**Title: A mathematically tractable framework for the inference of missing heritability**

**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 casual 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 central topic in statistical genetics1. From chip data to WGS data has been used to address the issue, and the statistical methods has also been proposed to improved the estimation. The method is largely on linear mixed model (LMM), which is powerful but obscure in its statistical properties.

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 , imputation2 and WGS.3

Given even large scale data, LMM is not that feasible in terms of computation, while Haseman-Elston regression, a least-squares method, have been employed to estimate the SNP-heritability4,5. In addition, HE promises analytical results for the estimated SNP-heritability. Under the HE framework, which can give (as shown below) analytical solution for and , it gives clear solution for this trends. Although provides upper-bound for , the estimation of and are largely orthogonal. So, provides much accurate estimation of heritability, but provides insight for the possible distribution of allele frequency of the causal SNPs.

**Methods**

**Heritability**

under the infinitesimal model, the high-order term trades off each other, .

Although there are various possible 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 sib pair in linkage analysis, and via unrelated samples.

The current observation supports .

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.

**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 4. 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 6. 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.

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.

**Estimator I: IBS for unrelated samples**

For a single locus, orthogonal coding scheme can separate the genetic variance into additive and dominance genetic variance 7. 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**)

in which is a diagonal matrix, via which weights can be introduced, and and is the trace of .

i) When there is no weight introduced for the relatedness, an identity matrix, . It is often used in human genetics for the estimation of , denoted as thereafter.

ii) When the weight is introduced, for example and . In parallel to the unweighted form, the weighted form of SNP-based heritability, denoted as thereafter, leads to “genomic heritability”, a concept more often employed in animal genetics 8.

Using the stochastic modeling method introduced by Chen 5, as promised by Haseman-Elston regression, the analytical result for the estimated SNP-heritability is

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.

The standard error of the estimated heritability is .

**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.

**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.

**IBD for sib pair**

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 it over the genome we have . As expected, the averaged LD is Fitting IBD into Eq 1, we can have conventional linkage analysis.

**(Eq 3)**

in which indicating the averaged linkage between the marker and the causal variants and the averaged linkage between a pair of markers.

The sampling variance of is , often far larger than .

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

**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 casual 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.

**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 casual 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.



