**Title**:Structure of SNP-based Heritability and Its Application for the Inference of Genetic Architecture

**Author**:Ting Xu1,Guo-Bo Chen2 (ORCID: 0000-0001-5475-8237)

**Affiliations**:

1Department of Statistics, Zhejiang University, Hangzhou, Zhejiang Province, China,

2Clinical Research Institute, Zhejiang Provincial People’s Hospital, People’s Hospital of Hangzhou Medical College, Hangzhou, Zhejiang Province, China

**Correspondence**:G-B C

**Email**:[chenguobo@gmail.com](mailto:chenguobo@gmail.com)

**Post Address**:

Shangtang Rd 158

Hangzhou, Zhejiang Province 310014, China

**Running title**: Decode SNP-heritability

**Technical details for the modified Haseman-Elston Regression**

**SNP-based genetic relationships**

To a single locus, under the orthogonal coding scheme, it is common to partition the genetic variance into additive and dominance genetic variance (Vitezica *et al.*, 2017). Under the assumption of Hardy-Weinberg equilibrium (HWE), for the individual, the additive and the dominance code schemes for the locus are and for , respectively.

Under HWE, , , and , respectively. Furthermore, after standardization, the genotyped loci, we have vector , a vector of elements; for dominance effects, . Then, genetic relationship matrices (of dimension , the sample size) can be constructed in terms of the additive and the dominance effects. Between individual and their genome-wide relatedness can be written as

in which is a diagonal matrix and and is the trace of .

When there is no weight introduced, an identity matrix, . For a pair of individuals,

In contrast, when the weight is introduced (VanRaden, 2008), and , and

The weighted form is often employed in parallel to the unweighted form (de los Campos *et al.*, 2015; Goudet *et al.*, 2018).

Similarly, we can define the dominance genetic relationship between a pair of individuals. When there is no weight, an identity matrix,

When there is weight and , and

For the additive genetic relationship , and of the dominance relationship . Each variance term can be written in matrix form

in which an vector for unity. is an matrix representing the LD structure between markers. For additive relationship , and for dominance .

Furthermore, we define and , the effective number of markers in terms of additive and dominance relationships, respectively. When there is no weight , ranging between 1 (all markers are of the same allele frequencies and in complete LD between each other, which is impossible in practice for whole genome data) to (linkage equilibrium between any pair of markers), and ranging between 1 to too. As , it is expected that . and will be invoked in the derivation of the sampling variance of the regression coefficient below. Upon the populations, and can be different even for the same set of markers (see HapMap example below). Similarly, even and are weighted, we can define and .

In addition, if the samples are unrelated and no weight is applied, , in which is the expectation of an off-diagonal element of . Alternatively, ; as also represents the expected correlation between any pair of individuals, a negative value is expected because the current population is treated as the base population in which every individual is assumed to be independent. Intuitively, as the sampling variance of the current population is proportional to sample size, the relatedness of a pair of individual is more pronounced in small populations than in large populations. In practice, we have little knowledge on the relatedness between samples, so we can define the effective number of sample . If the sample is reasonaly independent, is very close to .

**Covariance between phenotype and**

The theoretical framework for deriving HE regression coefficient has been demonstrated in detail under the context of using whole genome-wide markers (Chen, 2014),

in which is the squared difference between a pair of individuals, is the intercept, the regression coefficient,only has its off-diagonal elements used, and the residual. In general, the regression coefficient for HE can be derived as

because has the mean of zero. The derivation is based on conditional probability for a pair of loci, and its layout can be found in **Table 1**.

**Table 1** Conditional probabilities for the four phases of the haplotypes consisting of a pair of biallelic loci

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Allele at the locus | |  |
|  |  |  |  | Marginal probability |
| Allele at the locus |  |  |  |  |
|  |  |  |  |

Notes:, and , the conditional probabilities for the two coupling haplotypes.

We have

The corresponding elements can be found in **Table 2**.

**Table 2** The joint distribution of the additive genetic relatedness between individual and

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Individual | | | |  | Individual | | | | |  | Relatedness for individual and | |
| Genotype |  |  |  |  |  | Genotype |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Notes:** As allele is set the reference allele, the frequency of which was , , , and were coded as 0, 1 and 2, respectively.

Similarly, we can derive for dominance effect by replacing with , , and

The corresponding elements can be found in **Table 3**. The variance and covariance terms can be integrated to give the regression coefficient of HE, which is a linear function of SNP-heritability, as will be shown below.

**Table 3** The joint distribution of the dominance genetic relatedness between individual and

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Individual | | | |  | Individual | | | | |  | Relatedness for individual and | |
| Genotype |  |  |  |  |  | Genotype |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**The analytical results for the regression coefficients**

The key step of the derivation is to find , which is upon the conditional probability between the marker and the QTL in LD (**Table 1**). The simplest case is single marker and single causal variant , which is in LD with QTL , then the analytical result is . Of note, and will cancel each other off for this single marker scenario. In practice, it often involves multiple markers and multiple causal variants but can be built from single marker and single causal variant scenario.. After rearrangement, shorthanded in matrix form

in which is the vector of elements and and , and matrix. , representing the LD between the marker and each causal variant. is a diagonal matrix with .

**Dominance variance component**

The additive effects cancels out due to orthogonal coding for the additive and dominance effects, and for single marker and single QTL,. For many markers and many causal variants, . In matrix algebra,

in which is the vector of elements for dominance effects and and . , in terms of squared Pearson’s correlation for dominance effects, between the marker and each causal variant. is a diagonal matrix, and .

**For a pair of traits**

If each individual has a pair of traits ( and ) observed, the can be modified to hav

And similarly for the dominance effects,

we have the analytical results

and

**Connection between heritability and**

The classic definition of heritability is (Lynch and Walsh, 1998)

in which and for dominance variance component is . For both and , the first summation (single summation) term is for the within-locus variance, and the second summation (double summation) term is for the between-locus covariance, or called disequilibrium covariance due to LD. When expressed in matrix form,

**(Eq 5)**

in which a symmetric square matrix characterizing the LD structure between additive effects and a symmetric square matrix characterizing the squared LD structure between dominance effects. Both and are likely sparse matrices if the trait of interest is controlled by many causal variants along a big genome.

From the above analyses, we can see the shared motifs between **Eq** 1~4 and **Eq** 5 that and are the elements shared between heritability and SNP-heritability. The difference between them is that between and for the additive SNP-heritability and that between and for the dominance SNP-heritability. However, those difference can be easily passed. If every marker is causal, between and we find that , in which the row of or ; between and we find that , in which the row of . A visualized summary for the connection between the classic heritability and SNP-heritability estimated in HE can be found in **Figure 1**.

**Literature cited**

Chen,G.-B. (2014) Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front. Genet.*, **5**, 107.

Goudet,J. *et al.* (2018) How to estimate kinship. *Mol. Ecol.*, **27**, 4121–4135.

de los Campos,G. *et al.* (2015) Genomic heritability: what is it? *PLoS Genet.*, **11**, e1005048.

Lynch,M. and Walsh,B. (1998) Genetics and Analysis of Quantitative Traits Sinauer Associates, Inc., Sunderland, MA, USA.

VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Vitezica,Z.G. *et al.* (2017) Orthogonal Estimates of Variances for Additive, Dominance and Epistatic Effects in Populations. *Genetics*, **206**, 1297–1307.