Noname manuscript No.
(will be inserted by the editor)

A simplified calculation of the correlations between relatives

Reginald D. Smith

Received: date / Accepted: date

Abstract The correlations between relatives is one of the fundamental ideas and earliest success of quantitative genetics. Whether using genomic data to infer relationships between individuals or estimating heritability from correlations of phenotypes amongst relatives, understanding the theoretical genetic correlations is a common task. Calculating the correlations between arbitrary relatives in an outbred population, however, can be a careful and somewhat complex task for increasingly distant relatives. This paper introduces an equation based method that consolidates the results of path analysis and uses easily obtainable data from non-inbred pedigrees to allow the rapid calculation of additive or dominance correlations between relatives even in more complicated situations such as cousins sharing more than two grandparents and inbreeding.

Keywords correlation of relatives · quantitative genetics · kinship

1 Introduction

One of the key achievements of the Modern Synthesis in evolutionary biology was the determination of the expected degree of genetic relationship between relatives given a known pedigree. First rigorously defined and expanded in Fisher's landmark 1918 paper [Fisher 1918], the correlation of relatives became one of the first applied tools of quantitative genetics valuable for selective breeding and analysis of complex genetic disorders in families. Important contributions by Sewall Wright, especially path analysis [Wright 1922, Wright 1934] and Gustave Malécot [Malécot 1948, Malécot 1970] helped create the modern methods and notations to investigate the correlations between relatives for loci with additive or dominance contributions to genetic variance.

R. Smith
Supreme Vinegar LLC, 3430 Progress Dr., Suite D, Bensalem, PA 19020
Tel.: +1-215-633-9355
E-mail: rsmith@supremevinegar.com

Further work on the correlation of relatives focused on the correlations due to epistatic effects [Kempthorne 1954, Cockerham 1954, Kempthorne 1955] including adjustments for linkage and linkage disequilibrium [Cockerham 1956, Gallais 1974]. While the correlations amongst relatives is well-known for almost all typical situations, the methods of calculating these correlations remain algorithmic and recursive. In this paper, an exact equation using the information from the structure of a known pedigree will be used to demonstrate a consistent mathematical formulations for the additive and dominance genetic correlations between relatives for all situations.

2 The relation of correlation and kinship to general pedigree variables

While several methods of calculating the correlation of relatives are known, almost all are recursive requiring the tracing of paths between relatives or calculating multiple types of identity coefficients in order to determine the relationship between two related individuals [Wright 1922, Gillois 1965, Jacquard 1972, Karigl 1981, Lange 2012]. However, once the correlations for various sets of relatives have been determined by analysis, it is possible to consolidate them into a general equation which combines previously recognized insights and understandings for different relative pairs.

Using notation similar to the original notation from [Wright 1922], for two relatives, X and Y with a set of most recent common ancestors A , let us define three variables: G_{XA} , the number of generations of descent from any member of A to X and G_{YA} , the number of generations of descent from any member of A to Y . G_{XA} and G_{YA} are n and n' in [Wright 1922]. For half-sibs or half-cousins, one additional path is required through the parent of the half-sib/cousin so also define $H(X, Y)$ as a binary variable that is one if either X or Y are half-siblings or half-cousins and zero otherwise.

Define as C , the number of elements of A which designates the number of common ancestors in the generation of A . For typical pedigrees $C = 1$ for direct descent (parents-offspring or grandparents-grandchildren) or $C = 2$ for other relatives. However, when cousins share more than two ancestors in A , such as double cousins sharing four grandparents, C can increase up to all ancestors for both relatives in A where the maximum value is reached at $C = 2^{G_{XA}}$. Finally, also including the possibility that the ancestral generation is inbred with coefficient F_A , we can state the additive correlation between relatives X and Y in logarithmic form:

$$-\log_2 r_{XY}^A = G_{XA} + G_{YA} + H(X, Y) - \log_2 C(1 + F_A) \quad (1)$$

In the typical case where $C = 2$ and $F_A = 0$ this reduces to

$$-\log_2 r_{XY}^A = G_{XA} + G_{YA} + H(X, Y) - 1 \quad (2)$$

For cases where X and Y are inbred

$$-\log_2 r_{XY}^A = G_{XA} + G_{YA} + H(X, Y) - \log_2 \frac{C(1 + F_A)}{\sqrt{(1 + F_X)(1 + F_Y)}} \quad (3)$$

In Table 1, the actual and compared values of $-\log_2 r_{XY}^A$ are shown for several situations where $C = 1$ or $C = 2$. Therefore using equation 1 one can quickly and accurately determine the correlation between relatives. The kinship coefficient is defined as $\theta_{XY} = r_{XY}^A/2$.

Relationship	G_{XA}	G_{YA}	$H(X, Y)$	C	$-\log_2 r$	r
Parent-offspring	0	1	0	1	1	1/2
Full Sibs	1	1	0	2	1	1/2
Grandfather-Grandson	0	2	0	1	2	1/4
Uncle-nephew	1	2	0	2	2	1/4
Grand Avuncular	1	3	0	2	3	1/8
First cousins	2	2	0	2	3	1/8
Great Grand Avuncular	1	4	0	2	4	1/16
First cousins once removed	2	3	0	2	4	1/16
Second cousins	3	3	0	2	5	1/32
First cousin twice removed	2	4	0	2	5	1/32
Second cousins once removed	3	4	0	2	6	1/64
Third cousins	4	4	0	2	7	1/128
Half Sibs	1	1	1	2	2	1/4
Half Avuncular	1	2	1	2	3	1/8
Double half cousins	2	2	1	2	3	1/8
Half Grand Avuncular	1	3	1	2	4	1/16
Half Cousins	2	2	1	2	4	1/16
Sibs whose parents are sibs ($F_A = 0.5$; $F_X = F_Y = 0$)	1	1	0	2	0.415	0.75

Table 1 A basic table of relatives relating the variables derived from pedigrees and the additive correlations amongst the relatives. Values from [Karigl 1981] and [Ramstetter 2017].

2.1 Complex cousins

The more complicated situation arises when cousins have more than two relatives in common in the generation of their most recent common ancestor. For example, double first cousins share all four grandparents since each of their parents are from one of two pairs of full siblings. “Sesqui” first cousins share three grandparents due to one parent coming from a pair of full siblings and another coming from a pair of half siblings. Except double cousins, the terminology for such cousins is not universally standardized. For purpose of analysis, however, normal cousins of any degree should only share two relatives in the generation of their most recent common ancestors. Those that share more have an increased level of relationship that is reflected in their

additive or dominance correlations. Using equation 1 we can derive the correlations for a variety of half or full cousins sharing more than one ancestor in the generation of A .

Having the ability to derive r_{XY}^A , we can additionally calculate the dominance covariance and correlation in tandem with certain insights about the probability of sharing pairs of alleles that are identical by descent. Malécot used the insights of Fisher and Wright to derive the following expression of the total covariance between two relatives

$$\sigma_{XY} = \frac{\phi + \phi'}{2} \sigma_A^2 + \phi\phi' \sigma_D^2 \quad (4)$$

In equation 4, σ_A^2 and σ_D^2 are the additive and dominant variance respectively, and ϕ and ϕ' are the correlations between each pair of loci in two relatives as to whether they both are represented by an allele that is identical by descent. For cousins, ϕ can be determined based on the number of grandparents shared between cousins. For first cousins for example, ϕ is always 1/4 even if two, three, or four grandparents are shared by cousins; the exception being half first cousins where ϕ is 1/8. For more distant cousins, however, ϕ increases in multiple groups of four grandparents shared. For example, in second cousins, for two to four great-grandparents shared, ϕ is 1/16 (at 1/32 for half second cousins), while for five to eight great-grandparents shared it is 1/8. It proceeds similarly for increasingly distant cousins.

For (non-removed) cousins that share common ancestors $G = G_{XA} = G_{YA}$ generations in the past, ϕ can be calculated as

$$-\log_2 \phi = 2G + H(X, Y) - \log_2 \text{ceil}\left(\frac{C}{4}\right) - 2 \quad (5)$$

The function ceil in equation 5 is the ceiling function which rounds up to the smallest integer greater than or equal to $C/4$. Once we have ϕ , we can calculate the dominance correlation between cousins, $r_{XY}^D = \phi\phi'$ using $\phi' = 2r_{XY} - \phi$. Table 2 shows how various degrees of relations can effect the additive and dominant correlations between different first through third cousins.

3 Epistatic Covariances

The decomposition of the covariances of relatives who have epistatic genetic effects amongst loci in linkage equilibrium was given by [Kempthorne 1954, Kempthorne 1955, Cockerham 1954]. In short, for additive epistatic effects of two (AA), three (AAA), or n (A^n) loci, the covariance is given by

$$\sigma_{XY;n} = r_{XY}^A \sigma_{A^n}^2 \quad (6)$$

This can be represented in its effects by

$$-\log_2 r_{XY}^A \sigma_{A^n}^2 = n[G_{XA} + G_{YA} + H(X, Y) - \log_2 C] \quad (7)$$

Cousin Degree	G_{XA}	G_{YA}	$H(X, Y)$	C	$\log_2 C$	$-\log_2 r$	r	ϕ	ϕ'	r_{XY}^D
First Cousin	2	2	0	2	1	3	$\frac{1}{8}$	$\frac{1}{4}$	0	0
	2	2	1	2	1	4	$\frac{1}{16}$	$\frac{1}{8}$	0	0
	2	2	0	3	1.585	2.415	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{32}$
	2	2	0	4	2	2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{16}$
	2	2	1	4	2	3	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{64}$
Second Cousin	3	3	0	2	1	5	$\frac{1}{32}$	$\frac{1}{16}$	0	0
	3	3	1	2	1	6	$\frac{1}{64}$	$\frac{1}{32}$	0	0
	3	3	0	3	1.585	4.415	$\frac{3}{64}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{512}$
	3	3	0	4	2	4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{256}$
	3	3	0	5	2.322	3.678	$\frac{5}{64}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{256}$
	3	3	0	6	2.585	3.415	$\frac{3}{32}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{128}$
	3	3	0	7	2.807	3.193	$\frac{7}{64}$	$\frac{1}{8}$	$\frac{3}{32}$	$\frac{3}{256}$
	3	3	0	8	3	3	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{64}$
Third Cousin	4	4	0	2	1	7	$\frac{5}{64}$	$\frac{1}{64}$	0	0
	4	4	1	2	1	8	$\frac{3}{256}$	$\frac{1}{128}$	0	0
	4	4	0	3	1.585	6.415	$\frac{3}{256}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{8192}$
	4	4	0	4	2	6	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{4096}$
	4	4	0	5	2.322	5.678	$\frac{5}{256}$	$\frac{1}{32}$	$\frac{1}{128}$	$\frac{1}{4096}$
	4	4	0	6	2.585	5.415	$\frac{3}{128}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{2048}$
	4	4	0	7	2.807	5.193	$\frac{7}{256}$	$\frac{1}{32}$	$\frac{3}{128}$	$\frac{3}{4096}$
	4	4	0	8	3	5	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{1024}$
	4	4	0	9	3.17	4.83	$\frac{9}{256}$	$\frac{1}{64}$	$\frac{3}{128}$	$\frac{9}{8192}$
	4	4	0	10	3.322	4.678	$\frac{5}{128}$	$\frac{1}{64}$	$\frac{3}{32}$	$\frac{5}{2048}$
	4	4	0	11	3.459	4.541	$\frac{11}{256}$	$\frac{1}{64}$	$\frac{5}{128}$	$\frac{11}{8192}$
	4	4	0	12	3.585	4.415	$\frac{3}{64}$	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{9}{4096}$
	4	4	0	13	3.7	4.3	$\frac{13}{256}$	$\frac{1}{16}$	$\frac{5}{128}$	$\frac{13}{2048}$
	4	4	0	14	3.807	4.193	$\frac{7}{128}$	$\frac{1}{16}$	$\frac{3}{64}$	$\frac{7}{1024}$
	4	4	0	15	3.907	4.093	$\frac{15}{256}$	$\frac{1}{16}$	$\frac{7}{128}$	$\frac{15}{2048}$
	4	4	0	16	4	4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{256}$

Table 2 The additive and dominance correlations between various first, second, and third cousins using equations 2 and 4 where C is the number of shared relatives in the generation of the most recent common ancestor. Of note are double cousins where $G_X = G_Y = 2$ and $C = 4$ which share the same four grandparents and double half cousins which share the same variables except $H(X, Y) = 1$.

where $n = 1$ is the base case with no epistasis considered. In full siblings or cousins with dominance correlation, for n loci in linkage equilibrium that have m loci with additive epistasis and $n - m$ loci with dominance epistasis the epistatic covariance for X and Y is

$$\sigma_{XY;n} = r_{XY}^A{}^m r_{XY}^D{}^{n-m} \sigma_{A^m D^{n-m}}^2 \quad (8)$$

4 Results and Discussion

This paper does not present any new results regarding the kinship coefficients or correlations between relatives themselves. This is already a path well worn and verified by both simulation and increasingly, genomic data of large pedigrees. However, it does demonstrate that the correlations between arbitrary relatives can be quickly and easily calculated in a non-recursive fashion taking into account a few aspects of the pedigree tree for the relatives in question. While with modern computers, computational overhead is not much of an issue, and tables of kinship coefficients abound, this method may provide a shorthand for both researchers and students of quantitative genetics to ease the process of pedigree analysis.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- [Cockerham 1954] Cockerham, CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39: 859-882.
- [Cockerham 1956] Cockerham, CC (1956) Effects of linkage on the covariances between relatives. *Genetics* 41: 138-141.
- [Fisher 1918] Fisher, RA (1919) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52: 399-433.
- [Gallais 1974] Gallais, A (1974) Covariances between arbitrary relatives with linkage and epistasis in the case of linkage disequilibrium. *Biometrics* 30:429-446. <https://doi.org/10.2307/2529199>
- [Gillois 1965] Gillois, M (1965) Relation d'identité en génétique I.-Postulats et axiomes mendéliens. *Ann. Inst. H. Poincaré Probab. Statist.* 2: 1-94.
- [Jacquard 1972] Jacquard, A (1972) Genetic information given by a relative. *Biometrics* 28: 1101-1114. <https://doi.org/10.2307/2528643>
- [Karigl 1981] Karigl, G (1981) A recursive algorithm for the calculation of identity coefficients. *Ann. Hum. Genet.* 45: 299-305. <https://doi.org/10.1111/j.1469-1809.1981.tb00341.x>
- [Kempthorne 1954] Kempthorne, O (1954) The correlation between relatives in a random mating population. *J. R. Soc. B* 143: 103-113. <https://doi.org/10.1098/rspb.1954.0056>
- [Kempthorne 1955] Kempthorne, O (1955) The theoretical values of correlations between relatives in random mating populations. *Genetics* 40:153-167.
- [Lange 2012] Lange, K (2012) *Mathematical and Statistical Methods for Genetic Analysis*. 3rd Edition. Springer, New York.
- [Malécot 1948] Malécot, G (1948) *Les mathématiques de l'hérédité*. Masson, Paris.
- [Malécot 1970] Malécot, G (1970) *The Mathematics of Heredity*. W.H. Freeman, San Francisco.
- [Ramstetter 2017] Ramstetter, MD (2017) High resolution relative detection via inference of identical by descent sharing of sample ancestors. Dissertation, Cornell University.
- [Wright 1922] Wright, S (1922) Coefficients of inbreeding and relationship. *Am. Nat.* 56: 330-338.
- [Wright 1934] Wright, S (1934) The method of path coefficients. *Ann. Math. Stat.* 5: 161-215.