**Title:** Stochastic modeling and interpretation for SNP-based heritability using Haseman-Elston regression

Abstract 3

Introduction 4

Methods 5

Heritability in definition 5

Haseman-Elston regression 5

Estimator I : SNP-heritability for unrelated samples – population-based design 6

Additive SNP-heritability for unrelated samples 6

Dominance SNP-heritability 7

Estimator II : SNP-heritability for linkage analysis – sib pair design 7

Additive SNP-heritability for sibling in linkage analysis 7

Dominance SNP-heritability for sib pair design 9

Multiple regression analysis 9

Genetic architecture 9

Infinitesimal model 9

Local structure 10

Results 11

Simulation 11

Discussion 13

# Abstract

It exists a gap between the heritability estimated from pedigree studies () and SNP-heritability estimated from population-based studies (), and leads to missing heritability that . Various data and approaches have been employed to narrow down the gap and consequently to address the missing heritability problem, but no conclusive results have been reached. In this study, we address this problem under the framework of Haseman-Elston regression, which promises analytical resolution. We derived the analytical results for both and , and their statistical structures have been elucidated. As revealed by the uncovered structures, and tag causal variants via linkage and linkage disequilibrium, two different basic but nearly independent genetic mechanisms. Although these two estimates are little correlated due to different genetic mechanisms underlying, can be served as the upper bound for .

In order control possible perturbation of the estimation due to genetic architecture and population structure, nuclear family-based design are employed that the parental population is treated as unrelated samples for the estimation of and the sibpairs are used as linkage analysis for the estimation of . In consistent to the previous observational studies, including more variants will help push towards that is relatively invariable to tag SNPs.

# Introduction

Missing heritability question has been one of the recent topics in statistical genetics(Manolio *et al.* 2009). One of the tasks in searching missing heritability is to narrow the gap between pedigree-based estimation of heritability, such as via sib-pair design(Visscher *et al.* 2006), and the counterpart that via population-based design using unrelated samples(Yang *et al.* 2010). One of the myth in missing heritability is the gap between the pedigree-based estimation of heritability and the population-based estimation of the heritability . Various data had been tested to drive towards , imputation(Yang *et al.* 2015) and WGS.(Wainschtein *et al.* 2019)

Given even large scale data, LMM becomes infeasible in terms of computation and its statistical properties also obscure the properties of the estimated SNP-heritability. In contrast, Haseman-Elston regression, a least-squares method, have been employed to estimate the SNP-heritability(Haseman and Elston 1972; Chen 2014). HE promises analytical results for the estimated SNP-heritability. As presented in this work, HE can give (as shown below) analytical solution for and , and consequently paves an alternative path for searching missing heritability.

Historically, has been developed earlier than . When the genetic relatedness is defined on the origin of the alleles, IBD has been the major metric in estimating , especially in such as sib pair design. is based on the similarity of the observed markers, rather than their origins. Although these two metrics are correlated(Manichaikul *et al.* 2010), but their application in the estimation of SNP-based heritability leads to and .

Previous studies have been focused on the statistical estimation of SNP-based heritability, while leaves the genetic essence of SNP-heritability rarely touched, or only partially know its approximation. In this study, we use stochastic modeling method that used modified Haseman-Elston regression to provide the analysis framework. As indicated by theory, simulation, and real data, it can be conclude that , robust to the allelic spectrum of the causal variants, provides a natural upper bound of heritability, though the sampling variance is larger, and is subject to allelic spectrum of the causal variants.

# Methods

## Heritability in definition

For a standardized quantitative trait , its genetic effect can be partitioned into additive and dominance effects, ignoring epistatsis.

After orthogonal parameterization for additive and dominance codes and (Vitezica *et al.* 2017), we have genetic variance due to additive and dominance components, or their respected heritability, respectively. For the causal locus, the narrow-sense heritability of a quantitative trait is defined as below

in which the allele frequency of the reference locus, the alternative allele, the additive effect of the locus, and the correlation between the and loci.

Similarly, the dominance heritability in terms of dominance effects can be defined correspondingly,

in which the dominance effect of the locus.

Under the infinitesimal model that casual variants are randomly distributed, the high-order term trades off each other, and .

Although there are various interpretation for missing heritability, but one of the research direction is to fill the gap between , the heritability that is estimated from pedigree via IBD (Visscher *et al.* 2006) and the counterpart heritability that is estimated from population-based sample using association via IBS(Yang *et al.* 2010). is typically estimated via family-based design, such as using sib pairs in linkage analysis, and can be estimated via unrelated samples.

The current observation supports for various traits, and further is considered the upper-bound and proxy for . Data-driven approach has been widely used to address the missing heritability problem, and we here use HE to give a much crystalized framework in addressing this question. Upon the metric that measures relatedness, we have two possible forms of fundamental relatedness, identical by state (IBS) and identity in descent (IBD), and consequently make possible three estimators as described below.

## Haseman-Elston regression

The seminal HE regresses the squared phenotypic difference to IBD scores for a pair of siblings; via least squares approach, the estimated HE regression coefficient has an elegant analytical result (Haseman and Elston 1972). In this study, upon the data used, we replace the pair of the conventional relatives to unrelated individuals (in term of its conventional meaning), and the modified linear regression is as below

(**Eq 1**)

is the squared-difference for a pair of unrelated samples and ; given samples, is a vector of elements. can be constructed as additive genetic relatedness or dominance genetic relatedness, respectively. We consider three kinds of forms for .

1. For association analysis, is measured on IBS for additive and dominance genetic components (Yang *et al.* 2010; Zhu *et al.* 2015).
2. For association analysis, the alternative form of is to introduce weights (VanRaden 2008), which can be introduced in various ways. In this study, we focus on the weights related to the allele frequency, which leads to “genomic heritability” (de los Campos *et al.* 2015).
3. For linkage analysis, can be constructed using IBD under sib pair design.

The form 1 and form 2 can be considered as the same category, and we will treat them under the same stochastic modeling framework. This method has been introduced in Chen (Chen 2014), but generalized to possibly include weights. Form 3 can be extended from the seminal HE regression, which only considered one single marker and one causal variant scenario or multiple causal variant scenario in linkage equilibrium. In order to find the structure, we need to extend the IBD-HE in two ways, i) considering multiple markers and ii) considering causal variants in LD.

## Estimator I : SNP-heritability for unrelated samples – population-based design

### Additive SNP-heritability for unrelated samples (without weights)

For a single locus, the orthogonal coding scheme can separate the genetic variance into additive and dominance genetic variances (Vitezica *et al.* 2017). For the individual, the additive and the dominance code schemes for the locus are and for , respectively. Under HWE, the genotypic frequencies of , , and are , , and , respectively. After standardization, we have , a vector of elements (loci).

A genetic relationship matrix, of dimension , can be constructed in terms of the additive effects; between individuals and , their SNP-based additive genetic relatedness can be written as

(**Eq 2**)

The off-diagonal elements of the matrix can be fitted into the regression below

(**Eq** **3**)

Using the stochastic modeling method introduced by Chen (Chen 2014), the analytical result for the estimated SNP-heritability can be derived. The key step is to find . and are the conditional phenotypic values given six possible combinations for and , respectively, as found in Table 1. For more detailed technical steps, please refer to Chen (Chen 2014). The derived additive SNP-heritability is as below

(**Eq** **4**)

When the numerator describes how each marker tags, via LD, all causal variants. Of note, as markers are aggregated linearly in , and the squared term in the bracket parenthesis, as how the casual variants are tagged via the marker is below

Asis an aggregated measure for markers, the numerator has such matrices.

The denominator can also be expressed as the summation of the LD, in terms of squared correlation, matrix below

The standard error of the estimated heritability is , in which .

### Additive SNP-heritability for unrelated samples (with weights)

As can be a very small quantity for rare or low-frequent variant, weights can be introduced for genetic relatedness. We can introduce (Table 1), . The corresponding weighted related genetic relatedness is

(**Eq 5**)

This definition of the relatedness is concordant to the one used in estimating “genomic heritability” (de los Campos *et al.* 2015). When fitting into HE, using stochastic modeling, we could derive the expression for the weighted SNP-heritability —“genomic heritability”,

(**Eq 6**)

Clearly, when there are no weights introduced, is identical to . The weights will differentiate the estimate of the SNP-based heritability for quantitative traits, but the squared term in parenthesis is identical as that for no weighted SNP-heritability.

So, use stochastic modeling technique, the regression coefficient of the estimated SNP-heritability with or without weights can be illustrated as above.

### Dominance SNP-heritability

After orthogonal coding, the dominance relatedness can be defined as

(**Eq 7**)

in which for dominance genetic relatedness, we have . Similarly, can be derived accordingly,

(**Eq 8**)

The detailed structure of the squared term in parenthesis is

and the denominator is the summation of th matrix below

and its standard error is , and .

Similarly, the weights for dominance effect can be introduced, .

(**Eq 9**)

and its standard error is , and .

## Estimator II : SNP-heritability for linkage analysis – sib pair design

### Additive SNP-heritability for sibling in linkage analysis

Alternatively, the relatedness of IBD can be measured for sib pairs, such as IBD. For each marker, we can have its realized IBD for a pair of sibs, and when average IBS over the genome we have

(**Eq 10**)

As expected, the averaged LD is 0.5 for sibpairs. Fitting IBD into Eq 1, we can have conventional linkage analysis.

It is well-known from the seminal HE regression that when the IBD is measured on the single marker, marker, the interpretation for the regression coefficient is, . If the is linkage with more than one causal variants that are in linkage equilibrium to each other, , please refer to Eq 22 in the original HE paper (Haseman and Elston 1972).

However, in practice, causal variants are often clustered and in linkage disequilibrium. Between a pair of sibs, the quantification of the covariance between a pair of loci is related to IBD of the causal locus and the causal locus. As the gametes of the sibs are independent, their IBD probability is . In addition, the decay of LD due to recombination will also reduce the covariance between the and causal loci by .

When taking the linkage disequilibrium into account, the generalized expectation of the HE regression coefficient is for the marker is

The term indicates the decay of LD compared with that observed in the parental population. For the aggregated measure of IBD , the regression coefficient is

**(Eq 11)**

In matrix form, the numerator of the marker can be written as

And the denominator is

Compared with that of unrelated samples, the difference is about how different measures of relatedness quantifies the genetic variation.

The sampling variance of is , often far larger than . .

### Dominance SNP-heritability for sib pair design

The dominance SNP-heritability for sib pair design can be derived accordingly, and it can be expressed as

## Multiple regression analysis

The above analysis indicates the theoretical ground for association and linkage using whole genomic markers as single variance component. In addition, markers have been partitioned into multiple variance components, and the popular method is to partition the markers according to their minor allele frequency, the so-called MAF-bin method. The detailed statistical properties of the multiple regression is obscure, but it is easy to anticipate. For very extreme case, if we treat each marker as a single “bin”, the estimated heritability approaches , the covariance components due to LD will be traded off.

In MAF-bin analysis, the sum of the estimated partial regression coefficients is considered as heritability. As will be shown in simulation, MAF-bin can have larger or smaller estimate than the real heritability.

## Genetic architecture

### Infinitesimal model

Under the infinitesimal model, all linkage-disequilibrium terms such as the additive- and dominance covariance terms will trade off. The covariance terms in the above equation can be dropped off under the infinitesimal model,

**(Eq 2)**

in which indicates the averaged LD between a marker and a causal variants and indicates the averaged LD between a pair of tagged markers, and is the true heritability.

**Remark I:** is upon the linkage disequilibrium. For a pair of marker the maximal LD is constrained , in which and are the minor allele frequencies of the pair of loci. Uponthe allelic spectrum of the causal variants, if the tag SNPs are different in allele frequency with the causal spectrum,. So, for a trait that has its effect size , if , negative selection, would approach if more rare variants are included in building .

**Remark II**: Often a statistical method that can give an increased estimate of , is considered a better. However, it is very easy to find, albeit pathological, an example that overestimates due to local structure.

**(Eq 3)**

in which indicates the averaged linkage between a marker and a causal variant and the averaged linkage between a pair of markers.

**Remark III**: the analytical result for the structure of has never been known previously. is upon IBD, which is independent of allele frequency. So captures regardless the allelic spectrum of the causal variants.

**Remark IV**: as recombination is rare on a chromosome, fewer markers are needed to capture heritability.

For both and , at the whole genome scale, the local structure that harbors variation of covariance between causal variants are traded off.

### Local structure

Look a single marker

**Comparison between and**

It is much easier to have the insight for this two estimators under the simplest scenario for one marker and one causal variant. Eq 3 and Eq 4 can be reduced to

and

upon the allele frequency disparity between the marker and the causal variants; is allele frequency dependent. In contrast, is upon , which immunes of allele frequencies between the marker and the causal variants.

Consider and , the maximal but can be as small as zero; whereas, if their linkage is in tight linkage, say < 0.01—irrelevant to allele frequency, . In other words, for a single locus, these two signals can be orthogonal to each other. ~~It can explain little consistency between linkage and association analysis in gene mapping.~~

# Results

## Simulation

2000 nuclear families were simulated, and each family had two parents and a pair of sibs. The parental population were treated as unrelated samples and their relatedness were measured in . The unrelated samples could be used to estimated .

For 2000 sib pairs, we could measure for each sib pair. This set of data could be used to estimate .

Under this scenario, the genetic architecture between the and are exactly the same.

Scenario 1

900 markers, which had MAF range from 0.1~0.9, and 100 causal variants, which had MAF range from 0.1~0.2, recombination between any two consecutive marker was 0.01 and Lewontin’s was 0.9.

|  |  |  |
| --- | --- | --- |
| Estimator |  |  |
| Estimate I |  |  |
| (exclude causal variants) |  |
| Estimate II |  |  |
|  | (exclude causal variants) |  |
| Estimate III |  |  |
| (exclude causal variants) |  |

Of note, and were often referred to each other in “missing heritability problem”. As aforementioned, and were not correlated. The correlation between (exclude causal variants) and (exclude causal variants) was -0.034 (p-value 0.8), not correlated.





Real data analysis

We can identify sibpair from UK biobank,

Procedure (/public/home/xuhm/data/Oxford/UKB; ~/manuscript/Linkage\_Association/submit/FrontGenet/sib\_ukb)

According to 22021, we could find the 147526 individuals that had relatives included (tagged as “1”). We filtered those individuals, and estimated their IBD via the procedures below.

1 Plink –bfile Oxford –genome full –out Oxford –min 0.1

2 Sibs were selected as below: PI > 0.38 & <0.7, and IBD2 > 0.1 & < 0.6

For families has more than one sibpairs, we only selected one pair and identified 18,118 sib pairs, and finally, we found 34 parents and sibpair. This result was consistent to UKB(Bycroft *et al.* 2018).

# Discussion

References

Bycroft C., Freeman C., Petkova D., Band G., Elliott L. T., Sharp K., Motyer A., Vukcevic D., Delaneau O., O’Connell J., Cortes A., Welsh S., Young A., Effingham M., McVean G., Leslie S., Allen N., Donnelly P., Marchini J., 2018 The UK Biobank resource with deep phenotyping and genomic data. Nature **562**: 203–209.

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

Haseman J. K., Elston R. C., 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2**: 3–19.

los Campos G. de, Sorensen D., Gianola D., 2015 Genomic heritability: what is it? PLoS Genet. **11**: e1005048.

Manichaikul A., Mychaleckyj J. C., Rich S. S., Daly K., Sale M., Chen W.-M., 2010 Robust relationship inference in genome-wide association studies. Bioinformatics **26**: 2867–73.

Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. a, Hunter D. J., McCarthy M. I., Ramos E. M., Cardon L. R., Chakravarti A., Cho J. H., Guttmacher A. E., Kong A., Kruglyak L., Mardis E., Rotimi C. N., Slatkin M., Valle D., Whittemore A. S., Boehnke M., Clark A. G., Eichler E. E., Gibson G., Haines J. L., Mackay T. F. C., McCarroll S. a, Visscher P. M., 2009 Finding the missing heritability of complex diseases. Nature **461**: 747–53.

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

Visscher P. M., Medland S. E., Ferreira M. A. R., Morley K. I., Zhu G., Cornes B. K., Montgomery G. W., Martin N. G., 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. **2**: e41.

Vitezica Z. G., Legarra A., Toro M. A., Varona L., 2017 Orthogonal Estimates of Variances for Additive, Dominance and Epistatic Effects in Populations. Genetics **206**: 1297–1307.

Wainschtein P., Jain D. P., Yengo L., Zheng Z., 2019 Recovery of trait heritability from whole genome sequence data. bioRxiv: 1–23.

Yang J., Benyamin B., McEvoy B. P., Gordon S., Henders A. K., Nyholt D. R., Madden P. A., Heath A. C., Martin N. G., Montgomery G. W., Goddard M. E., Visscher P. M., 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**: 565–569.

Yang J., Bakshi A., Zhu Z., Hemani G., Vinkhuyzen A. a E., Lee S. H., Robinson M. R., Perry J. R. B., Nolte I. M., Vliet-Ostaptchouk J. V van, Snieder H., Esko T., Milani L., Mägi R., Metspalu A., Hamsten A., Magnusson P. K. E., Pedersen N. L., Ingelsson E., Soranzo N., Keller M. C., Wray N. R., Goddard M. E., Visscher P. M., 2015 Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. **47**: 1114–1120.

Zhu Z., Bakshi A., Vinkhuyzen A. A. E., Hemani G., Lee S. H., Nolte I. M., Vliet-Ostaptchouk J. V. van, Snieder H., Esko T., Milani L., Mägi R., Metspalu A., Hill W. G., Weir B. S., Goddard M. E., Visscher P. M., Yang J., 2015 Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. Am. J. Hum. Genet. **96**: 377–385.

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

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

**Notes:** and have subscript “A”, for additive, dropped off here.

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

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

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

**Notes:** and have subscript “D”, for dominance, dropped off here.

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