**Title**:Structure of SNP-based Heritability

**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

**Abstract**

SNP-based heritability that quantifies the phenotypic variation captured by genome-wide single nucleotide polymorphism (SNP) markers is a device towards the understanding of complex traits. Given many empirical results published in recent years, the exact structure of SNP-based heritability remains a state of obscurity. In this study, using a modified Haseman-Elston regression (HE), a model-free method that promises analytical results, we have established the formal genetic interpretation of SNP-based heritability by crystalizing its mathematical structure, for both additive and dominance genetic components, respectively. The analytical results were shorthanded by matrix algebra and easy for further extension. Simulation examples and real data examples using UK Biobank and 1KG data is demonstrated.

**Keywords**: complex traits, SNP-based heritability, genomic-heritability, Haseman-Elston regression, genetic relationship matrix.

**Introduction**

As a parameter of broad interest and utility, SNP-based heritability () is a device towards various genetic applications, such as for the inference of “missing heritability” and genetic architecture (Manolio *et al.*, 2009; Yang *et al.*, 2010). For example, comparison has been implemented to evaluate via realized quantity of under different SNP markers (Yang *et al.*, 2015) and different methods (Speed *et al.*, 2017). Endeavours have been casted to crystalize under the context of the maximum likelihood framework (Lee and Chow, 2014; de los Campos *et al.*, 2015; Legarra, 2016; Steinsaltz *et al.*, 2018); due to the obscure nature of restricted maximum likelihood (REML), even though after nearly a decade search by employing larger and larger sample size, the interpretation or comparison of the suggested estimates of remains empirical (Yang *et al.*, 2017).

Besides REML, there is alternative route for the estimation of : the least-squares estimation, known as Hasemen-Elston regression (HE)—which seems less nourished. The original HE is identity by descent (IBD) for linkage analysis (Haseman and Elston, 1972), but has recently been modified for association analysis after its IBD was replaced by identical by state (IBS) (Chen, 2014). In two perspectives the HE route should be investigated or further clarified. In terms of computational complex in the estimation of , REML is approximately cubic of the sample size [] but HE the square of the sample size [], an computational advances that is in need for biobank-scale data (Bycroft *et al.*, 2018; Liu *et al.*, 2018), recent studies are focused on the modified HE and its feasible applications in estimating variance component for (Ge *et al.*, 2017; Sofer, 2017; Ott *et al.*, 2017).

Among many possible approaches to estimating . The central question is what the exact structure of them. So, in this work, we crystalize the mathematical structure of for the additive variance component and for the dominance variance component via HE, which promises analytical results for the structure of the SNP-heritability. In addition to the additive variance component (), we provide the analytical result for SNP-heritability for dominance variance component (). We provide the analytical results for both additive and dominance under the weighted schemes of the genetic relationship matrices (VanRaden, 2008; Goudet *et al.*, 2018); the weighted form of SNP-heritability is often used in animal genetics and referred as “genomic heritability” (de los Campos *et al.*, 2015).

From an alternative perspective, the goal of the study can be understood as below. Of a trait its real heritability is the “inner mass”, but is perceived “mass” upon the instrument (such as SNPs) and the method (such as HE). In addition, “inner mass” is define regardless how heritability is measured in practice. In this analysis, we will explore the relationship between the “inner mass” – the reference heritability and the “measured mass”— SNP-heritability.

**Materials and Methods**

**Data description**

1000 Genome reference populations and UK Biobank resource has been employed to demonstrate the results (Bycroft *et al.*, 2018; Altshuler *et al.*, 2010). Detailed analysis of these data will be mentioned in the results section below.

**Linear model**

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, 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 cross-product 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 (VanRaden, 2008). 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.

**SNP-based genetic relationship**

For a single locus, orthogonal coding scheme can separate 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. 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). 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 and and is the trace of .

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

ii) When the weigh is introduced (VanRaden, 2008), 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 (de los Campos *et al.*, 2015).

Using matrix algebra, we significantly simplified the symbols for the SNP-heritability. In **Table 1**, we define a set of symbols that will be used to represent , and . The detailed derivation can be found in the **Supplementary Notes**.

Let denote the off-diagonal elements of , and are key important parameters for at least two aspects. Itis easy to see that

**(Eq 3)**

, and we further denote

**(Eq 4)**

a quantity indicating the effective number of markers. As has been standardized, we have

**(Eq 5)**

These elements will be plugged into, then largely simplify, the least-squares estimation (LSE) for Eq 1.

**1KG samples**

We examined the and using HapMap data (Altshuler *et al.*, 2010) in three major ethnicities: Asians (103 Chinese in Beijing, CHB; 104 Japanese in Tokyo, JPT), Europeans (99, Utah residents with Northern and Western European ancestry, CEU; 107 Tuscani in Italy, TSI), and Africans (85 Mende in Sierra Leone, MSL; 108 Yoruba in Ibadan Nigeria, YRI).

When using 907,614 SNP markers that had MAF greater than 0.05 for these HapMap cohorts, was always very close the sample size for each of the six cohorts (**Table 3**). In contrast, the variance of GRM varied across the populations reflecting the different LD structure underlying each cohort.

We only used the alleles having minor allele frequency (MAF) greater than 0.05 in each cohort. The estimated was always very close the actual sample size for each of the six cohorts, indicating that each sample was randomly collected. was of little change. Interestingly, for its largest and smallest numbers were observed in MSL (82,638.06) and YRI (23,707.74), both African populations; while Asian and European population had relative median values about from 38,685.66 in to 46,202.62 in CHB. For , the pattern was similar, but both African cohorts showed largest values, about 225K~229K, while nearly halved in that of the other two ethnicities (**Table 3**). When was weighted by and was weighted by , the results resembled the one that without weight (**Table 4**).

Using least-squares, the parameters in Eq 1 can be estimated via

and

Of note, can be written as , which is the estimate of the regression coefficient in Eq 1.

**The structure of**

The structure of has been derived in the systematic method presented in Chen (2014), but shorthanded in matrix notations.

**(Eq 6)**

A comparative concept for SNP-heritability is “genomic-heritability” as emphasized by de los Campos et al (2015), in which the genetic relationship matrix has been weighted. Under the weighted version for , the genomic-heritability, denoted as herein, has it mathematical structure

**(Eq 7)**

it yields the so-called genomic heritability (de los Campos *et al.*, 2015).

Similarly, if is constructed for dominance effect,

or after has been weighted

So, four SNP-heritability/genomic-heritability can be derived.

Furthermore, for a pair of their, their corresponding SNP-based coheritability can be derived correspondingly.

**(Eq 7)**

Denote the denominator the mean squared LD between a pair of markers.

We can further partition the numerator into two components

There are two terms included, the left term is about the heritability of each causal variant, and the right term is about the correlation between a pair of causal variants.

The second term, a “high-order” term, can be ignored, if the causal variants are randomly distributed along the genome. However, examples, albeit pathological in practice probably, will be constructed to show the existence of the “high-order” term.

When the causal variants are randomly distributed, .

**(Eq 8)**

Of note, , the classic definition of , resembles Eq 7. the correlation matrix between causal variants. Eq 8 only holds when , it is equivalent to set the correlation matrix between causal variants. However, when causal variants are randomly distributed, is reduced to the identity matrix.

Eq 1 can be easily extended to multiple regression by fitting more than one GRM, such as by partitioning the genome by region or by stratifying MAF. The least-squares estimation can be written as

in which is a matrix, the first column is 1 and the rest columns for GRMs. .

, in which , indicating the inverse of the covariance for a pair of loci sampled from GRM and GRM , respectively.

and

, mark: , we have .

Mark: .

**matrix**

Expectation for single MAF bin

The whole MAF heritability is estimated as

When we only use a MAF bin , the heritability is estimated as

in which can be estimated directly from the data. If the QTLs’s effect are not associated with MAF, for consecutive MAF bins, we have

The denominator can be estimated directly from the data , and the numerator can be approximated by . We can test it using UKBiobank.

It can be estimated below:

1. impute the data to 1KG,
2. construct using markers with their MAF inside the bin.
3. construct using all markers;

LD-dependent?(Gazal *et al.*, 2017)

**The analytical results for SNP-based heritability**

Although for a single locus, the additive and dominance genetic relationship are orthogonal to each other, but not for a pair of loci in LD. In order to give a much clearer derivation, we only consider one genetic relatedness in the HE otherwise the analytical result is not easy to derive.

We have the analytical result for the regression coefficient for Eq 1 when only is present, the additive SNP-heritability can be found as

(**Eq 3)**

When there is no weight, , as often used in human genetics, we have a simplified expression

(**Eq 4**)

, an matrix, indicating how the hidden causal variants are in LD with each other via the marker; represents the LD between the marker and each of the causal variants (Table 1). . Although is unobservable, it is likely to be a sparse matrix. is an vector for 1.

The standard error of is , in which is as presented above. Of note, the corresponding standard error of the counterpart estimate of SNP-heritability via REML was , differing by a factor of (Visscher *et al.*, 2014), indicating slightly poor statistical efficiency of HE compared with MLE.

**Connection between heritability and**

The classic definition of heritability can be expressed in matrix (Lynch and Walsh, 1998),

**(Eq 9)**

in which a symmetric square matrix and its element characterizing the LD structure between causal loci.

**Genetic architecture**

The difference between and is that of in Eq 9 and in Eq 5, and between and is that of and . Under NGA that all causal variants are randomly distributed along the genome and their effects follow a normal distribution , in particular, the effect size is independent of allele frequency or LD, both and can be simplified:

**(Eq 9)**

and

**(Eq 10)**

in which is the averaged squared LD between the SNP markers and the causal variants and the averaged squared LD between SNPs. Equation 9 is used as a device in linear mixed model system; equation 10 is probably the most often used device in the inference of genetic architecture (Yang *et al.*, 2017).

Of note, under the NGA, the weighted and unweighted version of SNP-heritability were equivalent.

Alternatively, if not more ubiquitously, causal variants are not randomly distributed along the genome. Under SGA, the analytical result of is as Eq 3, with weight, or Eq 4, without weight has no linear transform between towards . We demonstrate that under NGA, with and without weights are theoretically identical, and consequently should not be statistically different. The difference between Eq 3 and Eq 4 is the crux of NGA and NNGA.

**Numerical example for the connection between the estimated heritability and the inner heritability**

Consider a special case that of two equal-frequent QTLs, both of which are genotyped. Their effect size are both 1, and the LD between them is . The heritability can be written as

and .

After rearrangement, . Although it is a very special case, it indicates that “inner” heritability and the estimated SNP-heritability is not necessary equal. O note, for this case, as the allele frequency is the same, introducing weight will not differ and .

“/Users/gc5k/manuscript/Linkage\_Association/submit/FrontGenet/2qtl”

**Numerical examples for SNP-based heritability**

**The data on cluster linfeng/ukb/British/NGA\_test**

**5000 sample, 2 loci, random maf, dprime -1 to 1.**

Given the analytical results, we first validated their accuracy in simulation studies. We simulated a pair of loci, both of which had their allele frequency randomly distributed, and they were in LD, which was capped by their allele frequencies. The genetic effects were also simulated randomly. The heritability, “inner mass”, was set to 0.5. Similarly, we also tested dominance model. The simulation results were consistent with our **Eq 3-6** as derived above (data not shown; /public/home/xuhm/linfeng/ukb/EqSimu/NGA\_NNGA & /public/home/xuhm/linfeng/ukb/EqSimu/NGA\_NNGA\_dom).

Given the analytical results above, we compared and evaluated the discrepancy between the real heritability, the “inner mass”, and the SNP-based heritability, the “measured mass”. To have manual example, we only consider a pair of casual variants only. In order to give a clear picture how SNP-based heritability worked. We provided a very simple numerical example for a pair of loci. The allele frequencies for these two loci were and , respectively, and the additive effect were and , respectively. The squared correlation between them was , and the maximum of was 0.577. We have the realized matrices below

, , , , and ; for the weighted SNP-based heritability, , , whereas the unweighted , and . Given and , the range for is -0.577 to 0.577.

We gave a manual example for ,

we had “inner mass” , the “measured mass” without weight is

and the weighted one is

The full range of the result was show in Figure X below.

This above example shown the discrepancy among three kinds of heritability because the genetic effects were negative associated with allele frequency, simulating signature of selection. In contrast, we simulated another example, in which and , so the genetic effect is not associated to the allele frequency. Under this example, when three estimates had the same value (Figure X /public/home/xuhm/linfeng/ukb/EqSimu/heProperties).

One conclusion from this simple simulation was that the estimated SNP-heritability could be smaller, in particular even larger, than the real heritability. However, this discrepancy will disappear when the trait is controlled by many loci.

Similarly, we can define the dominance genetic relationship, , between a pair of individuals, with and without weights,

(**Eq 3**)

in which . When there is no weight, , and ; when there is weight, and .

This study is to derive the precise genetic interpretation of the regression coefficient, which can be converted to () for additive (dominance) genetic variance. According to the linear model Eq 1, the regression coefficient is estimated as . The regression coefficient is denoted as and , regarding to and .

**UK Biobank analysis**

We used UK Biobank to investigate the impact of underlying genetic architecture and the possible different genetic architecture between Chinese (CH) and UK whites (UKW). There were 1,495 Chinese in UKBiobank, we further removed 60 samples who were genetically admixed and 1,435 samples were used for analysis. We only used 324,012 SNPs that were passed quality control (minor allele frequency > 0.01) in each ethnicity. In addition, we also sampled 1,000 UK whites as a comparison. The genetic relationship matrices were generated with or without weight. Then we estimated SNP-heritability for 65 quantitative traits (/public/home/xuhm/linfeng/ukb/HERE/CH & /public/home/xuhm/linfeng/ukb/HERE/UK).

For CH, the effective sample size was 1425.7, indicating little relatedness – as expected. The effective number of marker was of 28,841 and 24,631, with and without weight, respectively; was of 87081 and 54585, with and without weight, respectively. For UKW, the effective sample size was 990.1; its effective number for was of 43,216 and 32,417, respectively, and for was of 114,395 and 60,921, respectively. So, the selected SNP markers had on average much higher LD for CH and for UKW.

Although the sampling variance was large, but the overall comparison between with and without weight could reveal the neutrality of the traits. Except for a few traits, for both CH and UK, the estimates with or without weights were very consistent, the correlation with and without weight was 0.993 for CH, and 0.972 for UKW and indicating no overall rejection of neutrality of the traits.

**Discussion**

The purpose of the study is to derive the connection between heritability and SNP-heritability, and to clarify the circumstance that SNP-heritability can be approximated by heritability under the neutral genetic architecture. Using modified HE and based on our results established previously (Chen, 2014, 2016), we derived the mathematical expression of SNP-heritability for both additive and dominance terms. However, due to materials used – high density SNP markers and the method used – numerator relationship matrix, SNP-heritability differ from heritability by the way LD are represented. In its classic definition, heritability is perfectly measured on QTLs directly, but in estimation the LD to the casual variants is tagged via markers.

In addition, we introduced the matrix algebra that can shorthand the structure of SNP-heritability succinctly. Given the crystallization of SNP-heritability, it also paves the road to construct effective hypotheses test for genetic architecture ***in silicon***. The general idea in testing varying genetic architecture is to introduce proper weights, a commonly adopted technique in linear algebra. Although we only demonstrated one kind of weights, obviously a family of weights can be introduced systematically according to the genetic architecture of interest. It relies on the further work for refined exploration of genetic architecture and real data will be tested in the future.

For UK Biobank data, within either UK-CH or UKW the weighted and unweighted forms SNP-heritability similar to each other for 65 quantitative traits analyzed. However, the difference between them were more visible. So, in general, we conclude at the whole-genome scale, the dependency between allele frequency and casual effects should be marginal.

**Acknowledgements**

This work was supported by National Nature Science Foundation of China (31771392 to GBC), and Zhejiang Provincial People’s Hospital Research Startup (ZRY2018A004 to GBC). This research has been conducted using the UK Biobank Resource (application number 41376).

**Conflict of interests**: None.

**Literature cited**

Altshuler,D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–8.

Bycroft,C. *et al.* (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.

Chen,G.-B. (2016) On the reconciliation of missing heritability for genome-wide association studies. *Eur. J. Hum. Genet.*, **24**, 1810–6.

Gazal,S. *et al.* (2017) Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.*, **49**, 1421–1427.

Ge,T. *et al.* (2017) Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.*, **13**, e1006711.

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

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

Lee,J.J. and Chow,C.C. (2014) Conditions for the validity of SNP-based heritability estimation. *Hum. Genet.*, **133**, 1011–22.

Legarra,A. (2016) Comparing estimates of genetic variance across different relationship models. *Theor. Popul. Biol.*, **107**, 26–30.

Liu,S. *et al.* (2018) Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell*, **175**, 347–359.

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.

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

Ott,A. *et al.* (2017) tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic Acids Res.*, **45**, e178.

Sofer,T. (2017) Confidence intervals for heritability via Haseman-Elston regression. *Stat. Appl. Genet. Mol. Biol.*, **16**, 259–273.

Speed,D. *et al.* (2017) Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, **49**, 986–992.

Steinsaltz,D. *et al.* (2018) Statistical properties of simple random-effects models for genetic heritability. *Electron. J. Stat.*, **12**, 321–358.

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

Visscher,P.M. *et al.* (2014) Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genet.*, **10**, e1004269.

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

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

Yang,J. *et al.* (2017) Concepts, estimation and interpretation of SNP- based heritability. *Nat. Genet.*, **49**, 1304–10.

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

**Figure legends**

**Figure 1 A minimal example for illustrating how SNP-heritability upon allele frequency and linkage disequilibrium.** The reference heritability is as defined in equation 9, and the estimates were scaled to have 0.5.



**Figure 2 SNP-heritability for 1436 UK Biobank Chinese (CH) and 1,000 UK whites (UKW) for 65 quantitative traits.** Only core SNPs that were minor allele frequencies greater than 0.01 in both CH and UK were used for the estimation. The first row: **Left**) SNP-heritability, without weight (x-axis) and with weight (y-axis), for UKBiobank Chinese over 65 quantitative traits; **Right**) estimated SNP-heritability for 1,000 UK Biobank whites for 65 quantitative traits. The second row: **Left**) Contrast for the SNP-heritability, without weight, for CH and UKW; **Right**) Contrast for SNP-heritability, with weight, for CH and UKW.



**Table 1 Symbols for the structure of various SNP-based heritabilities**

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Additive-related** | **Dominance-related** |
| Effect | is the vector of elements and . and are additive and dominance effects, respectively, of the locus. | is the vector of elements and its element is . |
| Allele frequency | a diagonal matrix, and . | a diagonal matrix, and . |
| Linkage disequilibrium | represents the linkage disequilibrium between the marker and each of the causal variants. | represents the squared correlation between the marker and each of the causal variants. |
| is an matrix, and . | is an matrix, and . |
| Weighting | , an vector. | , an vector. |
| a diagonal matrix, and . | a diagonal matrix, and . |
| Allele frequency | a diagonal matrix, and . | a diagonal matrix, and . |

Of note, ,, ,and are all hidden variables; the number of causal loci, the number of SNP loci.

**Table 2 The matrix presentation of SNP-heritability for additive and dominance genetic components**

|  |  |  |
| --- | --- | --- |
|  | **Additive** | **Dominance** |
| **Single traits** | | |
| No weight |  |  |
| Weighted |  |  |
| **A pair of traits** | | |
| No weight |  |  |
| Weighted |  |  |

**Table 3 Evaluation of and using HapMap populations without weight**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population** |  |  |  |  |  |  |  |  |  |
| **Scenario I** |  |  |  |  |  |  |  |  |  |
| CHB (103) | -0.0098 | 103.36 |  | 907,614 | 2.16e-5 | 46,202.62 |  | 9.58e-6 | 104,345.31 |
| JPT (104) | -0.0097 | 104.27 |  | 2.95e-5 | 43,603.59 |  | 1.04e-5 | 96,411.15 |
|  |  |  |  |  |  |  |  |  |
| CEU (99) | -0.0102 | 99.37 |  | 2.58e-5 | 38,685.66 |  | 3.98e-5 | 106,740.08 |
| TSI (107) | -0.0094 | 107.31 |  | 2.35e-5 | 44,933.03 |  | 9.01e-6 | 110,963.01 |
|  |  |  |  |  |  |  |  |  |
| MSL (85) | -0.0118 | 85.46 |  | 4.22e-5 | 23,707.74 |  | 4.35e-6 | 229,848.07 |
| YRI (108) | -0.0093 | 108.43 |  | 1.21e-5 | 82,638.06 |  | 4.44e-6 | 225,221.38 |
| **Scenario II** (MAF > 0.05) | | | | |  |  |  |  |  |
| CHB (103) | -0.0098 | 103.36 |  | 824,710 | 2.47e-5 | 40,520.49 |  | 1.13e-5 | 80,725.48 |
| JPT (104) | -0.0097 | 104.27 |  | 820,982 | 2.61e-5 | 38,244.57 |  | 1.23e-5 | 80,997.89 |
|  |  |  |  |  |  |  |  |  |  |
| CEU (99) | -0.0102 | 99.37 |  | 851,342 | 2.79e-5 | 35,841.94 |  | 1.04e-5 | 96,601.30 |
| TSI (107) | -0.0094 | 107.31 |  | 852,354 | 2.42e-5 | 41,277.66 |  | 1.00e-5 | 99,675.29 |
|  |  |  |  |  |  |  |  |  |  |
| MSL (85) | -0.0118 | 85.46 |  | 816,393 | 4.58e-5 | 21,817.77 |  | 5.20e-6 | 192,279.34 |
| YRI (108) | -0.0093 | 108.43 |  | 819,671 | 1.37e-5 | 72,781.09 |  | 5.27e-6 | 189,590.34 |

**Table 4 Evaluation of and using 1KG samples with weights**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population** | |  |  | |  |  |  |  |  |  |  |
| **Scenario I** | |  |  | |  |  |  |  |  |  |  |
| CHB (103) | | -0.0098 | 103.36 | |  | 907,614 | 2.81e-5 | 35,552.12 |  | 1.64e-5 | 61,095.14 |
| JPT (104) | | -0.0097 | 104.28 | |  | 2.92e-5 | 34,231.47 |  | 1.77e-5 | 56,463.71 |
|  | |  |  | |  |  |  |  |  |  |
| CEU (99) | | -0.0102 | 99.39 | |  | 3.12e-5 | 32,043.18 |  | 1.47e-5 | 67,815.78 |
| TSI (107) | | -0.0094 | 107.33 | |  | 2.75e-5 | 36,318.51 |  | 1.43e-5 | 70,057.59 |
|  | |  |  | |  |  |  |  |  |  |
| MSL (85) | | -0.0118 | 85.47 | |  | 4.75e-5 | 21,066.03 |  | 7.47e-6 | 133,810.81 |
| YRI (108) | | -0.0093 | 108.43 | |  | 1.54e-5 | 65,115.82 |  | 7.75e-6 | 129,057.08 |
| **Scenario II** (MAF > 0.05) | | | | | | |  |  |  |  |  |
| CHB (103) | -0.0098 | | 103.36 |  | | 824,710 | 2.89e-5 | 34,655.78 |  | 1.64e-5 | 60,805.68 |
| JPT (104) | -0.0097 | | 104.28 |  | | 820,982 | 2.99e-5 | 33,411.80 |  | 1.78e-5 | 56,193.86 |
|  |  | |  |  | |  |  |  |  |  |  |
| CEU (99) | -0.0102 | | 99.37 |  | | 851,342 | 3.16e-5 | 31,635.78 |  | 1.48e-5 | 67,628.51 |
| TSI (107) | -0.0094 | | 107.32 |  | | 852,354 | 2.79e-5 | 35,782.68 |  | 1.43e-5 | 69,852.85 |
|  |  | |  |  | |  |  |  |  |  |  |
| MSL (85) | -0.0118 | | 85.47 |  | | 816,393 | 4.81e-5 | 20,811.19 |  | 7.51e-6 | 133,189.46 |
| YRI (108) | -0.0093 | | 108.44 |  | | 819,671 | 1.57e-5 | 63,781.53 |  | 7.78e-6 | 128,491.57 |

**Notes**: CHB, Chinese in Beijing; JPT, Japanese in Tokyo; MSL, Mende in Sierra Leone; YRI, Yoruba in Ibadan, Nigeria; CEU, Utah residents with Northern and Western European ancestry; TSI, Tuscani in Italy.

**Table 4 1000 samples for SNP markers with MAF > 0.01 in all three populations**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample size | Marker |  |  | **(weighted)** |  |  | **(weighted)** |
| 1,000 Africans | 324,012 |  | 1,925.44 | 35,82.87 |  | 27,486.49 | 74,809.44 |
| 1,000 Chinese | 324,012 |  | 18,665.02 | 20,286.95 |  | 79,020.11 | 53,832.35 |
| 1,000 UK | 324,012 |  | 43,220.85 | 32,412.13 |  | 114,417.03 | 60,919.29 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Center | Marker |  | (weighted) |  | **(weighted)** |
| Manchester | 476499 | 1.81E-05 | 2.75E-05 | 55127.40 | 36343.28 |
| Oxford | 476499 | 1.76E-05 | 2.69E-05 | 56737.27 | 37126.14 |
| Cardiff | 476336 | 1.87E-05 | 2.84E-05 | 53452.78 | 35196.40 |
| Glasgow | 476558 | 1.88E-05 | 2.88E-05 | 53260.90 | 34697.64 |
| Edinburgh | 476486 | 1.84E-05 | 2.81E-05 | 54430.95 | 35595.42 |
| Stoke | 476708 | 1.90E-05 | 2.82E-05 | 52681.77 | 35479.49 |
| Reading | 476297 | 1.67E-05 | 2.59E-05 | 59735.02 | 38584.27 |
| Bury | 476672 | 1.77E-05 | 2.70E-05 | 56631.88 | 36996.07 |
| Newcastle | 476670 | 1.78E-05 | 2.72E-05 | 56210.72 | 36789.19 |
| Leeds | 476620 | 1.72E-05 | 2.64E-05 | 58116.22 | 37831.07 |
| Bristol | 476454 | 1.71E-05 | 2.63E-05 | 58423.16 | 37969.25 |
| Barts | 476010 | 1.88E-05 | 2.76E-05 | 53107.31 | 36200.93 |
| Nottingham | 476440 | 1.73E-05 | 2.65E-05 | 57740.05 | 37673.44 |
| Sheffield | 476740 | 1.76E-05 | 2.69E-05 | 56811.73 | 37122.83 |
| Middlesborough | 476575 | 1.82E-05 | 2.76E-05 | 54883.24 | 36296.19 |
| Hounslow | 476166 | 1.71E-05 | 2.63E-05 | 58394.84 | 38062.47 |
| Croydon | 476209 | 1.69E-05 | 2.61E-05 | 59041.98 | 38360.47 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Additive |  |  |  |  |
| UK1K | 0.602 |  |  |  |  |
|  | 0.4043 |  |  |  |  |
|  |  |  |  |  |  |
| CH1K | 1.605 |  | CHUK (1435) | 1.809 |  |
|  | 1.57 |  |  | 1.57 |  |
|  |  |  |  |  |  |
| AF1K | 0.2288 |  | AFUK (2819) | 0.7689 |  |
|  | 0.3337 |  |  | 0.7507 |  |

CH1K and AF1K only used SNPs passed QC of MAF > 0.01 in three ethnicities, and in total 324,012 SNPs are found.

CHUK: using projected PC with 1kg as the reference to find Chinese and using 324012 SNPs above.

AFUK: using projected PC with 1KG as the reference to find African and using 324012 SNPs above.