Rationale for estimating genealogical coancestry from molecular markers

Genetics Selection Evolution 2011, 43:27

Toro, M.A., García-Cortés, L.A., Legarra, A.

ETSIA UPM, Ciudad Universitaria 28040 Madrid, Spain.
INIA, Ctra. Coruña Km 7.5 28040 Madrid.
INRA, UR 631 SAGA, F-31326 Castanet Tolosan, France.

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*

	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

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

There are many other measures of molecular similarity
Why to choose these ones?

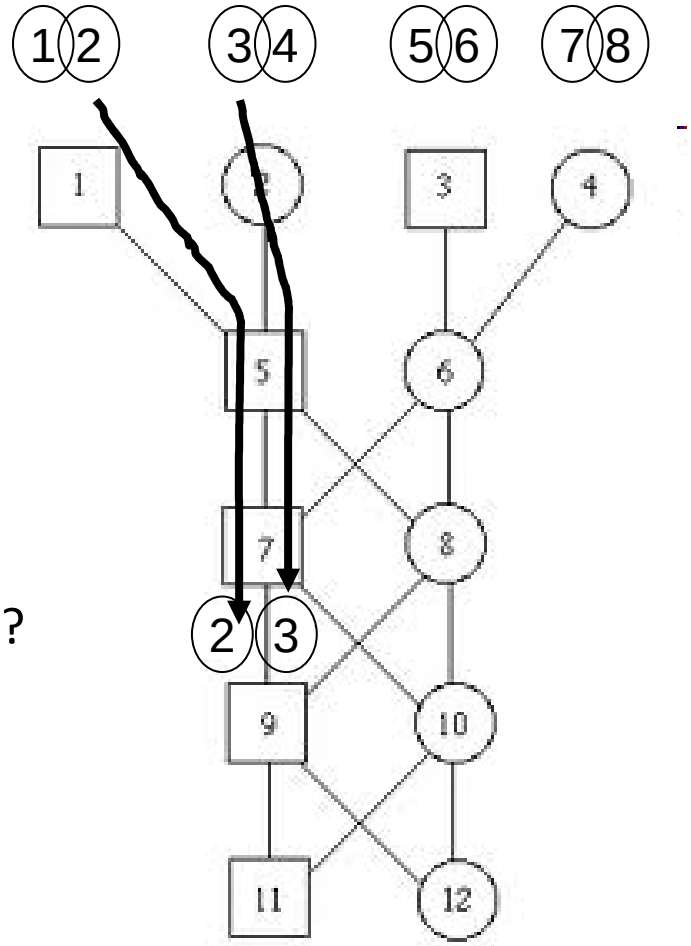
Let consider two individuals

	Individual i	Individual j
Locus 1	AA	Aa
Locus 2	BB	Bb
Locus 3	CC	Cc
.	.	.
.	mm	Mm
Locus L	nn	Nn
	oo	Oo

Molecular self-coancestry	***1.00***	***0.50***	Inbreeding
Molecular variance	***0.25***	***0.00***	Genetic drift

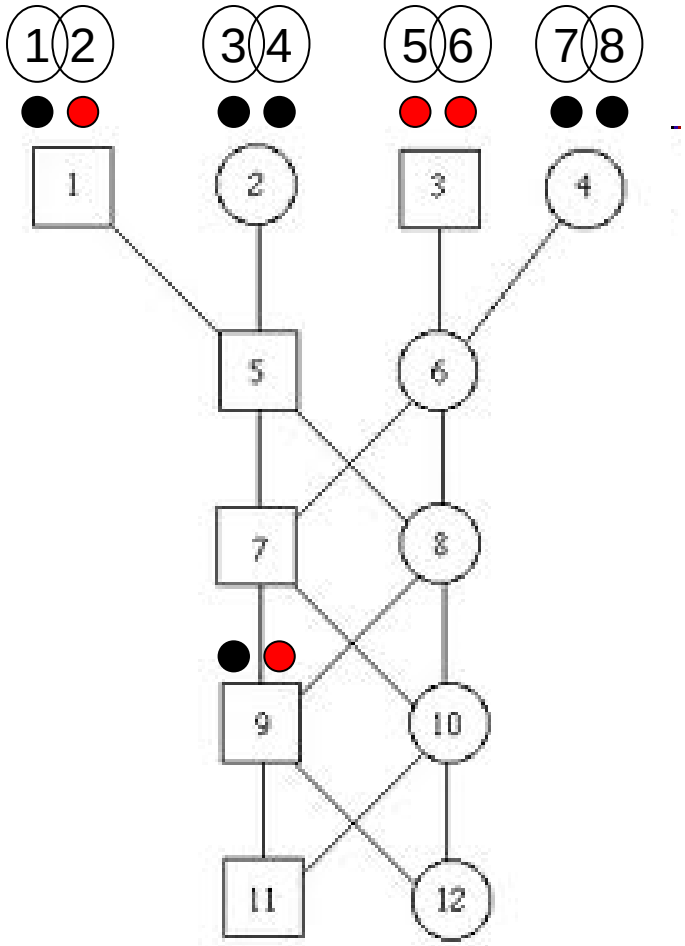
Equivalences

- Malécot 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…

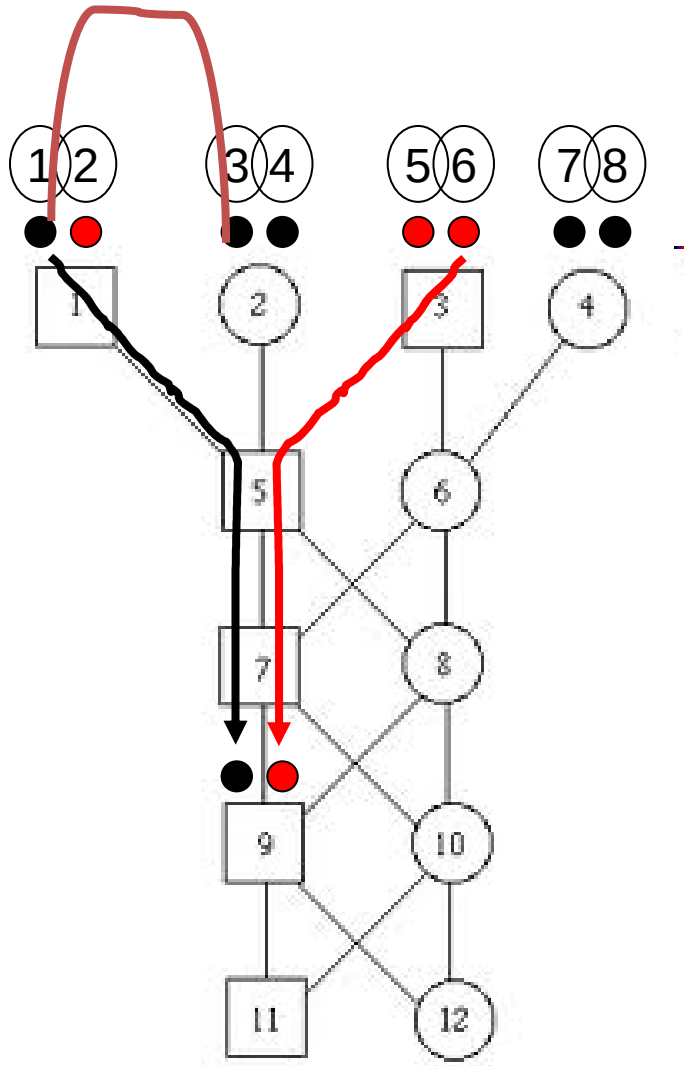
- Let us imagine that to each one of the 2M founder alleles we assign at random a tag saying if the allele is A or a with probability p and q=1-p
- Then we genotype 9
- Can we say which ancestral allele (1 to 8) inherited 9?



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!!
- But allelic frequencies are typically variable…
  - Can be thought of as coming from a random (beta) distribution

Variation of allelic frequencies

- it can be shown that, *on expectation across the distribution of allelic frequencies*,

$$E\left(Cov_{Mij}\right) = Var\left(p\right) + f_{ij}\left(\overline{p}\,\overline{q} - Var\left(p\right)\right)$$

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

- Reversing these formulae, estimators of coancestry f_{ij} can be easily derived

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

Not averaged within-individual but (possibly) within loci

1st ➡ $\hat{f}_{VR1ij} = \frac{1}{n} \frac{\sum (g_{ik} - p_k)(g_{jk} - p_k)}{\sum p_k (1 - p_k)}$

allelic frequencies are « fixed » (not random)

2nd ➡ $\hat{f}_{VR2ij} = \frac{1}{n} \sum \frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k (1 - p_k)}$

numerically unstable if p ~ 0

INRA

Testing

- Simulation **(drift):** 20 individuals x 10 generations, 10000 SNPs

- Data: 1827 Holstein bulls (~6 generations), 51325 SNP
  - MAF >0.0000….

Quality of estimators: simulated drift

$$f_{ij} = a + b\hat{f}_{ij}$$

	Ours	VanRaden's
Intercept	0.09	0.09
Slope	0.90	0.90
R^2	0.99	0.58

Alleviate drift (50 x 4 generations) R^2 =0.96

Drift creates:

estimation of allelic frequencies difficult
bias (understimates relationships)
slope (inflate them)

Quality of estimators: Real data

$$f_{ij} = a + b\hat{f}_{ij}$$

Pedigree relationships were taken as reference

	Ours	VanRaden's 1	VanRaden's 2
Intercept	0.04	0.04	0.04
Slope	0.45	0.80	0.28
R^2	0.45	0.76	0.23

Within-individual averaging loses information

Numerical instability gives lots of problems

Conclusions

- Relationships between IBD and molecular relationships are easily established
  - Building estimators is thus simple
  - Need to consider *p*'s as random
- Lack of knowledge of allelic frequencies is a problem
  - But not for practical purposes

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 coefficient, « kinship »):
  - probability (IBD)
  - But also: excess from H-W equilibrium, « correlation between uniting gametes » (Wright; *can be negative* !!)
- Remember: IBD is a proxy to the true (unknown) IBS at the gene