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

Abstract 3

Introduction 4

Methods 4

Heritability 4

Haseman-Elston regression 5

Estimator I : SNP-heritability for unrelated samples – population-based design 5

Additive SNP-heritability for unrelated samples 5

Dominance SNP-heritability for unrelated samples 6

Estimator II : SNP-heritability for linkage analysis – sib pair design 6

Additive SNP-heritability for sibling in linkage analysis 6

Dominance SNP-heritability for sib pair design 7

Multiple regression analysis 8

Genetic architecture 8

Infinitesimal model 8

Local structure 9

Results 10

Simulation 10

Discussion 12

# 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 genetics1. One of the tasks in searching missing heritability is to narrow the gap between pedigree-based estimation of heritability, such as via sib-pair design2, and the counterpart that via population-based design using unrelated samples3. 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 , imputation4 and WGS.5

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-heritability6,7. 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 correlated8, but their application in the estimation of SNP-based heritability leads to and .

In this study, we use analytical 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

under the infinitesimal model, the high-order term trades off each other, . Although there are various interpretation for missing heritability, but one of the research direction is to fill the gap between and . 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 .

## Haseman-Elston regression

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.

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 6. 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. Furthermore, weights can be introduced for 9. So, in total four possible forms of are considered here: additive relatedness, dominance relatedness, additive relatedness with weights, and dominance relatedness with weights. As The purpose here is to derive pure mathematical expectation of the regression coefficient, we fit only one in Eq 1; in real data analysis, however, multiple , in terms of additive or dominance relationship matrices,can be fitted concurrently.

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

### Additive SNP-heritability for unrelated samples

For a single locus, orthogonal coding scheme can separate the genetic variance into additive and dominance genetic variance 10. 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. The allele frequency of the locus is . Under HWE, the genotypic frequencies of , , and are , , and , respectively. After standardization, we have , a vector of elements (loci); for dominance genetic relatedness, we have .

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**)

Using the stochastic modeling method introduced by Chen 7, as promised by Haseman-Elston regression, the analytical result for the estimated SNP-heritability is

When numerator describes how each marker tags, via LD, all causal variants. For example, the marker tags the causal variants as below

Asis an aggregated measure for markers, the numerator has such matrices. The standard error of the estimated heritability is , and .

### Dominance SNP-heritability for unrelated samples

Similarly, can be derived accordingly,

and its standard error is , and .

**IBS for related samples**

Nevertheless, can be used to infer the genetic relatedness for related samples, such as sib pairs,

a decayed LD compared to the founder population.

## 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 . 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 paper6.

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

When takes the value of 1, otherwise 0. The term indicates the decay of LD compared with the parental population.

**(Eq 3)**

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

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 UKB11.

# Discussion

References

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).

2. Visscher, P. M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).

3. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

4. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

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

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

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

8. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–73 (2010).

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

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

11. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).