Genomic relationships

Andrés Legarra,
UR 631 Saga, INRA Toulouse, France

Acknowledgements

- ANR projects Amasgen, Rules&Tools; Apisgene
- Toulouse bioinformatics platform (bioinfo.genotoul.fr)
- GENOMIA funding:

  www.poctefa.eu

Measurements of relationships

- Coancestry r_{xy} (Malécot coefficent, coefficient de parenté ou d'apparenté) : probability that a randomly drawn gene from x is *identical by descent* with a gene randomly drawn from y
- Inbreeding f_z (coefficient de consanguinité): the probability that the two genes in z descending from x and y are *identical by descent*. $r_{xy} = f_z$
  - and $r_{zz} = (1+f_z)/2$
- Additive relationship: covariance between additive genetic values (u) of individuals x and y
  - $Cov(u_x,u_y)=a_{xy}\sigma^2_u$
  - Twice the coancestry

Measurements of relationships

- The additive or numerator relationship matrix (a_{xy}, numerator)
  - is *not* a matrix with probabilities
  - but of 2 * coancestries (r_{xy})
  - describes covariances between individuals due to additive variation
- Inbreeding and relationships
  - are defined with respect to a base population (usually founders)
  - where an arbitrary relationship across individuals is defined (usually 0).

Measurements of relationships

- Wright (and Cockerham later) was very open in his interpretation of F
  - F (=inbreeding, relationship) can be, depending on the context,
    - a correlation (and as such, it can be negative),
    - a variance component (positive),
    - a measure of the structure of populations (F_{st}) or
    - a relationship between individuals
- But in all cases, it measures the excess from Hardy-Weinberg equilibrium
  - mind, if mate animals *against* homozigosity inbreeding can be negative

	A	a
A	p^2+Fpq	$pq(1-F)$
a	$pq(1-F)$	q^2+Fpq

Measurements of relationships

- How do we conciliate negative « F »'s (inbreeding, whatever) with our **A** which has (positive) probabilities only?
- Remember: IBD is a proxy to the true (unknown) IBS at the gene
  - Coancestry is usually positive as a byproduct of considering founders as unrelated
    - certainly this is false: founders are always related
  - But there is no need to *impose* coancestry to be positive

Molecular relationships

- In conservation genetics, molecular markers have often been used to estimate relationships
  - Either estimates of r_{xy}, or estimates of « the most likely relation » (son-daughter, cousins, whatever)
  - Not very accurate
  - e.g. Ritland, 1996

- Some formulae pop out in later works

genomic relationships

- Two ways of deriving the genomic relationship matrix
  - The first is an extension of BLUP_SNP
    - SNP have effects
    - Individuals are similar because they share SNP effects
  - SNPs give clues on « family » relationships
    - i.e., two individuals sharing lots of genotypes at SNPs are likely because they belong to the same family
    - We estimate a relationship that is more accurate than the one estimated by genealogy

The genomic relationship matrix

VanRaden, 2007, 2008

- Assume **g** = **Za**

Using centered coding !

(genetic value = sum of SNP effects).

- If we assume Var(**a**)=**I**σ^2_a , it follows from theory that
  - Var(**g**)=**ZZ'**σ^2_a
    - This is the covariance matrix of **g**, individual genetic values (or BVs)
  - This is not very informative because σ^2_a has no interpretation for us (it is just the variance of SNP effects)
  - And also, we would like the covariance of individuals to look like a relationship matrix (~1 in the diagonal and not something that depends on the number of SNPs)
  - Also, there is an issue with **Z** (which coding should I use?)

Centering Z

- value of « 1 » allele = $-p_i\ a_i$
- value of « 2 » allele = $(1-p_i)\ a_i$, where a_i is the effect of the SNP at that locus, and p_i is the frequence of the allele 2
- Thus results in centered **Z** matrix (E(**Za**)=0 for any **a**)
- e.g. the sum of each column of **Z** is 0
  - « 11 » = $-2p_i$
  - « 12 » = $1-2p_i$
  - « 22 » = $2-2p_i$
- i.e. we force the average BV to be 0.

Variance of the genetic values

Gianola et al., 2009 (Genetics)

- Suppose a population in Hardy-Weinberg, Linkage Equilibrium
  - The variance of the genetic values of the pool of individuals that form this population is σ^2_u
- If SNP effects are drawn from a distribution with Var(**a**)=**I**σ^2_a
- Then $$\sigma_u^2 = 2 \sum_{all\ SNPs} p_i \left(1 - p_i\right) \sigma_a^2$$
- Remember, this is an approximation because HW, LE do not always hold
  - For instance if the population is a cross of two lines

The genomic relationship matrix

- We want to transform
  - Var(**g**)=**ZZ'**σ^2_a into
  - Var(**g**)=**G**σ^2_u
  - where σ^2_u is the genetic variance of the population

- One way is to use $\mathbf{G} = \mathbf{ZZ'}/2\sum_{all\ SNPs} p_i(1-p_i)$

  in this case we <u>assume</u> $\sigma^2_u = 2\sum_{all\ SNPs} p_i(1-p_i)\sigma^2_a$

- And we have « declared » a base population with average 0, allelic frequencies ***p*** and « <u>genomic</u> » genetic variance σ^2_u

- Usually this **G** is compatible with pedigree (as in the Single Step) but it won't be so for extreme cases (too much drift, strong selection, crosses)

The genomic relationship matrix

- Different weights by SNP can be given by using Var(**a**)=**D** where **D** is a matrix with different variances (weights) for each SNP.
- Thus
  - Var(**g**)=**ZDZ'** = $\mathbf{G}\sigma^2_u$
- These weights can be obtained by another method (Bayesian Lasso, BayesB: Zhang et al., 2010, Legarra et al., 2011)
- In this case one should use $\mathbf{G} = \mathbf{ZDZ}'/\sigma^2_u$

Some properties of **G**

- In H-W, Linkage equilibrium
  - Average of Diag(**G**) = 1
  - Average off-diagonal(**G**) =0
  - Average genetic value of genotyped individuals =0
  - This corresponds to the definition of base population

$$\mathbf{G} = \mathbf{ZZ}' / 2 \sum_{all\ SNPs} p_i \left(1 - p_i\right)$$

- With average inbreeding F
  - Average of Diag(**G**) = 1+F

	AA	Aa	aa
freq	$q^2 + pqF$	$2pq(1-F)$	$p^2 + pqF$

Average genetic value=0

Let matrix $\mathbf{Z}$ be composed of $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots \mathbf{z}_n)$ columns. Each $\mathbf{z}_i$ column has $2(1-p_i)$ occurrences of allele 1 (with effect $-p_i a_i$) and $2p_i$ of allele 2 (with effect $(1-p_i)a_i$); therefore, the sum of $\mathbf{z}_i a_i$ cancels out for any column.

Average off-diagonal=0

Also, in case of LE, terms out of the diagonal of $\mathbf{Z}'\mathbf{Z}$ are null, for the following. These are the crossproducts of covariables associated with loci i and j. In LE, these crossproducts occur with frequency $(1-p_i)(1-p_j)$ for the co-occurrence of alleles "1" in i and "1" in j, $(p_i)(1-p_j)$ for "2" and "1", and so on. Then, by summing in order genotypes at respective loci i and j "1" and "1", "1" and "2", "2" and "1", and "2" and "2", weighted by the respective frequencies:

$$\begin{aligned}\mathbf{z}_i'\mathbf{z}_j =&(1-p_i)(1-p_j)(-p_i)(-p_j)+\\ &(p_i)(1-p_j)(1-p_i)(-p_j)+\\ &(1-p_i)(p_j)(-p_i)(1-p_j)+\\ &(p_i)(p_j)(1-p_i)(1-p_j)=0\end{aligned}$$

Not positive definite

- Strandén & Christensen (2011) showed that **G** constructed with « centered » coding is not positive definite (has no inverse)
- We could use BLUP equations with non-inverted **G** (Henderson, 1984)
- Instead, we use $\mathbf{G} = 0.99\frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} + 0.01\mathbf{A}$

  or something similar

Take-home message 1

- By defining a genomic relationship matrix, we define a genetic base
  - All inference will refer to this genetic base.
  - For instance, reliabilities computed from inverse of the MME will be different depending on the assumed *p*'s.
  - It seems that « observed » *p*'s (or even better, *p*'s at the base population) are a good reference

GBLUP

- GBLUP is a « BLUP » constructed with **G** so defined
  - Sustitute **A** for **G**

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{G}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- As in regular BLUP, we can include animals with genotype but without phenotype

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$$

GBLUP

- GBLUP gives *identical* results to BLUP_SNP *if* we fit equivalent variances in both

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{I}\sigma_a^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\sigma_u^2 = 2\sum_{all\ SNPs} p_i\left(1-p_i\right)\sigma_a^2$$

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i\left(1-p_i\right)}$$

$$\hat{\mathbf{g}}\ \text{from GBLUP} = \mathbf{Z}\hat{\mathbf{a}}\ \text{from BLUP\_SNP}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{I} \\ \mathbf{I}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{I}'\mathbf{R}^{-1}\mathbf{I}+\mathbf{G}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{I}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

GBLUP

- We can jump from GBLUP to BLUP_SNP

$$\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{a}}$$

GEBV's from SNP effects

SNP effects from GEBV's (Henderson, 1973; Strandén and Garrick, 2009)

$$\hat{\mathbf{a}} = \mathbf{I}\sigma_a^2\,\mathbf{Z}'\mathbf{G}^{-1}\sigma_u^{-2}\hat{\mathbf{g}}$$

GBLUP

Some advantages of GBLUP:

- It fits nicely into existing BLUP software
- …and into existing theory (REML, multiple traits…Single Step)
- Provides measures of accuracy from the inverse of the LHS
- Accomodates all animals

Inconvients:

- Can't easily accomodate major genes (unless using weights in the construction of **G**)
- Computation of **G** and inversion might be challenging

GREML, G-Gibbs…

Use of **G** to estimate variance components…
It can be done with `remlf90, gibbs*f90, AsReml, TM...`

The result will refer to an ideal population with whatever allelic frequencies we introduced in the computation of **G**.

Remember: the simplest is to use « observed » (centered) allelic frequencies

Otherwise (for instance fixing all p=0.5) your estimated variances will be too high.

- What has this **G** to do with pedigree relationships?

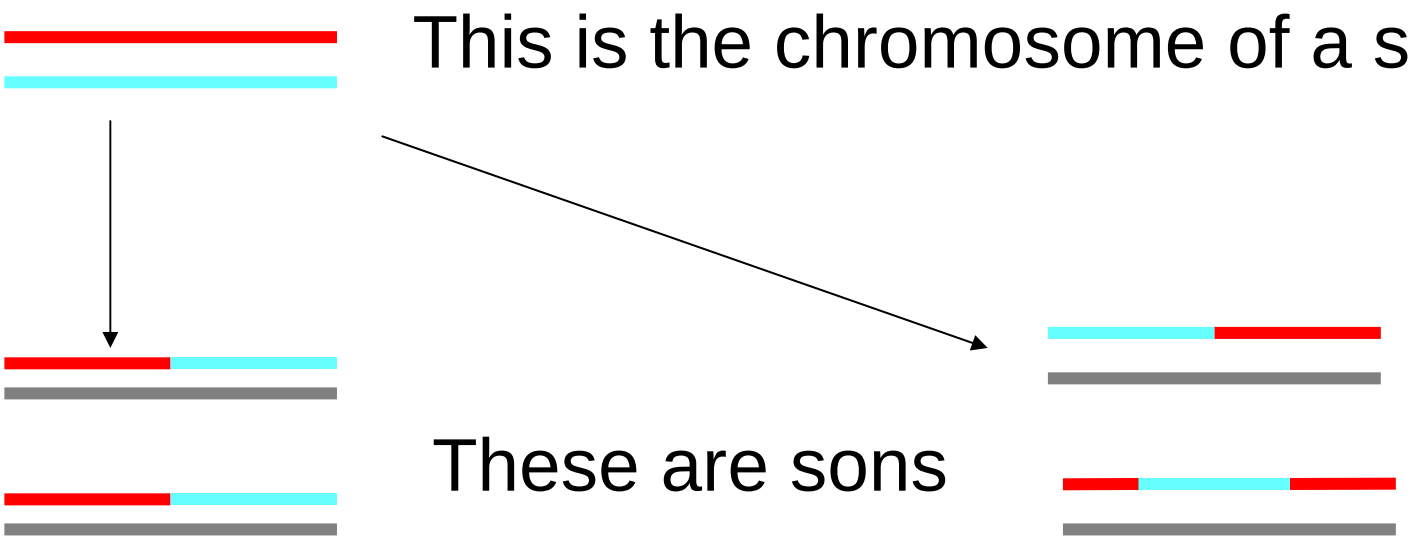
The genomic relationship matrix

- The other way around
  - SNPs are genotyped, and thus follow Mendel rules in transmission
  - So, we can use this Mendel rules…
  - to deduce « true » relationships
- But what is a « true » relationship?

The genomic relationship matrix

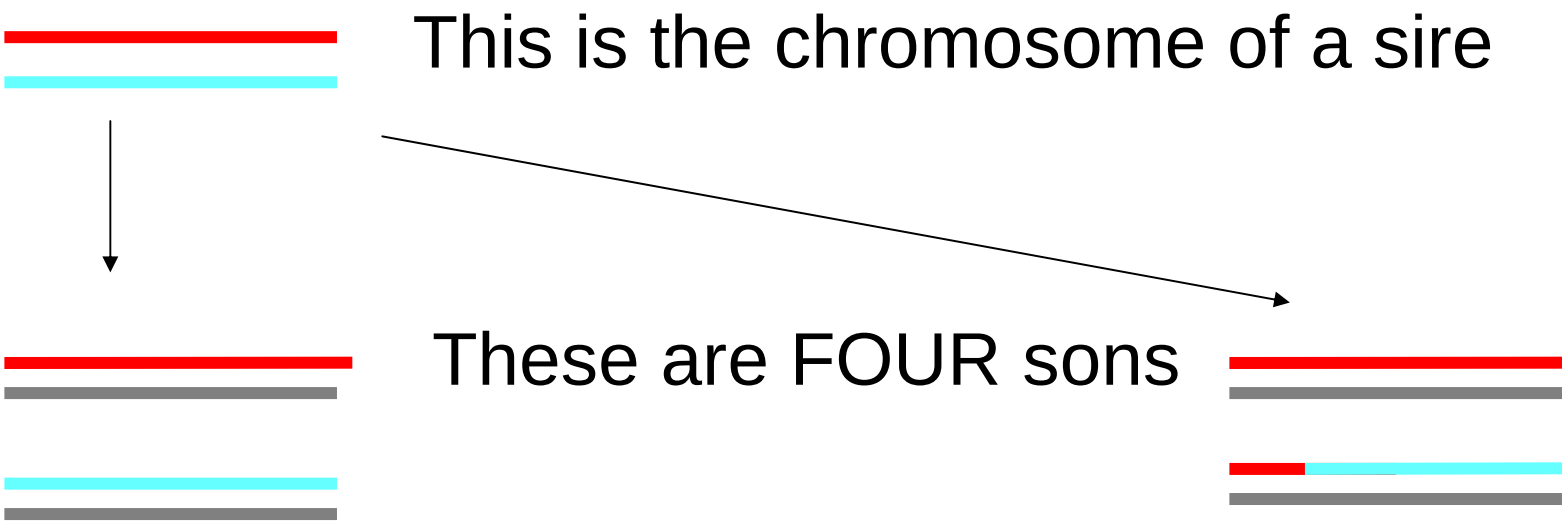
- The relationship matrix **A** based on pedigree is an average relationship which assumes infinite loci.
- « Real » relationships are a bit different due to finite genome size (Hill and Weir, 2010)
- Therefore **A** is the expectation of realized relationhips
- SNPs more informative than **A**.
  - Two half-sibs might have a correlation of 0.3 or 0.2
- You need many markers to get these « fine relationships »

Example



- In the infinitesimal model, each son receives exactly half the sire.
- But we don’t even know if the halves are identical or not. This « noise » is the mendelian sampling

Example



- In reality, two sons are identical and other two are very different from the first two but alike among them.
  - Somehow with SNP we can see this and catch mendelian sampling

Realized vs. expected

- With **G**, we estimate this realized relationship matrix
- **A** is a matrix of expected relationships

- E(**G**)=**A**

First derivation

- VanRaden (2008) explains (without much detail) that **G** (if derived properly) and the pedigree relationship (**A**) are somehow « compatible »
- The idea behind is that genetic base and variances are the same
- He provides three derivations; I'll show two
  - I will provide first the rationale why **G** is related to **A** (Toro et al., 2011 GSE)

Molecular measures of similarity

1) Molecular coancestry

	Individual i	*Individual j*	$f_{M(i,j)}$
Locus 1	AA	AA	1
Locus 2	Bb	Bb	0.5
Locus 3	Cc	CC	0.5
.	.	.	.
.	.	.	.
Locus L	mm	MM	0

the probability that two alleles taken at random, one from each individual, are equal

$$f_{M(i,j)} = \frac{\sum_{L} f_{l(i,j)}}{L}$$

In more formal terms if g_{ik} is the frequency (= gene content/2) of an allele (A, B,C,..) in individual *i*

Note that g's are half Z's in G=ZZ'/2sumpq

	Individual i	*Individual j*	g_{ik}	g_{jk}
Locus 1	**AA**	**AA**	**1**	**1**
Locus 2	**Bb**	**Bb**	**0.5**	**0.5**
Locus 3	**Cc**	**CC**	**0.5**	**1**
.	**.**	**.**	**.**	**.**
.	**.**	**.**	**.**	**.**
Locus L	**mm**	**MM**	**0**	**0**

$$f_{M(i,j)} = \frac{1}{L}\sum_{k} g_{ik}g_{jk} + (1-g_{ik})(1-g_{jk})$$

2) Molecular covariance

If g_{ik} is the frequency of allele BIG (A, B,C,..) in individual i

	Individual i	*Individual j*	g_{ik}	g_{jk}
Locus 1	**AA**	**AA**	**1**	**1**
Locus 2	**Bb**	**Bb**	**0.5**	**0.5**
Locus 3	**Cc**	**CC**	**0.5**	**1.0**
.	.	.	.	.
.	.	.	.	.
Locus L	**mm**	**MM**	**0**	**0**

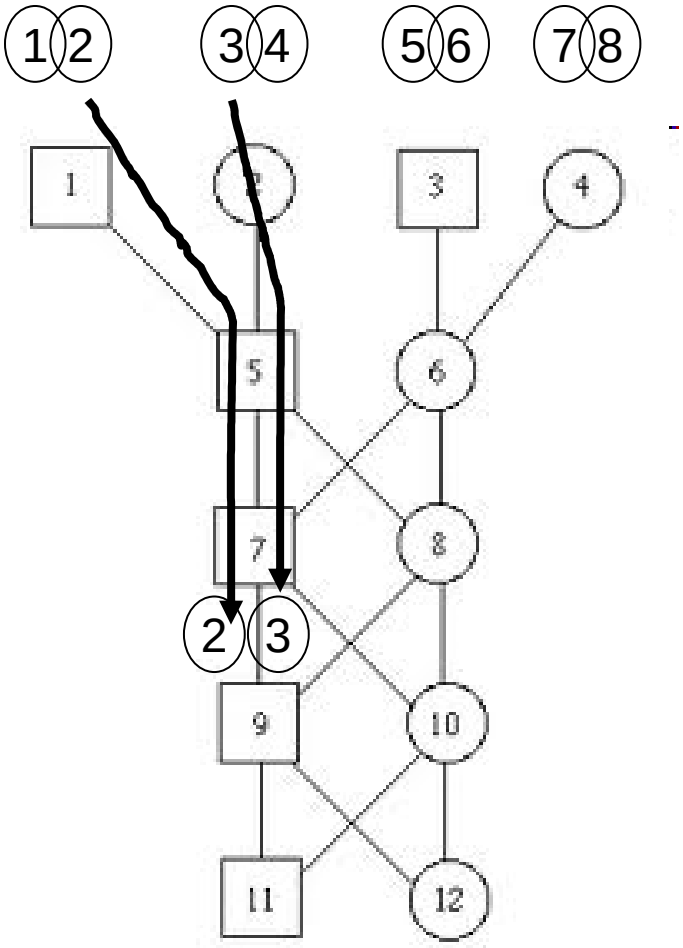
$$Cov_{M(i,j)} = Cov(g_{i\ldots}, g_{j\ldots}) = \frac{1}{L}\sum_{k}(g_{ik} - \bar{g}_i)(g_{jk} - \bar{g}_j)$$

Within-individual average allelic frequency

$$\bar{g}_i = \frac{1}{L}\sum_{k} g_{ik}$$

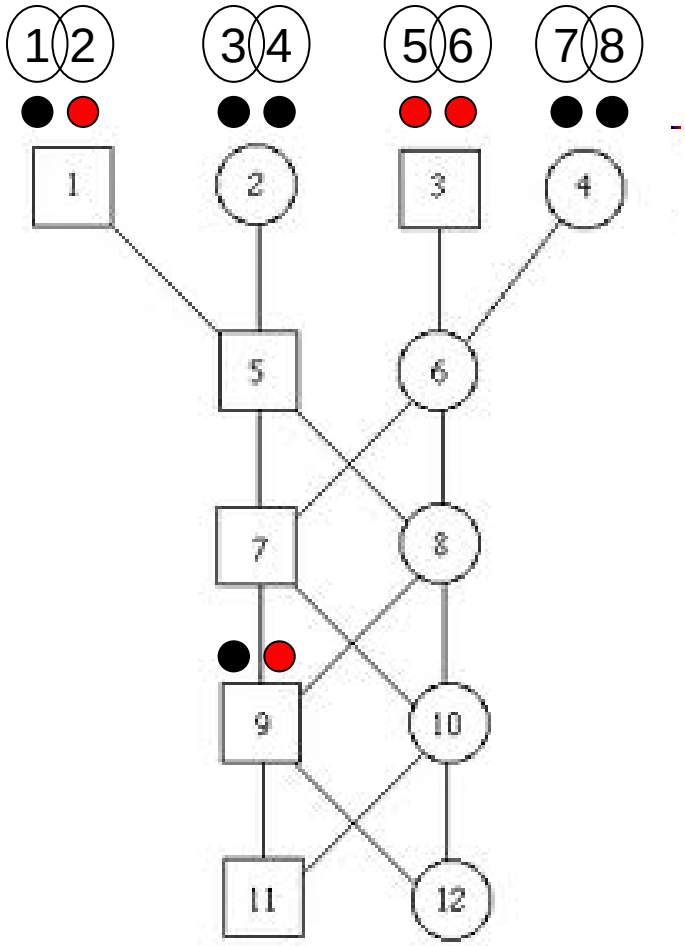
Equivalences

- Pedigree (Malécot) relationships assumes we have 2N founder alleles
- Then we genotype individual 9
- *In this case*,
  - molecular coancestry = Malécot IBD coancestry
- However SNPs have 2 alleles
  - How are then these equivalences?



With SNPs…

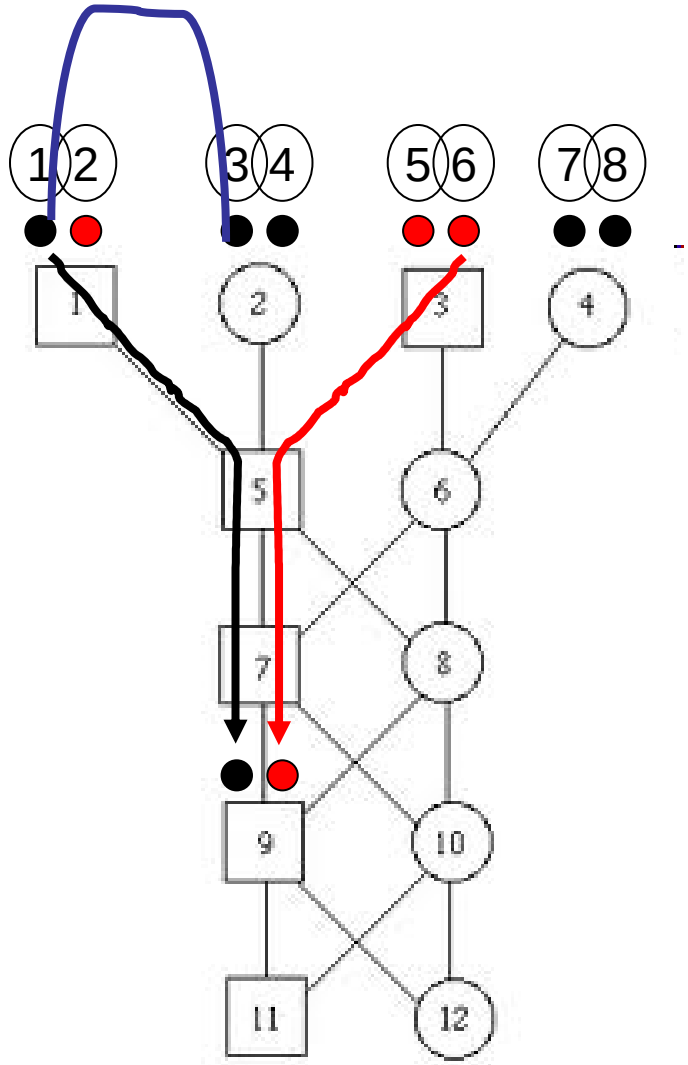
- Let us imagine that to each one of the 2M founder alleles we assign at random a tag saying if the allele is A or a with probability *p* and *q*=1-*p*
- Then we genotype 9
- Can we say which ancestral allele (1 to 8) inherited 9?



with SNPs…

- The molecular coancestry between two individuals *i* and *j* will be
  - probability that two alleles are equal (alike in state) f_{Mij},
    - either because they have become identical by descent or
    - either because they are not identical by descent but equal in the base population.

$$f_{M_{ij}} = p^2 + q^2 + 2pqf_{ij}$$



Doing the algebra (Cockerham, 1969) …

- it can be shown that, *on expectation*,

$$E\left(Cov_{Mij}\right) = f_{ij}\,pq$$

Molecular covariance

Coancestry

$$E\left(f_{Mij}\right) = p^2 + q^2 + 2\,pq f_{ij}$$

Molecular coancestry

Coancestry

- In other words
  - $Cov(g_i, g_j) = r_{ij} pq$

$r_{ij} = A_{ij} / 2$

- with allelic frequency *p* in the base population!!

Compare with VanRaden's **G**

$$E\left(Cov_{Mij}\right) = f_{ij}\,pq \Rightarrow \hat{f}_{ij} = Cov_{Mij} / pq = \frac{1}{2}\mathbf{z}_i'\mathbf{z}_j / 2\sum p_i q_i$$

for two animals

$$\hat{G}_{ij} = 2\,\hat{f}_{ij} = \mathbf{z}_i'\mathbf{z}_j / 2\sum p_i q_i$$

for ALL animals

$$\mathbf{G} = \mathbf{Z}\mathbf{Z}' / 2 \sum_{all\ SNPs} p_i\left(1 - p_i\right)$$

Therefore, **G** is an estimator, based on SNP, of « true » relationships; whereas as **A** is another estimator based on pedigree

Note that either one can be very bad (too little SNPs, incomplete pedigree)

Compare with VanRaden's **G**'s

Actually VanRaden suggests <u>three</u> G's

1st ⇨ $$\mathbf{G} = \mathbf{Z}\mathbf{Z}' / 2 \sum_{all\ SNPs} p_i \left(1 - p_i\right)$$

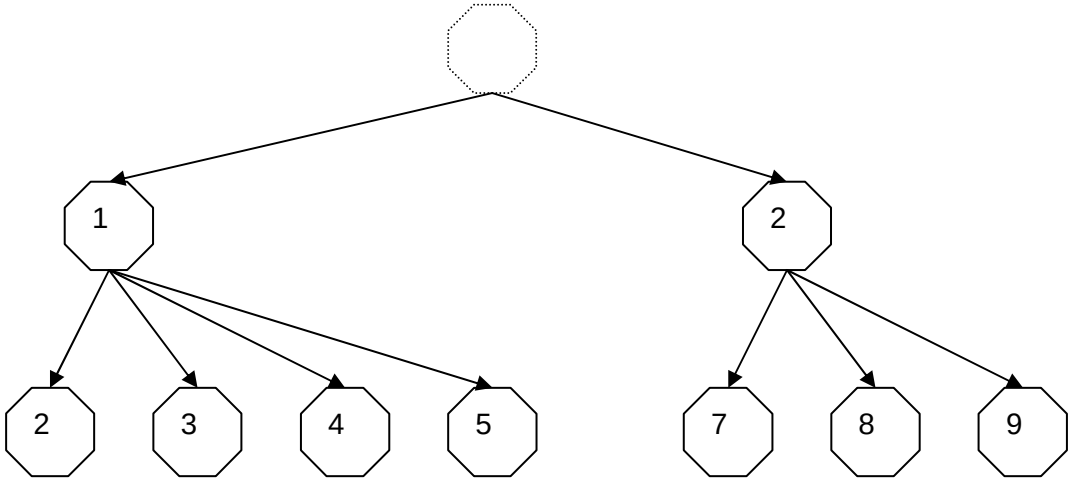
2nd ⇨ $$\mathbf{G} = \frac{1}{nsnp} \sum \frac{\mathbf{z}_i \mathbf{z}_i'}{p_i q_i}$$

Very used in human genetics; numerically unstable if p ~ 0, it does not give better results in our experience

Real results (AMASGEN)

- 9 real French bulls among 1827 genotyped, ~50000 SNPs
- Very complex pedigree
- All genotyped bulls are included in genomic estimations
- Genomic relationships as explained before
- Population means for allelic frequencies
- Programming by (most) I Aguilar and (a little) myself

Figure 3. Direct genealogical paths of the nine animals in the example.



Pedigree-based relationship

Little inbreeding

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1.00	0.51	0.57	0.51	0.26	0.15	0.15	0.14	0.14
[2,]	0.51	1.01	0.30	0.33	0.17	0.17	0.12	0.11	0.11
[3,]	0.57	0.30	1.07	0.30	0.20	0.12	0.18	0.11	0.12
[4,]	0.51	0.33	0.30	1.01	0.17	0.18	0.11	0.11	0.11
[5,]	0.26	0.17	0.20	0.17	1.00	0.56	0.51	0.52	0.53
[6,]	0.15	0.17	0.12	0.18	0.56	1.06	0.31	0.32	0.32
[7,]	0.15	0.12	0.18	0.11	0.51	0.31	1.01	0.30	0.29
[8,]	0.14	0.11	0.11	0.11	0.52	0.32	0.30	1.02	0.30
[9,]	0.14	0.11	0.12	0.11	0.53	0.32	0.29	0.30	1.03

Cousin relationships ~0.125

“first **G**” genomic relationship

Less than 1 in the diagonal

Negative coefficients

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.82	0.40	0.43	0.38	0.12	0.04	0.04	0.01	0.10
[2,]	0.40	0.91	0.18	0.24	0.02	0.05	-0.04	-0.04	0.04
[3,]	0.43	0.18	0.88	0.19	0.07	0.00	0.07	-0.02	0.05
[4,]	0.38	0.24	0.19	0.86	0.02	-0.01	-0.02	0.01	0.03
[5,]	0.12	0.02	0.07	0.02	0.73	0.34	0.30	0.31	0.35
[6,]	0.04	0.05	0.00	-0.01	0.34	0.85	0.15	0.14	0.18
[7,]	0.04	-0.04	0.07	-0.02	0.30	0.15	0.80	0.14	0.17
[8,]	0.01	-0.04	-0.02	0.01	0.31	0.14	0.14	0.80	0.17
[9,]	0.10	0.04	0.05	0.03	0.35	0.18	0.17	0.17	0.85

Relationships among cousins are ~0

$$\mathbf{G} = \mathbf{ZZ}'/2\sum_{all\ SNPs} p_i\left(1-p_i\right)$$

“Second **G**” genomic relationship

Closer to 1 in the diagonal

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.91	0.44	0.47	0.42	0.14	0.05	0.05	0.02	0.11
[2,]	0.44	1.01	0.20	0.27	0.02	0.06	-0.04	-0.04	0.04
[3,]	0.47	0.20	0.98	0.21	0.07	0.00	0.08	-0.02	0.05
[4,]	0.42	0.27	0.21	0.96	0.02	-0.01	-0.02	0.01	0.04
[5,]	0.14	0.02	0.07	0.02	0.81	0.37	0.33	0.35	0.39
[6,]	0.05	0.06	0.00	-0.01	0.37	0.94	0.16	0.15	0.20
[7,]	0.05	-0.04	0.08	-0.02	0.33	0.16	0.88	0.15	0.19
[8,]	0.02	-0.04	-0.02	0.01	0.35	0.15	0.15	0.88	0.18
[9,]	0.11	0.04	0.05	0.04	0.39	0.20	0.19	0.18	0.94

Very similar but more “exaggerated”

$$\mathbf{G} = \frac{1}{nsnp} \sum \frac{\mathbf{z}_i \mathbf{z}_i'}{p_i q_i}$$

Use of **G**

- Genomic selection (GBLUP)
- Estimation of genomic parameters (GREML, G-Gibbs)
  - In populations with no pedigree recording
  - With pedigree recording: how much variance due to SNPs, how to pedigree

Genetic parameters estimates using chicken data (resistance to salmonella) and ~900 SNPs

Legarra et al., 2010, Poult. Sci.

Table 1. Estimates of genetic parameters of analyses with pedigree only (top, Control) or pedigree and SNP markers (bottom, Combined) for chicken caecal load ($Young_{log(cfu)}$), adult liver ($Adult_{liver}$), spleen ($Adult_{Spleen}$) or caecal ($Adult_{Caeca}$) contamination as well as animal contamination ($Adult_{0-1}$). Residual variance and heritability explained by pedigree (h^2_u) or markers (h^2_h).

Method of genetic evaluation		$Young_{log(cfu)}$	$Adult_{liver}$	$Adult_{Spleen}$	$Adult_{Caeca}$	$Adult_{0-1}$
Control	Var(e)	1.72	0.011	0.022	0.17	0.18
	h^2_u	0.17	0.04	0.17	0.18	0.21
Combined	Var(e)	1.85	0.011	0.024	0.18	0.19
Pedigree: **A** →	h^2_u	0.048	0.002	0.019	0.02	0.02
Markers: **G** →	h^2_h	0.034	0.009	0.079	0.13	0.19

Same residual variances

Markers + pedigree do not capture all h^2

Markers explain almost everything

Genetic parameters estimates using mice data and pedigree (**A**) or **G**

G

A

Very different!

	varg	varu	varc	vare
		Body weight		
1		4.726	2.091	0.087
2	1.421		3.24	1.89
3	0.991	3.649	2.115	0.1273

Very similar

		Body length		
1		0.038	0.048	0.147
2	0.035		0.050	0.149
3				

Take home

- Genomic relationships work very well and are well defined
- With >50K SNP chips they are similar, but more exact, than pedigree relationships
- The exact formula for **G** depends on the interpretation but results do not change much
  - Unless somebody wants to combine pedigree and molecular relationships