Theoretical Interpretation of Genetic Association Studies in Admixed Populations

Hanbin Lee[1,2,§,†] and Moo Hyuk Lee[1,§]

[1]Department of Medicine, Seoul National University College of Medicine

[2]Department of Mathematical Sciences, Seoul National University

[§]These authors contributed equally.

[†]Corresponding Author: hanbin973@snu.ac.kr

Abstract

Admixed populations are formed through complicated evolutionary processes resulting in heterogeneous linkage disequilibrium patterns. This hampered the interpretation of genetic association studies followed by the development of various methods tailored for the purpose without clear guidance. In this work, we propose a unifying theory built upon population genetics and causal inference that brings a straightforward interpretation of GWAS findings in admixed populations. Furthermore, the theory establishes the connection between distinct methods in admixed GWAS. Finally, we propose two novel tests motivated by our theory that combine association and admixture signals.

Keywords: Genome-wide association study, GWAS, Admixture, Admixture Mapping, Population Genetics, Causal Inference

1 Introduction

Populations that experienced recent admixture events provides an unique opportunity to understand the genetic basis of complex traits. When two or more populations are admixed, individuals inherit their genetic material from multiple ancestral populations. The more recent the admixture event is, the longer the range of blocks of genetic materials originating from the same ancestry is. Hence, using these blocks (hereafter, admixture blocks) for genome-wide scan remained popular until high-density single nucleotide polymorphism (SNP) markers became widely available.

Admixture mapping (ADM) measures the association between the number of haplotypes from a particular ancestry (or the *local ancestry*) at a region with the trait to localize genetic signals [1, 2]. When the underlying causal variant is more common compared to others, an individual is more likely to carry the causal variant with increasing number of haplotypes inheriting from that population. The approach has been successful in detecting disease loci associated from obesity to prostate cancer [3, 4].

Shortly after the success of genome-wide ADM, *genome-wide association study* (GWAS), a association mapping approach became popular [5]. As SNP LD blocks were generally shorter than admixture blocks, association mapping provided higher resolution in localizing the susceptibility loci compared to admixture mapping. However, GWAS have been performed predominantly on populations from single continental origins such as Europeans and East Asians [6, 7].

Despite the decline in diversity in genome-wide genetic studies, statistical methods for admixed populations have been steadily developed in the past decade. Pasaniuc and colleagues, together with a novel combined test for association and admixture, benchmarked various methods for detecting genetic signals in admixed populations [8]. As cohorts comprised of admixed populations (e.g. African Americans) gradually grew in the past years, methods such as Tractor has been proposed [9]. These methods differ in several ways. An important topic is whether to include local ancestry into the statistical model [10, 11, 12]. However, these arguments are largely based on simulations and no theoretical guarantees have been made. As a result, there have been no agreement in which method to use.

In this work, we propose a unifying theory built upon population genetics and causal inference that provides a general guidance for GWAS in admixed populations. We assume that the population follows the famous admixture model [13, 14] and derive the target parameters (hereafter, estimands) of distinct methods proposed for admixed GWAS. The result provides provable interpretation of statistical estiamtes delivered from admixed populations which gives a general guide on how meta-analysis should be performed when incorporating admixed populations. As a byproduct, estimands of admixture mapping is derived. Furthermore, statistical efficiency and power are analyzed under

the asymptotic regime. Finally, we propose provably powerful tests designed by theory to perform admixed GWAS that combines association and admixture signals. Best to our knowledge, the novel tests are the first to combine the two designs in quantitative traits.

2 Results

Trait and Population Models

The central argument of this work is based on the distinction between the causal variants C_{ij} and the marker variants M_{ik}. $i = 1, \ldots, n_I$, $j = 1, \ldots, n_J$ and $k = 1, \ldots, n_K$ are the indices of individuals, causal variants and marker variants respectively. The boldface variables are the vector of variants $\mathbf{C}_i = \{C_{ij}\}_{j=1,\ldots,n_J}$ and $\mathbf{M}_i = \{M_{ik}\}_{k=1,\ldots,n_K}$. Y_i denotes the quantitative trait of an individual which is a linear function respect to the causal variants.

$$Y_i = \alpha_0 + \sum_{j=1}^{n_J} C_{ij}\alpha_j + \epsilon_i \tag{1}$$

where α_j is the causal effect size of variant j and ϵ_i is the random error such that $\mathbb{E}[\epsilon_i \mid \mathbf{C}_i] = 0$. We call this model the *causal additive model* (CAM) [15].

Suppose that there are n_L ancestral populations indexed by $l = 1, \ldots, n_L$. The admixture model we consider is determined by the ancestral reference allele frequency of markers g_{lk} and the individual's ancestral proportion P_{il} (or the *global ancestry*). $\mathbf{P}_i = \{P_{il}\}_{l=1,\ldots,n_L}$ is the random vector of global ancestry ($\sum_{l=1}^{n_L} \mathbf{P}_i = 1$). For causal variants, we write f_{lj} for the reference allele frequencies. We define admixture blocks $b = 1, \ldots, n_B$ across the genome where each block has a ancestral origin called *local ancestry* $\mathbf{L}_{ib} = \{L_{ibl}\}_{l=1,\ldots,n_L}$. Here, L_{ibl} is the number of haplotypes of an individual inheriting from ancestral population l. For each variant j and k, they belong to a particular block b in which we write as $j, k \in b$ with a slight abuse of notation. Variables are often written respect to the parental origin (or haplotypes). For $\mathbf{L}_{ib}$, $\mathbf{C}_i$ and $\mathbf{M}_i$, the superscript $\mathbf{h} = \mathbf{m}, \mathbf{p}$ (maternal and paternal) denotes the haplotype. For example, $C_{ij}^{\mathbf{h}} = 0, 1$ is the reference allele count of the causal variant j at the $\mathbf{h}$-th haplotype.

The value at a given locus of an individual M_{ik} (or C_{ij}) is determined through a two step process. In the first step, an individual inherits its genetic material at the locus from a particular ancestral population according to its global ancestry. Next, the reference allele count is assigned according to the reference allele frequency of the selected ancestral population. This process is summarized in the following causal graph (**Figure 1**). The bidirectional arrow denotes the linkage between M_{ik} and C_{ij}. The famous Pritchard-Stephens-Donnelly (PSD) model follows the described process with additional probabilistic assumptions [13]. The detailed mathematical description is deferred to **Online Methods**.

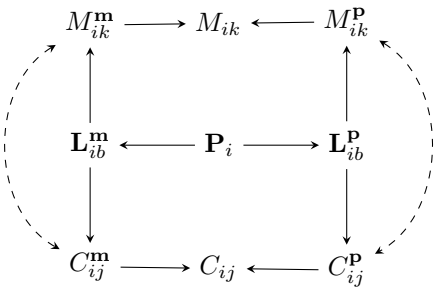

Figure 1: The causal graph of population genetic variables in admixed populations.

Marginal Effect Sizes in Admixed Populations

Let h_{ljk} be the joint allele frequency of C_{ij} and M_{ik} in the ancestral population l. The *linkage disequilibrium* (LD) covariance is $D_{ljk} = h_{ljk} - f_{lj}g_{lk}$. Then the marginal effect size of marker M_{ik} restricted to an ancestral population l is the product of the LD covariance and the causal effect size [15].

$$\beta_{lk}^{\text{anc}} = \sum_{j=1}^{n_J} \frac{D_{ljk}}{g_{lk}(1-g_{lk})}\alpha_j \tag{2}$$

However, the marginal effect size in admixed GWAS has been unknown.

Two famous regression estimators have been advocated in performing association studies in admixed populations. The first estimator regresses the trait Y_i with multiple reference allele counts for each ancestral population [8, 9]. Let M_{ikl} be the number of reference allele counts from ancestry l. Therefore, $M_{ik} = \sum_{l=1}^{n_L} M_{ikl}$. Together, local ancestry $\mathbf{L}_{ib}$ for $b \in k$ is included to the regression.

$$Y_i = \beta_0 + \sum_{l=1}^{n_L} M_{ikl}\beta_{lk} + \sum_{l=2}^{n_L} L_{ibl}\gamma_{lk} + \sum_{l=2}^{n_L} P_{il}\delta_{lk} + \zeta_{ik} \tag{3}$$

Note that the latter two summations start from $l = 2$ to avoid multicolinearity due to $\sum_{l=1}^{n_L} L_{ibl} = 2$ and $\sum_{l=1}^{n_L} P_{il} = 1$. Then,

$$\beta_{lk} = \beta_{lk}^{\text{anc}} \tag{4}$$

Hence, regression (3) recovers the ancestry specific marginal effect sizes that would have been obtained if the regression had been performed exclusively in a particular ancestral population. Furthermore,

the coefficient of L_{ibl}, the local ancestry, is

$$\gamma_{lk} = \sum_{j=1}^{n_J} \left(\frac{f_{lj} - h_{ljk}}{1 - g_{lk}} - \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} \right) \alpha_j \tag{5}$$

Hence, the local ancestry captures the effect of the causal variants that is missed by the marker. This formula also uncovers the source of the signal of admixture mapping.

The second estimator does not separate the minor allele counts and does not include local ancestry [8, 10]. Only global ancestry is incldued as covariates.

$$Y_i = \beta_0^{\text{agg}} + M_{ik}\beta_k^{\text{agg}} + \sum_{l=1}^{n_L} P_{il}\delta_{lk}^{\text{agg}} + \zeta_{ik}^{\text{agg}} \tag{6}$$

This regression recovers a variance weighted mean of $\beta_{lk}(\mathbf{P}_i)$

$$\beta_k^{\text{agg}} = \frac{\mathbb{E}[\beta_k(\mathbf{P}_i)\text{Var}(M_{ik} \mid \mathbf{P}_i)]}{\mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)]} \tag{7}$$

where

$$\beta_k(\mathbf{P}_i) = \sum_{j=1}^{n_J} \frac{\sum_l D_{ljk}P_{il} + \sum_l g_{lk}f_{lj}P_{il}(1 - P_{il}) - \sum_{l \neq l'} g_{lk}f_{l'k}P_{il}P_{il'}}{\sum_l g_{lk}(1 - g_{lk})P_{il} + \sum_l g_{lk}^2 P_{il}(1 - P_{il}) - \sum_{l \neq l'} g_{lk}g_{l'k}P_{il}P_{il'}} \alpha_j \tag{8}$$

Although equation (8) is complicated at glance, for an individual that inherits all of its genome from a single ancestral population, i.e. $P_{il} = 1$ and $P_{il'} = 0$ for $l \neq l'$, $\beta_k(\mathbf{P}_i) = \beta_{lk}^{\text{anc}}$ (see **Online Methods**).

The Role of Local Ancestry

The inclusion of local ancestry has been a central discourse in association mapping of admixed populations [10, 11, 12]. We first analyze the effect of local ancestry on the interpretability of the marginal effect sizes. Recently, Hou et al. argued that excluding local ancestry from regression (3) is better in terms of power to detect marginal effect size heterogeneity across ancestries [12]. We show that this comes at the cost of lost interpretability. When the local ancestry is omitted in regression (3), β_{lk} is no more β_{lk}^{anc}, and even worse, becomes a combination of parameters from multiple ancestries (see **Online Methods**). It has a closed form formula when $n_L = 2$ which is

$$\begin{aligned}
\beta_{1k}^{\text{nolanc}} &= \beta_{1k}^{\text{anc}} + \sum_{j=1}^{n_J} \frac{g_{1k}g_{2k}(1 - g_{2k})\mathbb{E}[P_{i1}P_{i2}]\mathbb{E}[P_{i2}]}{\Lambda_{12}} \left(\frac{f_{1j} - h_{1jk}}{1 - g_{1k}} - \frac{f_{2j} - h_{2jk}}{1 - g_{2k}} \right) \alpha_j \\
\beta_{2k}^{\text{nolanc}} &= \beta_{2k}^{\text{anc}} + \sum_{j=1}^{n_J} \frac{g_{1k}(1 - g_{1k})g_{2k}\mathbb{E}[P_{i1}]\mathbb{E}[P_{i1}P_{i2}]}{\Lambda_{12}} \left(\frac{f_{2j} - h_{2jk}}{1 - g_{2k}} - \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} \right) \alpha_j
\end{aligned} \tag{9}$$

As a result, comparing β_{lk} to conclude inter-ancestry difference becomes invalid. The test is valid if and only if the marker is causal itself or unlinked with any causal variants.

Another important issue is statistical power. Earlier studies have found that inclusion of local ancestry alters the power of discovery. In **Online Methods**, we provide a detailed analysis in terms of asymptotic efficiency. Briefly, it depends not only on the effect size heterogeneity but also to reference allele frequencies and LD patterns.

Aggregating Signals from Association and Admixture

Assuming that the genotyped marker M_{ik} is itself a causal variant and is not in linkage with other causal variants, one can derive a powerful test that aggregates both association and admixture mapping. A similar idea has been proposed earlier in the case of binary traits [8]. Here, we propose the first example for quantitative traits.

When variant M_{ik} is the only causal variant in block b (i.e. $M_{ik} = C_{ij}$), equation (7) reduces to

$$\beta_{lk}^{\text{anc}} = \alpha_j \tag{10}$$

which is the causal effect size of variant j. Excluding M_{ikl} in regression (3) gives

$$\gamma_{lk} = (f_{lj} - f_{1j}) \cdot \alpha_j \text{ for } l = 2, \ldots, n_L \tag{11}$$

which shows that the two regressions have the same estimand that only differs by a factor of $f_{lj} - f_{1j}$.

Since the reference allele frequency can be obtained from an external reference, $f_{lj} - f_{1j}$ is known *a priori*. This means that we have two different regressions that estimate an essentially the same parameter. Therefore, the information from the two regressions can be combined to improve power through a Z-estimator approach (see **Online Methods**).

Formula (4) and (5) uncover another test that combines both association and admixture signals. The coefficient of local ancestry γ_{lk} is proportional to the causal effect size α_j of the underlying causal variant. Therefore, testing $\gamma_{lk} = 0$ is also a valid test for the genetic signal. Hence, a $(2n_L - 1)$-degrees of freedom test combining n_L coefficients of M_{ikl} and $n_L - 1$ coefficients of L_{ibl} can test the genetic effect. This can be done through standard Wald, log-likelihood ratio and F-test.

3 Discussion

In this work, we derive formal interpretation of regression coefficients of famous methods proposed for association studies in admixed population. The interpretation provides a clear guidance on when to include or exclude covariates such as local ancestry. Further characterization of the variance structure reveals the power properties of the methods. Finally, we propose powerful tests as extensions of existing methods based on the theoretical results we have derived.

A notable character of our work is that regressions are proved rather than assumed at first place. We setup a trait model (the CAM) and the demographic model (the admixture model) and derive the estimators from scratch. This way, we were able to shed light on the properties of the estimators in the light of population genetics and causal inference. We hope that this approach becomes common in genetic methodology development which is likely to promote new findings and enhance interpretability of the results.

There are several limitations in this work. The current setup only considers population structure due to admixture. Therefore, other population processes such as inbreeding and assortative mating is not addressed [16]. Such processes breaks down the conditional independence between haplotypes which is central to our proofs. Second, this work targeted population designs, so the estimands of sibling designs should be addressed [17]. Further work should be made to address these issues.

In the causal inference perspective, the trait model we consider is rather restrictive. The model assumes that the causal effect of the variant is constant across individuals. As heterogeneous causal effects are currently gaining interest in many fields, the extension of the trait model must be studied in the future [18, 19]. In genetic terms, this is essential to understand epistasis, dominance and gene-environment interaction effects under the admixture scenario.

Acknowledgements

This research was supported by a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development (KHDI), funded by the Ministry of Health and Welfare, Republic of Korea. We thank Doc Edge (University of Southern California, USA) for providing helpful advices after reading an early version of the manuscript.

Online Methods

Note that all vectors are column vectors unless mentioned otherwise.

The Admixture Model

As described in **Results**, the reference allele counts C_{ij} and M_{ik} are determined through a two step process in admixed populations (**Figure 1**). Therefore, the joint probability of $(\mathbf{P}_i, \mathbf{L}_{ib}, C_{ij}, M_{ik})$ obeys the following factorization.

$$\mathbb{P}(\mathbf{P}_i, \mathbf{L}_{ib}, C_{ij}, M_{ik}) = \mathbb{P}(\mathbf{P}_i) \cdot \mathbb{P}(\mathbf{L}_{ib} \mid \mathbf{P}_i) \cdot \mathbb{P}(C_{ij}, M_{ik} \mid \mathbf{L}_{ib}) \tag{M1}$$

$\mathbb{P}(\mathbf{P}_i)$ is the usually referred as the prior distribution of global ancestry. In the PSD model, this is given as the Dirichlet distribution [13].

$$\mathbf{P}_i \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{n_L}) \text{ for } \alpha_1, \ldots, \alpha_{n_L} > 0 \tag{M2}$$

In this work, we mostly condition on $\mathbf{P}_i$ so the distribution of $\mathbf{P}_i$ does not alter the conclusions. Hence, we remain agnostic about the distribution of $\mathbf{P}_i$.

To determine $\mathbb{P}(\mathbf{L}_{ib} \mid \mathbf{P}_i)$, we must assume that the local ancestry of two haplotypes are independent given $\mathbf{P}_i$. In other words, we allow population structure by $\mathbf{P}_i$ but no assortative mating or inbreeding within a fixed $\mathbf{P}_i$. Given the independence, $\mathbb{P}(\mathbf{L}_{ib} \mid \mathbf{P}_i)$ follows a multinomial distribution with two trials and the probability of each choice given by the global ancestry $\mathbf{P}_i$.

$$\mathbf{L}_{ib} \mid \mathbf{P}_i \sim \text{Multinomial}(n = 2, p = \mathbf{P}_i) \tag{M3}$$

Determining the joint conditional probability $\mathbb{P}(C_{ij}, M_{ik} \mid \mathbf{L}_{ib})$ will be deferred to the next section when we derive the estimands of the regressions (3) and (6). Here, estimand stands for the parameter in which the regression estimates [cite]. Instead, we only write the marginal conditional expectations $\mathbb{E}[C_{ij} \mid \mathbf{L}_{ib}]$ and $\mathbb{E}[M_{ik} \mid \mathbf{L}_{ib}]$.

$$\mathbb{E}[C_{ij} \mid \mathbf{L}_{ib}] = \sum_{l=1}^{n_L} f_{lj} L_{ibl} \quad \text{and} \quad \mathbb{E}[M_{ik} \mid \mathbf{L}_{ib}] = \sum_{l=1}^{n_L} g_{lk} L_{ibl} \tag{M4}$$

The conditional expectation conditional only on the global ancestry is also important. This is obtained

through the law of total expectation.

$$\begin{aligned}
\mathbb{E}[C_{ij} \mid \mathbf{P}_i] &= \mathbb{E}_{\mathbf{L}}[\mathbb{E}[C_{ij} \mid \mathbf{P}_i, \mathbf{L}_{ib}] \mid \mathbf{P}_i] \\
&= \mathbb{E}_{\mathbf{L}}[\mathbb{E}[C_{ij} \mid \mathbf{L}_{ib}] \mid \mathbf{P}_i] \\
&= \mathbb{E}_{\mathbf{L}}\left[\sum_{l=1}^{n_L} f_{lj} L_{ibl} \;\middle|\; \mathbf{P}_i\right] \\
&= \sum_{l=1}^{n_L} f_{lj} \mathbb{E}_{\mathbf{L}}[L_{ibl} \mid \mathbf{P}_i] \\
&= 2\sum_{l=1}^{n_L} f_{lj} P_{il}
\end{aligned} \tag{M5}$$

and similarly $\mathbb{E}[M_{ik} \mid \mathbf{P}_i] = 2\sum_{l=1}^{n_L} g_{lk} P_{il}$. The second line follows from the conditional independence $C_{ij} \perp\!\!\!\perp \mathbf{P}_i \mid \mathbf{L}_{ib}$ encoded in **Figure 1**.

Models, Estimators and Regressions

Models and estimators are frequently conflated in literature as in the case of linear regression and a data generating process following a linear function. Although the confusion may not have notable consequences when the estimator is directly derived from the model, it is of central importance in our work because we have multiple estimators for a single model.

The model describes the data generating process. The causal additive model postulates that the quantitative trait of interest is generated in terms of an additive function of causal variants.

$$Y_i = \alpha_0 + \sum_{j=1}^{n_J} C_{ij}\alpha_j + \epsilon_i \tag{M6}$$

This is an example of a data generating process that follows a linear functional relationship.

An estimator is an algorithm that estimates the parameter of a given model from data. If we had known which variants were the causal variants, it would have been possible to estimate $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1,\ldots,n_J}$ directly through the *ordinary least square* (OLS) estimator. Let $\mathbf{Y} = \{Y_i\}_{i=1,\ldots,n_I}$ and $\mathbf{X} = \{\mathbf{X}_i\}_{i=1\,\ldots,n_I}$ be the dependent and the explanatory variable respectively. The OLS estimator is

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} \tag{M7}$$

When the model assumes that Y_i is linear respect to $\mathbf{X}_i$, then the OLS estimator is unbiased for the coefficients of $\mathbf{X}_i$. In the causal additive model, the OLS estimator produces a estimate $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\{\mathbf{Z}_i\}_{i=1,\ldots,n_I})$ (according to equation (M7)), a function of the data $\mathbf{Z}_i = (1, Y_i, \mathbf{C}_i)$, that estimates

the model parameter $\boldsymbol{\alpha}$ (here, 1 is the intercept). Explicitly, let $\mathbf{W}_i = (1, \mathbf{C}_i)$ and $\mathbf{W} = \{\mathbf{W}_i\}_{i=1,\ldots,n_I}$. Then,

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}\mathbf{Y} \tag{M8}$$

Since $\widehat{\boldsymbol{\alpha}}$ is an unbiased estimator for $\boldsymbol{\alpha}$, people frequently write equation (M6) to denote the OLS estimator (M8).

In real data, the distinction between the model and the estimator becomes clearer. Causal variants are seldom observed themselves and usually the marker variants are observed. Therefore, regression is performed based on the observed markers as in regression (3) and (6). As the true model is equation (M6), the two regressions are not a description of the data generating process. Rather, they are referring to the OLS estimators, both with Y_i as the dependent variable, $(1, M_{ikl}, L_{ibl}, P_{il})$ and $(1, M_{ik}, P_{il})$ as the explanatory variables respectively. The formulas are given identical to equation (M7).

In summary, a model is the data generating process and an estimator is an algorithm that estimates the parameter of the model using data. Regression is frequently used as a substitute for estimator when the estimator can be expressed in a linear functional form. In practice, with an abuse of language, regression is often described as if it is a data generating process. Therefore, the readers should discern between a model and an estimator when the regression is described in such a way.

Estimands of Regressions

Now we prove equation (2), (5) and (7). The first two results are proved together by showing that

$$\begin{aligned}\mathbb{E}[Y_i \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i] = \beta_0 & \\ + \sum_{l=1}^{n_L} M_{ikl} \frac{D_{ljk}}{g_{lk}(1-g_{lk})}\alpha_j & \\ + \sum_{l=2}^{n_L} L_{ibl}\left(\frac{f_{lj}-h_{ljk}}{1-g_{lk}} - \frac{f_{1j}-h_{1jk}}{1-g_{1k}}\right)\alpha_j & \\ + \sum_{l=2}^{n_L} P_{il}\delta_{lk} &\end{aligned} \tag{M9}$$

for some constant β_0, δ_{lk} and a vector variable $\mathbf{M}_{ik} = \{M_{ikl}\}_{l=1,\ldots,n_L}$.

Proof.

$$\begin{aligned}
\mathbb{E}[Y_i \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i] &= \mathbb{E}\left[\alpha_0 + \sum_{j=1}^{n_J} C_{ij}\alpha_j + \epsilon_i \;\middle|\; \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i\right] \\
&= \alpha_0 + \sum_{j=1}^{n_J} \mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i]\alpha_j \\
&= \alpha_0 + \sum_{j\in b} \mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i]\alpha_j + \sum_{j\notin b} \mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}, \mathbf{P}_i]\alpha_j \qquad \text{(M10)} \\
&= \alpha_0 + \sum_{j\in b} \mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}]\alpha_j + \sum_{j\notin b} \mathbb{E}[C_{ij} \mid \mathbf{P}_i]\alpha_j \\
&= \alpha_0 + \sum_{j\in b} \mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}]\alpha_j + \sum_{j\notin b} \left[\sum_{l=1}^{n_L} 2 f_{lj} P_{il}\right]\alpha_j
\end{aligned}$$

Now, it remains to compute $\mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}]$ for $j, k \in b$.

$$\begin{aligned}
\mathbb{E}[C_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}] &= 2\mathbb{E}[C^{\mathbf{h}}_{ij} \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}] \\
&= 2\mathbb{P}(C^{\mathbf{h}}_{ij} = 1 \mid \mathbf{M}_{ik}, \mathbf{L}_{ib}) \qquad \text{(M11)} \\
&= 2 \cdot \frac{\mathbb{P}(C^{\mathbf{h}}_{ij} = 1, \mathbf{M}_{ik} \mid \mathbf{L}_{ib})}{\mathbb{P}(\mathbf{M}_{ik} \mid \mathbf{L}_{ib})}
\end{aligned}$$

Since $\sum_{l=1}^{n_L} L_{ibl} = 2$, only two cases arise. It is either $L_{ibl} = 2$ for some l and $L_{ibl'} = 0$ for all $l' \neq l$, or $L_{ibl} = L_{ibl'} = 1$ for some l, l' and 0 otherwise. Therefore, it suffice to prove for the $n_L = 2$ case.

Case 1: $(L_{ib1}, L_{ib2}) = (2, 0)$**.** The possible combinations of (M_{ik1}, M_{ik2}) are $(0, 0)$, $(1, 0)$ and $(2, 0)$.

$$\begin{aligned}
2 \cdot \frac{\mathbb{P}(C^{\mathbf{h}}_{ij} = 1, \mathbf{M}_{ik} \mid \mathbf{L}_{ib})}{\mathbb{P}(\mathbf{M}_{ik} \mid \mathbf{L}_{ib})} &= 2 \cdot \frac{\mathbb{P}(C^{\mathbf{h}}_{ij} = 1, M_{ik1} \mid \mathbf{L}_{ib})}{\mathbb{P}(M_{ik1} \mid \mathbf{L}_{ib})} \\
&= \begin{cases} 2 \cdot \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} & : M_{ik1} = 0 \\ \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} + \frac{h_{1jk}}{g_{1k}} & : M_{ik1} = 1 \\ 2 \cdot \frac{h_{1jk}}{g_{1k}} & : M_{ik1} = 2 \end{cases} \qquad \text{(M12)} \\
&= \frac{D_{1jk}}{g_{1k}(1 - g_{1k})} M_{ik1} + 2 \cdot \frac{f_{1j} - h_{1jk}}{1 - g_{1k}}
\end{aligned}$$

Case 2: $(L_{ib1}, L_{ib2}) = (1,1)$**.** The possible combinations of (M_{ik1}, M_{ik2}) are $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$.

$$
\begin{aligned}
2 \cdot \frac{\mathbb{P}(C_{ij}^{\mathbf{h}} = 1, \mathbf{M}_{ik} \mid \mathbf{L}_{ib})}{\mathbb{P}(\mathbf{M}_{ik} \mid \mathbf{L}_{ib})} &= 2 \cdot \frac{\mathbb{P}(C_{ij}^{\mathbf{h}} = 1, M_{ik1} \mid \mathbf{L}_{ib})}{\mathbb{P}(M_{ik1} \mid \mathbf{L}_{ib})} \\
&= \begin{cases} \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} + \frac{f_{2j} - h_{2jk}}{1 - g_{2k}} & : (M_{ik1}, M_{ik2}) = (0,0) \\ \frac{h_{1jk}}{g_{1k}} + \frac{f_{2j} - h_{2jk}}{1 - g_{2k}} & : (M_{ik1}, M_{ik2}) = (1,0) \\ \frac{f_{1j} - h_{1jk}}{1 - g_{1k}} + \frac{h_{2jk}}{g_{2k}} & : (M_{ik1}, M_{ik2}) = (0,1) \\ \frac{h_{1jk}}{g_{1k}} + \frac{h_{2jk}}{g_{2k}} & : (M_{ik1}, M_{ik2}) = (1,1) \end{cases} \\
&= \sum_{l=1,2} \frac{D_{ljk}}{g_{lk}(1 - g_{lk})} M_{ikl} + \sum_{l=1,2} \frac{f_{lj} - h_{ljk}}{1 - g_{lk}} L_{ibl}
\end{aligned}
\tag{M13}
$$

Finally, substituting $L_{ib1} = 2 - L_{ib2}$ gives the desired result. □

Formula (7) is proved in Lee and Lee [15]. Here, we prove equation (8).

Proof. Note that

$$
\beta_k(\mathbf{P}_i) = \sum_{j=1}^{n_J} \frac{D_{jk}(\mathbf{P}_i)}{\mathrm{Var}[M_{ik} \mid \mathbf{P}_i]} \tag{M14}
$$

where $D_{jk}(\mathbf{P}_i) = \mathbb{P}(M_{ik}^{\mathbf{h}} = 1, C_{ij}^{\mathbf{h}} = 1 \mid \mathbf{P}_i) - \mathbb{P}(M_{ik}^{\mathbf{h}} = 1 \mid \mathbf{P}_i) \cdot \mathbb{P}(C_{ij}^{\mathbf{h}} = 1 \mid \mathbf{P}_i) = \mathrm{Cov}[M_{ik}^{\mathbf{h}}, C_{ij}^{\mathbf{h}} \mid \mathbf{P}_i]$ [15]. The numerator can be computed using the law of total covariance.

$$
\begin{aligned}
\mathrm{Cov}[M_{ik}^{\mathbf{h}}, C_{ij}^{\mathbf{h}} \mid \mathbf{P}_i] &= \mathbb{E}[\mathrm{Cov}(M_{ik}^{\mathbf{h}}, C_{ij}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}, \mathbf{P}_i) \mid \mathbf{P}_i] \\
&\quad + \mathrm{Cov}[\mathbb{E}(M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}, \mathbf{P}_i), \mathbb{E}(C_{ij}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}, \mathbf{P}_i) \mid \mathbf{P}_i] \\
&= \mathbb{E}[\mathrm{Cov}(M_{ik}^{\mathbf{h}}, C_{ij}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}) \mid \mathbf{P}_i] \\
&\quad + \mathrm{Cov}[\mathbb{E}(M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}), \mathbb{E}(C_{ij}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}) \mid \mathbf{P}_i] \\
&= \mathbb{E}\left[\sum_{l=1}^{n_L} D_{ljk} L_{ibl}^{\mathbf{h}} \,\middle|\, \mathbf{P}_i \right] + \mathrm{Cov}\left[\sum_{l=1}^{n_L} g_{lk} L_{ibl}^{\mathbf{h}}, \sum_{l'=1}^{n_L} f_{l'k} L_{ibl}^{\mathbf{h}} \,\middle|\, \mathbf{P}_i \right] \\
&= \sum_{l=1}^{n_L} D_{ljk} \mathbb{E}[L_{ibl}^{\mathbf{h}} \mid \mathbf{P}_i] + \sum_{l,l'} g_{lk} f_{l'j} \mathrm{Cov}[L_{ibl}^{\mathbf{h}}, L_{ibl'}^{\mathbf{h}} \mid \mathbf{P}_i]
\end{aligned}
\tag{M15}
$$

Since $\mathbf{L}_{ib}^{\mathbf{h}}$ follows a multinomial distribution with $n = 1$ and probability given by $\mathbf{P}_i$, $\mathbb{E}[L_{ibl}^{\mathbf{h}} \mid \mathbf{P}_i] = P_{il}$. Also, $\mathrm{Cov}[L_{ibl}^{\mathbf{h}}, L_{ibl'}^{\mathbf{h}} \mid \mathbf{P}_i]$ is $P_{il}(1 - P_{il})$ if $l = l'$ and $-P_{il}P_{il'}$ if $l \neq 1'$ respectively. Substituting these to equation (M15) gives the numerator as desired.

The denominator can be computed similarly using the law of total variance.

$$\begin{aligned}
\mathrm{Var}[M_{ik}^{\mathbf{h}} \mid \mathbf{P}_i] &= \mathbb{E}[\mathrm{Var}(M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}) \mid \mathbf{P}_i] + \mathrm{Var}[\mathbb{E}(M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}) \mid \mathbf{P}_i] \\
&= \mathbb{E}\left[\sum_{l=1}^{n_L} g_{lk}(1-g_{lk}) L_{ibl}^{\mathbf{h}} \,\middle|\, \mathbf{P}_i\right] + \mathrm{Var}\left[\sum_{l=1}^{n_L} g_{lk} L_{ibl}^{\mathbf{h}} \,\middle|\, \mathbf{P}_i\right] \\
&= \sum_{l=1}^{n_L} g_{lk}(1-g_{lk}) \mathbb{E}[L_{ibl}^{\mathbf{h}} \mid \mathbf{P}_i] + \sum_{l,l'} g_{lk} g_{l'k} \mathrm{Cov}[L_{ibl}^{\mathbf{h}}, L_{ibl'}^{\mathbf{h}} \mid \mathbf{P}_i]
\end{aligned} \tag{M16}$$

Using the covariance formula of the multinomial distribution as the numerator finishes the proof.

□

Exclusion of Local Ancestry and Interpretability

Proof. Obtaining the coefficient of M_{ikl} without L_{ibl} requires computing $\mathcal{P}[L_{ibl} \mid \mathbf{M}_{ik}, \mathbf{P}_i]$. Here, $\mathcal{P}$ is the linear projection [20]. The coefficient of $\mathbf{M}_{ikl}$ of this projection is the bias term. Following the Frisch-Waugh-Lovell (FWL) theorem, the coefficient is equal to the coefficient of $\mathcal{P}[L_{ibl} \mid \mathbf{M}_{ik} - \mathbb{E}(\mathbf{M}_{ik} \mid \mathbf{P}_i)]$ [20, 21]. By the definition of linear projection, this coefficient is

$$\mathrm{Var}[\mathbf{M}_{ik} - \mathbb{E}(\mathbf{M}_{ik} \mid \mathbf{P}_i)]^{-1} \mathrm{Cov}[\mathbf{M}_{ik} - \mathbb{E}(\mathbf{M}_{ik} \mid \mathbf{P}_i), L_{ibl}] \tag{M17}$$

This hold because $\mathbb{E} = \mathcal{P}$ when $\mathbb{E}$ is linear respect to the conditional variables.

We first compute $\mathrm{Var}[\mathbf{M}_{ik} - \mathbb{E}(M_{ik} \mid \mathbf{P}_i)]$. The l-th diagonal element is

$$\begin{aligned}
\mathrm{Var}[M_{ikl} - \mathbb{E}(M_{ikl} \mid \mathbf{P}_i)] &= \mathbb{E}[\mathrm{Var}(M_{ikl} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Var}(M_{ikl}^{\mathbf{h}} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Var}(\mathbb{E}[M_{ikl}^{\mathbf{h}} \mid \mathbf{L}_{ibl}^{\mathbf{h}}] \mid \mathbf{P}_i) + \mathbb{E}(\mathrm{Var}[M_{ikl}^{\mathbf{h}} \mid \mathbf{L}_{ibl}^{\mathbf{h}}] \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Var}(g_{lk} L_{ibl}^{\mathbf{h}} \mid \mathbf{P}_i) + \mathbb{E}(g_{lk}[1-g_{lk}] L_{bil}^{\mathbf{h}} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[g_{lk}^2 \mathrm{Var}(L_{ibl}^{\mathbf{h}} \mid \mathbf{P}_i) + g_{lk}(1-g_{lk}) \mathbb{E}(L_{bil}^{\mathbf{h}} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[g_{lk}^2 P_{il}(1-P_{il}) + g_{lk}(1-g_{lk}) P_{il}] \\
&= 2g_{lk}^2 \mathbb{E}[P_{il}(1-P_{il})] + 2g_{lk}(1-g_{lk}) \mathbb{E}[P_{il}]
\end{aligned} \tag{M18}$$

and the value at (l, l') is

$$\begin{aligned}
\mathrm{Cov}[M_{ikl} - \mathbb{E}(M_{ikl} \mid \mathbf{P}_i), M_{ikl'} - \mathbb{E}(M_{ikl'} \mid \mathbf{P}_i)] &= \mathbb{E}[\mathrm{Cov}(M_{ikl}, M_{ikl'} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Cov}(M^{\mathbf{h}}_{ikl}, M^{\mathbf{h}}_{ikl'} \mid \mathbf{P}_i)] \\
&= 2g_{lk}g_{l'k}\mathbb{E}[\mathrm{Cov}(L^{\mathbf{h}}_{ibl}, L^{\mathbf{h}}_{ibl'} \mid \mathbf{P}_i)] \\
&= 2g_{lk}g_{l'k}\mathbb{E}[-P_{il}P_{il'}] \\
&= -2g_{lk}g_{l'k}\mathbb{E}[P_{il}P_{il'}]
\end{aligned} \tag{M19}$$

due to $\mathrm{Cov}[M^{\mathbf{h}}_{ikl}, M^{\mathbf{h}}_{ikl'} \mid \mathbf{L}^{\mathbf{h}}_i] = 0$ if $l \neq l'$. The inverse matrix is complicated for general l but can be computed in the $l = 2$ case. The formula is provided at the end of the section.

Next, we compute $\mathrm{Cov}[\mathbf{M}_{ik} - \mathbb{E}(\mathbf{M}_{ik} \mid \mathbf{P}_i), L_{ibl}]$. The l'-th element is

$$\begin{aligned}
\mathrm{Cov}[M_{ikl'} - \mathbb{E}(M_{ikl'} \mid \mathbf{P}_i), L_{ibl}] &= \mathbb{E}[\mathrm{Cov}(M_{ikl'}, L_{ibl} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Cov}(M^{\mathbf{h}}_{ikl'}, L^{\mathbf{h}}_{ibl} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\mathrm{Cov}(g_{l'k}L^{\mathbf{h}}_{ibl'}, L^{\mathbf{h}}_{ibl} \mid \mathbf{P}_i)] \\
&= 2g_{l'k}\mathbb{E}[\mathrm{Cov}(L^{\mathbf{h}}_{ibl'}, L^{\mathbf{h}}_{ibl} \mid \mathbf{P}_i)] \\
&= \begin{cases} 2g_{l'k}\mathbb{E}[P_{il}(1 - P_{il})] & \text{if } l = l' \\ -2g_{l'k}\mathbb{E}[P_{il}P_{il'}] & \text{if } l \neq l' \end{cases}
\end{aligned} \tag{M20}$$

□

For general n_L, it is hard to derive the closed form solution of this result. Nevertheless, $n_L = 2$ case can be computed using the inverse formula of 2×2 matrices. Additionally, further simplification can be made by $P_{i1} + P_{i2} = 1$.

$$\begin{aligned}
&\begin{bmatrix} g^2_{1k}\mathbb{E}[P_{i1}(1 - P_{i1})] + g_{1k}(1 - g_{1k})\mathbb{E}[P_{i1}] & -g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] \\ -g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] & g^2_{2k}\mathbb{E}[P_{i2}(1 - P_{i2})] + g_{2k}(1 - g_{2k})\mathbb{E}[P_{i2}] \end{bmatrix}^{-1} \\
&= \frac{1}{\Delta_{12}} \begin{bmatrix} g^2_{2k}\mathbb{E}[P_{i2}(1 - P_{i2})] + g_{2k}(1 - g_{2k})\mathbb{E}[P_{i2}] & g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] \\ g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] & g^2_{1k}\mathbb{E}[P_{i1}(1 - P_{i1})] + g_{1k}(1 - g_{1k})\mathbb{E}[P_{i1}] \end{bmatrix}
\end{aligned} \tag{M21}$$

where

$$\begin{aligned}
\Delta_{12} = {} & g_{1k}(1 - g_{lk})g_{2k}\mathbb{E}[P_{i1}]\mathbb{E}[P_{i1}P_{i2}] \\
& + g_{1k}g_{2k}(1 - g_{2k})\mathbb{E}[P_{i1}P_{i2}]\mathbb{E}[P_{i2}] \\
& + g_{1k}(1 - g_{1k})g_{2k}(1 - g_{2k})\mathbb{E}[P_{i1}]\mathbb{E}[P_{i2}]
\end{aligned} \tag{M22}$$

Then we multiply the inverse matrix to the second term.

$$\begin{bmatrix} g_{2k}^2\mathbb{E}[P_{i2}(1-P_{i2})] + g_{2k}(1-g_{2k})\mathbb{E}[P_{i2}] & g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] \\ g_{1k}g_{2k}\mathbb{E}[P_{i1}P_{i2}] & g_{1k}^2\mathbb{E}[P_{i1}(1-P_{i1})] + g_{1k}(1-g_{1k})\mathbb{E}[P_{i1}] \end{bmatrix} \times \begin{bmatrix} -g_{1k}\mathbb{E}[P_{i1}P_{i2}] \\ g_{2k}\mathbb{E}[P_{i1}P_{i2}] \end{bmatrix} = \begin{bmatrix} -g_{1k}g_{2k}(1-g_{2k})\mathbb{E}[P_{i1}P_{i2}]\mathbb{E}[P_{i2}] \\ g_{1k}(1-g_{1k})g_{2k}\mathbb{E}[P_{i1}]\mathbb{E}[P_{i1}P_{i2}] \end{bmatrix} \tag{M23}$$

Therefore, gathering the terms gives

$$\beta_{1k}^{\text{nolanc}} = \beta_{1k}^{\text{anc}} + \sum_j \frac{g_{1k}g_{2k}(1-g_{2k})\mathbb{E}[P_{i1}P_{i2}]\mathbb{E}[P_{i2}]}{\Delta_{12}} \left(\frac{f_{1j}-h_{1jk}}{1-g_{1k}} - \frac{f_{2j}-h_{2jk}}{1-g_{2k}} \right) \alpha_j \tag{M24}$$

and

$$\beta_{2k}^{\text{nolanc}} = \beta_{2k}^{\text{anc}} + \sum_j \frac{g_{1k}(1-g_{1k})g_{2k}\mathbb{E}[P_{i1}]\mathbb{E}[P_{i1}P_{i2}]}{\Delta_{12}} \left(\frac{f_{2j}-h_{2jk}}{1-g_{2k}} - \frac{f_{1j}-h_{1jk}}{1-g_{1k}} \right) \alpha_j \tag{M25}$$

The Impact of Local Ancestry on Precision

Let $\widetilde{M}_{ik} = M_{ik} - \mathbb{E}[M_{ik} \mid \mathbf{P}_i]$ and $\widetilde{\mathbf{M}}_{ik} = \mathbf{M}_{ik} - \mathbb{E}[M_{ik} \mid \mathbf{L}_{ib}]$. We borrow the results of Ding (2021) and Buja et al. (2014) [21, 22].

$$\begin{aligned} \sqrt{n_I}\left(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right) &\xrightarrow[n_I\to\infty]{} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right]^{-1}\mathbb{E}\left[\zeta_{ik}^2\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right]\mathbb{E}\left[\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right]^{-1}\right) \\ \sqrt{n_I}\left(\widehat{\beta}_k^{\text{agg}} - \beta_k^{\text{agg}}\right) &\xrightarrow[n_I\to\infty]{} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\widetilde{M}_{ik}^2\right]^{-2}\mathbb{E}\left[\left(\zeta_{ik}^{\text{agg}}\right)^2\widetilde{M}_{ik}^2\right]\right) \end{aligned} \tag{M26}$$

where ζ_{ik} and ζ_{ik}^{agg} are the residuals appearing in regression (3) and (6) respectively. The expectations in the right hand side are difficult to compute. The calculations can be greatly eased by assuming that the variance explained by the causal variants linked to variant k is small relative to the total genetic variance [cite]. Furthermore, we must assume that ϵ_i has constant variance. Then,

$$\begin{aligned} \sqrt{n_I}\left(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right) &\xrightarrow[n_i\to\infty]{} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right]^{-1}\mathbb{E}\left[\eta_i^2\right]\right) \\ \sqrt{n_I}\left(\widehat{\beta}_k^{\text{agg}} - \beta_k^{\text{agg}}\right) &\xrightarrow[n_i\to\infty]{} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\widetilde{M}_{ik}^2\right]^{-1}\mathbb{E}\left[\eta_i^2\right]\right) \end{aligned} \tag{M27}$$

where $\eta_i = \epsilon_i + \sum_{j=1}^{n_J}\left[C_{ij} - \sum_{l=1}^{n_L} 2f_{lj}P_{il}\right]$.

Now, the expectations appearing the limit distributions can be expressed in terms of population genetic parameters. $\mathbb{E}\left[\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right] = \text{Var}[\mathbf{M}_{ik} - \mathbb{E}(\mathbf{M}_{ik} \mid \mathbf{P}_i)]$ so the diagonal elements are given by equation (M18) and the non-diagonal elements are given by equation (M19). The remaining terms are

$$\begin{aligned}
\mathbb{E}\left[\widetilde{M}_{ik}^2\right] &= \mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\text{Var}(M_{ik}^{\mathbf{h}} \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}[\text{Var}(\mathbb{E}[M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}] \mid \mathbf{P}_i) + \mathbb{E}(\text{Var}[M_{ik}^{\mathbf{h}} \mid \mathbf{L}_{ib}^{\mathbf{h}}] \mid \mathbf{P}_i)] \\
&= 2\mathbb{E}\left[\text{Var}\left(\sum_{l=1}^{n_L} g_{lk} L_{ibl}^{\mathbf{h}} \;\middle|\; \mathbf{P}_i\right) + \mathbb{E}\left(\sum_{l=1}^{n_L} g_{lk}(1-g_{lk}) L_{bil}^{\mathbf{h}} \;\middle|\; \mathbf{P}_i\right)\right] \\
&= 2\mathbb{E}\left[\sum_{l,l'} g_{lk}g_{l'k}\text{Cov}\left(L_{ibl}^{\mathbf{h}}, L_{ibl'}^{\mathbf{h}} \mid \mathbf{P}_i\right) + \sum_{l=1}^{n_L} g_{lk}(1-g_{lk})\mathbb{E}\left(L_{bil}^{\mathbf{h}} \mid \mathbf{P}_i\right)\right] \\
&= 2\mathbb{E}\left[\sum_{l=1}^{n_L} g_{lk}^2 P_{il}(1-P_{il}) + \sum_{l\neq l'} g_{lk}g_{l'k}(-P_{il}P_{il'}) + \sum_{l=1}^{n_L} g_{lk}(1-g_{lk})P_{il}\right] \\
&= 2\sum_{l=1}^{n_L} g_{lk}^2 \mathbb{E}[P_{il}(1-P_{il})] - 2\sum_{l\neq l'} g_{lk}g_{l'k}\mathbb{E}[P_{il}P_{il'}] + 2\sum_{l=1}^{n_L} g_{lk}(1-g_{lk})\mathbb{E}[P_{il}]
\end{aligned} \tag{M28}$$

and

$$\begin{aligned}
\mathbb{E}[\eta_i^2] &= \mathbb{E}[\epsilon_i^2] + \mathbb{E}\left[\left(\sum_{j=1}^{n_J}\left[C_{ij} - \sum_{l=1}^{n_L} 2f_{lj}P_{il}\right]\right)^2\right] \\
&= \mathbb{E}[\epsilon_i^2] + \mathbb{E}\left[\text{Var}\left(\sum_{j=1}^{n_J} C_{ij} \;\middle|\; \mathbf{P}_i\right)\right] \\
&= \mathbb{E}[\epsilon_i^2] + 2\mathbb{E}\left[\text{Var}\left(\sum_{j=1}^{n_J} C_{ij}^{\mathbf{h}} \;\middle|\; \mathbf{P}_i\right)\right] \\
&= \mathbb{E}[\epsilon_i^2] + 2\mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\sum_{j=1}^{n_J} C_{ij}^{\mathbf{h}} \;\middle|\; \mathbf{L}_{ib}^{\mathbf{h}}\right] \;\middle|\; \mathbf{P}_i\right) + \mathbb{E}\left(\text{Var}\left[\sum_{j=1}^{n_J} C_{ij}^{\mathbf{h}} \;\middle|\; \mathbf{L}_{ib}^{\mathbf{h}}\right] \;\middle|\; \mathbf{P}_i\right)\right] \\
&=
\end{aligned} \tag{M29}$$

We focus on the Wald test because the test, log-likelihood test (LRT) and F-test are all asymptotically equivalent at the true parameter [23]. Regression (3) tests $\boldsymbol{\beta}_k = \mathbf{0} \in \mathbb{R}^{n_L}$ and regression (6)

tests $\beta_k^{\text{agg}} = 0$. The corresponding test statistics asymptotically ($n_I \to \infty$) obeys

$$\begin{aligned} T_n &= \sqrt{n_I}\left(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)^T \left(\mathbb{E}\left[\widetilde{\mathbf{M}}_{ik}\widetilde{\mathbf{M}}_{ik}^T\right]^{-1}\mathbb{E}\left[\eta_i^2\right]\right)^{-1}\sqrt{n_I}\left(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right) \sim \boldsymbol{\chi}^2_{n_L} \\ T_n^{\text{agg}} &= \sqrt{n_I}\left(\widehat{\beta}_k^{\text{agg}} - \beta_k^{\text{agg}}\right)\left(\mathbb{E}\left[\widetilde{M}_{ik}^2\right]^{-1}\mathbb{E}\left[\eta_i^2\right]\right)^{-1}\sqrt{n_I}\left(\widehat{\beta}_k^{\text{agg}} - \beta_k^{\text{agg}}\right) \sim \boldsymbol{\chi}^2_1 \end{aligned} \tag{M30}$$

Note that we used the asymptotic variances themselves instead of the estimator. Also, we assumed that the inverted matrices are non-singular.

Combining Association and Admixture Signals

We prove equation (10) and (11).

Proof. When $M_{ik} = C_{ij}$ and variant j is the only causal variant (in block b), $f_{ljk} = f_{lj} = g_{lk}$ and $D_{ljk} = g_{lk} - g_{lk}g_{lk} = g_{lk}(1 - g_{lk})$. Substituting the above equation to equation (8) gives

$$\begin{aligned} \beta_k(\mathbf{P}_i) &= \frac{\sum_l D_{ljk}P_{il} + \sum_l g_{lk}f_{lj}P_{il}(1-P_{il}) - \sum_{l\neq l'} g_{lk}f_{l'k}P_{il}P_{il'}}{\sum_l g_{lk}(1-g_{lk})P_{il} + \sum_l g_{lk}^2 P_{il}(1-P_{il}) - \sum_{l\neq l'} g_{lk}g_{l'k}P_{il}P_{il'}}\alpha_j \\ &= \frac{\sum_l g_{lk}(1-g_{lk})P_{il} + \sum_l g_{lk}^2 P_{il}(1-P_{il}) - \sum_{l\neq l'} g_{lk}g_{l'k}P_{il}P_{il'}}{\sum_l g_{lk}(1-g_{lk})P_{il} + \sum_l g_{lk}^2 P_{il}(1-P_{il}) - \sum_{l\neq l'} g_{lk}g_{l'k}P_{il}P_{il'}}\alpha_j \\ &= \alpha_j \end{aligned} \tag{M31}$$

As a result,

$$\begin{aligned} \beta_k^{\text{agg}} &= \frac{\mathbb{E}[\beta_k(\mathbf{P}_i)\text{Var}(M_{ik} \mid \mathbf{P}_i)]}{\mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)]} \\ &= \frac{\mathbb{E}[\alpha_j \text{Var}(M_{ik} \mid \mathbf{P}_i)]}{\mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)]} \\ &= \alpha_j \frac{\mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)]}{\mathbb{E}[\text{Var}(M_{ik} \mid \mathbf{P}_i)]} \\ &= \alpha_j \end{aligned} \tag{M32}$$

Hence, we proved equation (10).

To prove equation (11), we consider the case when variant j is the only causal variant (in block b), and marker k is unlinked to j. Then $\beta_{lk} = 0$ for all l and $f_{lj} - h_{ljk} = f_{lj} - f_{lj}g_{lk} = f_{lj}(1 - g_{lk})$.

Substituting the latter equation to equation (5) gives

$$\begin{aligned}\gamma_{lk} &= \left(\frac{f_{lj}-h_{ljk}}{1-g_{lk}} - \frac{f_{1j}-h_{1jk}}{1-g_{1k}}\right)\alpha_j \\ &= \left[\frac{f_{lj}(1-g_{lk})}{1-g_{lk}} - \frac{f_{1j}(1-g_{1k})}{1-g_{1k}}\right]\alpha_j \\ &= (f_{lj}-f_{1j})\alpha_j\end{aligned} \tag{M33}$$

as desired. □

Based on the above two results, we formulate a Z-estimator combining the association and admixture [23]. We first define two composite random vectors

$$\mathbf{W}_i = (1, M_{ik}, \mathbf{P}_{i(-1)}) \text{ and } \mathbf{Z}_i = (1, \mathbf{N}_{ib(-1)}, \mathbf{P}_{i(-1)}) \tag{M34}$$

where $\mathbf{N}_{ib} = \{(f_{lj} - f_{1j})L_{ibl}\}_{l=1,\ldots,n_L}$ and the subscript (-1) stands for the omission of the first element. The corresponding coefficients are

$$\boldsymbol{\beta}_{\mathbf{W},(-1)} = (\beta_0^{\text{agg}}, \alpha_j, \boldsymbol{\delta}_{k(-1)}^{\text{agg}}) \text{ and } \boldsymbol{\beta}_{\mathbf{Z},(-1)} = (\beta_0, \underbrace{\alpha_j, \ldots, \alpha_j}_{(n_L-1)\text{-copies}}, \boldsymbol{\delta}_{k(-1)}) \tag{M35}$$

The moment function $\boldsymbol{\psi}$ is

$$\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Z}_i \,;\, \boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}) = \begin{bmatrix} \mathbf{W}_i \left(Y_i - \mathbf{W}_i^T \boldsymbol{\beta}_{\mathbf{W},(-1)}\right) \\ \mathbf{Z}_i \left(Y_i - \mathbf{Z}_i^T \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right) \end{bmatrix} \in \mathbb{R}^{3n_L \times 1} \tag{M36}$$

which obeys $\mathbb{E}\left[\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Z}_i \,;\, \boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)})\right] = \mathbf{0}$. As a result, we have total n_L-constraints on α_j. Note that considering only the first $(n_L + 1)$ components $\mathbf{W}_i^T \left(Y_i - \boldsymbol{\beta}_{\mathbf{W},(-1)}\mathbf{W}_i\right)$ are identical to OLS. Augmenting additional $(2n_L - 1)$ constraints is at least as asymptotically efficient as the OLS [24, 25, 26]. The latter constraints only is simply the admixture mapping.

The estimators of $\boldsymbol{\beta}_{\mathbf{W},(-1)}$ and $\boldsymbol{\beta}_{\mathbf{Z},(-1)}$ of defined as

$$\begin{aligned}&\left(\boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right) \\ &= \underset{\boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}}{\text{argmin}} \left[\sum_{i=1}^{n_I} \boldsymbol{\psi}_i \left(\boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right)\right]^T \boldsymbol{\Omega} \left[\sum_{i=1}^{n_I} \boldsymbol{\psi}_i \left(\boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right)\right]\end{aligned} \tag{M37}$$

for some positive definite weighting matrix $\boldsymbol{\Omega} \in \mathbb{R}^{3n_L \times 3n_L}$, and due to space constraints, we write $\boldsymbol{\psi}_i \left(\boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right) = \boldsymbol{\psi}\left(\mathbf{W}_i, \mathbf{Z}_i \,;\, \boldsymbol{\beta}_{\mathbf{W},(-1)}, \boldsymbol{\beta}_{\mathbf{Z},(-1)}\right)$.

References

[1] Giovanni Montana and Jonathan K. Pritchard. Statistical tests for admixture mapping with case-control and cases-only data. *The American Journal of Human Genetics*, 75(5):771–789, nov 2004.

[2] Michael W. Smith and Stephen J. O'Brien. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet*, 6(8):623–632, jul 2005.

[3] Ching-Yu Cheng, W. H. Linda Kao, Nick Patterson, Arti Tandon, Christopher A. Haiman, Tamara B. Harris, Chao Xing, Esther M. John, Christine B. Ambrosone, Frederick L. Brancati, Josef Coresh, Michael F. Press, Rulan S. Parekh, Michael J. Klag, Lucy A. Meoni, Wen-Chi Hsueh, Laura Fejerman, Ludmila Pawlikowska, Matthew L. Freedman, Lina H. Jandorf, Elisa V. Bandera, Gregory L. Ciupak, Michael A. Nalls, Ermeg L. Akylbekova, Eric S. Orwoll, Tennille S. Leak, Iva Miljkovic, Rongling Li, Giske Ursin, Leslie Bernstein, Kristin Ardlie, Herman A. Taylor, Eric Boerwinckle, Joseph M. Zmuda, Brian E. Henderson, James G. Wilson, and David Reich. Admixture mapping of 15,280 african americans identifies obesity susceptibility loci on chromosomes 5 and x. *PLoS Genet Genetics*, 5(5):e1000490, may 2009.

[4] Matthew L. Freedman, Christopher A. Haiman, Nick Patterson, Gavin J. McDonald, Arti Tandon, Alicja Waliszewska, Kathryn Penney, Robert G. Steen, Kristin Ardlie, Esther M. John, Ingrid Oakley-Girvan, Alice S. Whittemore, Kathleen A. Cooney, Sue A. Ingles, David Altshuler, Brian E. Henderson, and David Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in african-american men. *Proc. Natl. Acad. Sci. U.S.A.*, 103(38):14068–14073, sep 2006.

[5] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, jul 2017.

[6] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, mar 2019.

[7] Roseann E. Peterson, Karoline Kuchenbaecker, Raymond K. Walters, Chia-Yen Chen, Alice B. Popejoy, Sathish Periyasamy, Max Lam, Conrad Iyegbe, Rona J. Strawbridge, Leslie Brick, Caitlin E. Carey, Alicia R. Martin, Jacquelyn L. Meyers, Jinni Su, Junfang Chen, Alexis C. Edwards, Allan Kalungi, Nastassja Koen, Lerato Majara, Emanuel Schwarz, Jordan W. Smoller, Eli A. Stahl, Patrick F. Sullivan, Evangelos Vassos, Bryan Mowry, Miguel L. Prieto, Alfredo Cuellar-Barboza, Tim B. Bigdeli, Howard J. Edenberg, Hailiang Huang, and Laramie E. Dun-

can. Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*, 179(3):589–603, oct 2019.

[8] Bogdan Pasaniuc, Noah Zaitlen, Guillaume Lettre, Gary K. Chen, Arti Tandon, W. H. Linda Kao, Ingo Ruczinski, Myriam Fornage, David S. Siscovick, Xiaofeng Zhu, Emma Larkin, Leslie A. Lange, L. Adrienne Cupples, Qiong Yang, Ermeg L. Akylbekova, Solomon K. Musani, Jasmin Divers, Joe Mychaleckyj, Mingyao Li, George J. Papanicolaou, Robert C. Millikan, Christine B. Ambrosone, Esther M. John, Leslie Bernstein, Wei Zheng, Jennifer J. Hu, Regina G. Ziegler, Sarah J. Nyante, Elisa V. Bandera, Sue A. Ingles, Michael F. Press, Stephen J. Chanock, Sandra L. Deming, Jorge L. Rodriguez-Gil, Cameron D. Palmer, Sarah Buxbaum, Lynette Ekunwe, Joel N. Hirschhorn, Brian E. Henderson, Simon Myers, Christopher A. Haiman, David Reich, Nick Patterson, James G. Wilson, and Alkes L. Price. Enhanced statistical tests for GWAS in admixed populations: Assessment using african americans from CARe and a breast cancer consortium. *PLoS Genet Genetics*, 7(4):e1001371, apr 2011.

[9] Elizabeth G. Atkinson, Adam X. Maihofer, Masahiro Kanai, Alicia R. Martin, Konrad J. Karczewski, Marcos L. Santoro, Jacob C. Ulirsch, Yoichiro Kamatani, Yukinori Okada, Hilary K. Finucane, Karestan C. Koenen, Caroline M. Nievergelt, Mark J. Daly, and Benjamin M. Neale. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet*, 53(2):195–204, jan 2021.

[10] Kangcheng Hou, Arjun Bhattacharya, Rachel Mester, Kathryn S. Burch, and Bogdan Pasaniuc. On powerful GWAS in admixed populations. *Nat Genet*, 53(12):1631–1633, nov 2021.

[11] Elizabeth G. Atkinson, Alex Bloemendal, Adam X. Maihofer, Caroline M. Nievergelt, Mark J. Daly, and Benjamin M. Neale. Reply to: On powerful GWAS in admixed populations. *Nat Genet*, 53(12):1634–1635, nov 2021.

[12] Kangcheng Hou, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian Belbin, David Conti, Burcu F. Darst, Myriam Fornage, Chris Gignoux, Xiuqing Guo, Christopher Haiman, Eimear Kenny, Michelle Kim, Charles Kooperberg, Leslie Lange, Ani Manichaikul, Kari E. North, Natalie Nudelman, Ulrike Peters, Laura J. Rasmussen-Torvik, Stephen S. Rich, Jerome I. Rotter, Heather E. Wheeler, Ying Zhou, Sriram Sankararaman, and Bogdan Pasaniuc. Causal effects on complex traits are similar across segments of different continental ancestries within admixed individuals. *bioRxiv*, aug 2022.

[13] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, jun 2000.

[14] Noah A Rosenberg and Magnus Nordborg. A general population-genetic model for the production by population structure of spurious genotype–phenotype associations in discrete, admixed or spatially distributed populations. *Genetics*, 173(3):1665–1678, jul 2006.

[15] Hanbin Lee and Moo Hyuk Lee. Disentangling linkage and population structure in association mapping. *Unpublished work*, 2023.

[16] James F Crow and Motoo Kimura. *An introduction to population genetics theory*. Blackburn Press, West Caldwell, NJ, January 2009.

[17] Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Helga Ask, Aris Baras, Christopher R. Bauer, Dorret I. Boomsma, Archie Campbell, Harry Campbell, Zhengming Chen, Paraskevi Christofidou, Elizabeth Corfield, Christina C. Dahm, Deepika R. Dokuru, Luke M. Evans, Eco J. C. de Geus, Sudheer Giddaluru, Scott D. Gordon, K. Paige Harden, W. David Hill, Amanda Hughes, Shona M. Kerr, Yongkang Kim, Hyeokmoon Kweon, Antti Latvala, Deborah A. Lawlor, Liming Li, Kuang Lin, Per Magnus, Patrik K. E. Magnusson, Travis T. Mallard, Pekka Martikainen, Melinda C. Mills, Pål Rasmus Njølstad, John D. Overton, Nancy L. Pedersen, David J. Porteous, Jeffrey Reid, Karri Silventoinen, Melissa C. Southey, Camilla Stoltenberg, Elliot M. Tucker-Drob, Margaret J. Wright, Hyeokmoon Kweon, Philipp D. Koellinger, Daniel J. Benjamin, Patrick Turley, Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Helga Ask, Dorret I. Boomsma, Archie Campbell, Harry Campbell, Zhengming Chen, Paraskevi Christofidou, Elizabeth Corfield, Christina C. Dahm, Deepika R. Dokuru, Luke M. Evans, Eco J. C. de Geus, Sudheer Giddaluru, Scott D. Gordon, K. Paige Harden, W. David Hill, Amanda Hughes, Shona M. Kerr, Yongkang Kim, Antti Latvala, Deborah A. Lawlor, Liming Li, Kuang Lin, Per Magnus, Patrik K. E. Magnusson, Travis T. Mallard, Pekka Martikainen, Melinda C. Mills, Pål Rasmus Njølstad, Nancy L. Pedersen, David J. Porteous, Karri Silventoinen, Melissa C. Southey, Camilla Stoltenberg, Elliot M. Tucker-Drob, Margaret J. Wright, John K. Hewitt, Matthew C. Keller, Michael C. Stallings, James J. Lee, Kaare Christensen, Sharon L. R. Kardia,

Patricia A. Peyser, Jennifer A. Smith, James F. Wilson, John L. Hopper, Sara Hägg, Tim D. Spector, Jean-Baptiste Pingault, Robert Plomin, Alexandra Havdahl, Meike Bartels, Nicholas G. Martin, Sven Oskarsson, Anne E. Justice, Iona Y. Millwood, Kristian Hveem, Øyvind Naess, Cristen J. Willer, Bjørn Olav Åsvold, Jaakko Kaprio, Sarah E. Medland, Robin G. Walters, David M. Evans, George Davey Smith, Caroline Hayward, Ben Brumpton, Gibran Hemani, Neil M. Davies, John K. Hewitt, Matthew C. Keller, Michael C. Stallings, James J. Lee, Kaare Christensen, Sharon L. R. Kardia, Patricia A. Peyser, Jennifer A. Smith, James F. Wilson, John L. Hopper, Sara Hägg, Tim D. Spector, Jean-Baptiste Pingault, Robert Plomin, Alexandra Havdahl, Meike Bartels, Nicholas G. Martin, Sven Oskarsson, Anne E. Justice, Iona Y. Millwood, Kristian Hveem, Øyvind Naess, Cristen J. Willer, Bjørn Olav Åsvold, Philipp D. Koellinger, Jaakko Kaprio, Sarah E. Medland, Robin G. Walters, Daniel J. Benjamin, Patrick Turley, David M. Evans, George Davey Smith, Caroline Hayward, Ben Brumpton, Gibran Hemani, Neil M. Davies, and and. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat Genet*, 54(5):581–592, may 2022.

[18] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, mar 2009.

[19] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, dec 2021.

[20] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. The MIT Press. MIT Press, London, England, 2 edition, October 2010.

[21] Peng Ding. The frisch–waugh–lovell theorem for standard errors. *Statistics & Probability Letters*, 168:108945, jan 2021.

[22] Andreas Buja, Richard Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang. Models as approximations i: Consequences illustrated with linear regression, 2014.

[23] A W van der Vaart. *Cambridge series in statistical and probabilistic mathematics: Asymptotic statistics series number 3*. Cambridge University Press, Cambridge, England, June 2000.

[24] Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269, sep 1982.

[25] Roger Koenker and José A.F. Machado. GMM inference when the number of moment conditions is large. *Journal of Econometrics*, 93(2):327–344, December 1999.

[26] Chirok Han and Peter C. B. Phillips. GMM with many moment conditions. *Econometrica*, 74(1):147–192, jan 2006.