRM

When QTLs are in strong LD, using the unweighted genomic relationship matrix in G-BLUP can cause upward bias in the heritability estimation (Speed et al. 2012; Fernando et al. 2017; Legarra 2016)

Varying degree of LD between SNPs and QTLs in each may lead to biased heritability estimate (Yang et al. 2015; Gusev et al. 2013; Yang et al. 2017)

Practical session:
Building GRM in R

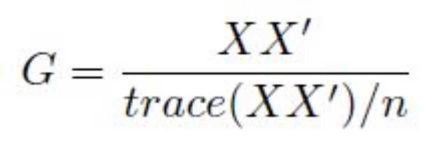
Overview

SNPready R package

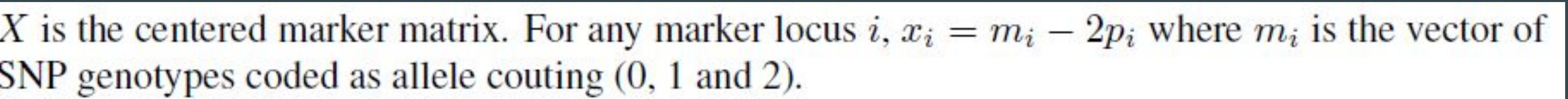
G.matrix → Four types of GRM

- VanRaden (2008) ,
- Yang (2010)
- Yang's modified
- Gaussian kernel

Van Raden



$$G = \frac{XX'}{trace(XX')/n}$$



X is the centered marker matrix. For any marker locus i, $x_i = m_i - 2p_i$ where m_i is the vector of SNP genotypes coded as allele couting (0, 1 and 2).

```
G_vanRaden <- G.matrix(X_alleles, method="VanRaden", plot = TRUE)
```

`X_alleles` don’t need to be centered

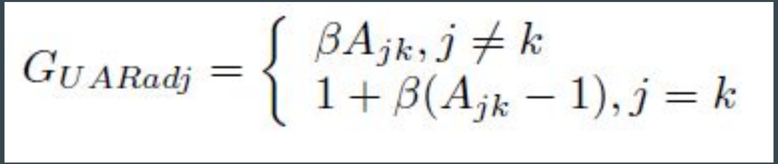
Yang

$$G_{UAR} = A_{jk} = \frac{1}{N}\sum_{i} A_{ijk} = \begin{cases} \frac{1}{N}\sum_{i} \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, j \neq k \\ 1 + \frac{1}{N}\sum_{i} \frac{x_{ij}^2(1+2p_i)x_{ij}+2p_i^2}{2p_i(1-p_i)}, j = k \end{cases}$$

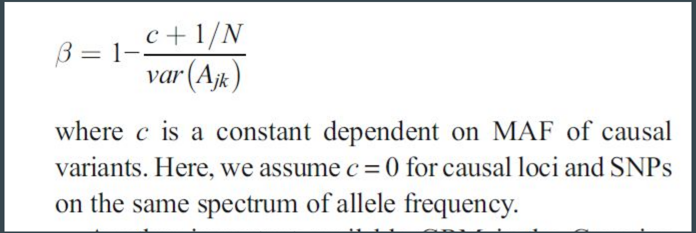
```
G_Yang <- G.matrix(X_alleles, method="UAR", plot = TRUE)
```

`X_alleles` don’t need to be centered

Yang - modified



$$G_{UARadj} = \begin{cases} \beta A_{jk}, j \neq k \\ 1 + \beta(A_{jk} - 1), j = k \end{cases}$$



$$\beta = 1 - \frac{c + 1/N}{var(A_{jk})}$$

where c is a constant dependent on MAF of causal variants. Here, we assume $c = 0$ for causal loci and SNPs on the same spectrum of allele frequency.

```
G_Yang <- G.matrix(X_alleles, method="UARadj", plot = TRUE)
```

`X_alleles` don’t need to be centered

Gaussian kernel

$$K(x_i, x_{i'}) = \frac{exp(-d^2_{ii'})}{quantile(d^2, 0.5)}$$

```
G_Gaussian <- G.matrix(X_alleles, method="GK", plot = TRUE)
```

`X_alleles` don’t need to be centered