Biometrics 000, 000–000 DOI: 000
000 0000

A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis

Jingchunzi Shi[1,*] and Seunggeun Lee[1,]**

[1]Department of Biostatistics, University of Michigan, Ann Arbor, U.S.A.

email: shijingc@umich.edu

**email:* leeshawn@umich.edu

Summary: Meta-analysis of trans-ethnic genome-wide association studies (GWAS) has proven to be a practical and profitable approach for identifying loci that contribute to the risk of complex diseases. However, the expected genetic effect heterogeneity cannot easily be accommodated through existing fixed-effects and random-effects methods. In response, we propose a novel random effect model for trans-ethnic meta-analysis with flexible modeling of the expected genetic effect heterogeneity across diverse populations. Specifically, we adopt a modified random effect model from the kernel regression framework, in which genetic effect coefficients are random variables whose correlation structure reflects the genetic distances across ancestry groups. In addition, we use the adaptive variance component test to achieve robust power regardless of the degree of genetic effect heterogeneity. Simulation studies show that our proposed method has well-calibrated type I error rates at very stringent significance levels and can improve power over the traditional meta-analysis methods. We re-analyzed the published type 2 diabetes GWAS meta-analysis (Consortium et al., 2014) and successfully identified one additional SNP that clearly exhibits genetic effect heterogeneity across different ancestry groups. Furthermore, our proposed method provides scalable computing time for genome-wide datasets, in which an analysis of one million SNPs would require less than 3 hours.

Key words: Effect-size heterogeneity; GWAS; Kernel regression; Meta-analysis; Random effect model; Trans-ethnic meta-analysis.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Although genome-wide association studies (GWAS) have successfully identified more than 2000 loci that influence the severity of human health outcomes, those identified loci account for only a small fraction of the genetic contribution to complex diseases and traits (McCarthy et al., 2008). It has been argued that numerous loci with very small effects can explain additional disease risk or trait heritability, and the challenge is to find those loci that can be identified only with very large numbers of samples (Eichler et al., 2010). Since it is too costly to design and conduct a single study with tens or hundreds of thousands of samples, a more practical alternative is to combine studies that have already been conducted through a meta-analysis (Evangelou and Ioannidis, 2013).

A natural extension of the traditional European-based meta-analysis is to include samples from as many studies as possible, even if they come from heterogeneous ancestries. With the increased sample size, trans-ethnic meta-analysis is expected to be more powerful at detecting novel loci without the cost of additional genotyping (Cooper et al., 2008). In fact, several trans-ethnic meta-analyses have been performed in the last few years with success in discovering risk alleles across ancestry groups. For example, five consortia (Consortium et al., 2014) aggregated published GWAS meta-analyses of type 2 diabetes (T2D) from four ancestry groups and successfully identified seven new loci with very small effect sizes.

To take full advantage of the profitability of trans-ethnic meta-analysis, improved statistical methods are required to account for the distinctive ancestral origins among data. Existing methods for GWAS meta-analysis include the classical fixed-effects and random-effects methods, as well as the recently introduced new random-effects method by Han and Eskin (2011) and the Bayesian approach by Morris (2011). The fixed-effects method (FE) (Evangelou and Ioannidis, 2013) is the most popular approach for synthesizing GWAS data. It assumes that the true effect of each risk allele is the same in each data set, and as a result, it

has limited power in the presence of genetic effect heterogeneity (Evangelou and Ioannidis, 2013). The random-effects method (RE) was developed explicitly to model the between-study heterogeneity; however, it implicitly assumes heterogeneity under the null hypothesis, which causes it to have far more limited power than FE (Han and Eskin, 2011). To relax the conservative assumption of RE, Han and Eskin (2011) developed a new random-effects model (RE-HE) which achieves higher power than RE. Morris (2011) developed a trans-ethnic meta-analysis method by means of a Bayesian partition model (MANTRA). MANTRA accounts for the relatedness of studies by grouping them into different ethnic clusters. Specifically, studies that are grouped into the same ethnic cluster share the same underlying genetic effect, while different ethnic clusters have different underlying genetic effects.

The aforementioned T2D trans-ethnic meta-analysis was carried out using the FE method. In addition to identifying novel T2D susceptibility loci, they analyzed 69 established T2D susceptibility loci using Cochran Q test (Cochran, 1954) to evaluate their genetic effect heterogeneity. Among the 69 loci, 3 had very strong evidence of the heterogeneity (Cochran Q p-value $< 10^{-3}$), and 12 had some evidence of the heterogeneity ($10^{-3} \leqslant$ Cochran Q p-value < 0.05). For those 15 loci, FE may not be sufficiently powerful to detect the association signals. In order to improve power, we developed a new trans-ethnic meta-analysis approach, referred to as TransMeta, and used it to reanalyze the T2D trans-ethnic meta-data.

As mentioned above, one of the challenging issues in trans-ethnic GWAS meta-analysis is to appropriately account for the expected genetic effect heterogeneity. There can be several reasons for the heterogeneous effect sizes. First, it is highly possible that the queried SNP is not the underlying causal SNP, but rather is correlated to the causal SNP through linkage disequilibrium (LD). Variations in the LD structures across ancestry groups can create the genetic effect heterogeneity. Second, the environmental risk factors may differ between ancestry groups. With the possibility of interaction between the causal variants and

the environmental factors, marginal genetic effects may vary between populations (Morris, 2011). To address the heterogeneity issue, we consider a modified random effect model based on a kernel machine framework (Liu et al., 2007). Specifically, we treat the genetic effect coefficients as random variables, with their correlation structure across ancestry groups reflecting the expected heterogeneity (or homogeneity) among ancestry groups. To test for associations, we derive a data-adaptive variance component test with adaptive selection of the degree of heterogeneity. This adaptive test combines models of homogeneous and heterogeneous genetic effects, and provides robust power regardless of the genetic effect distribution. For details of our proposed method, TransMeta, see Section 2, Methods.

The rest of this paper is organized as follows: In Section 3, to compare the performance of TransMeta with FE, RE, RE-HE and MANTRA for meta-analyzing GWAS across genetically diverse populations, we perform simulation studies and reanalyze the T2D trans-ethnic meta-analysis. We conclude the paper with a discussion in Section 4.

2. Methods

2.1 *Statistical Models for GWAS Meta-Analysis*

In this section, we first introduce statistical models of the existing GWAS meta-analysis methods. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_n)'$ be the effect-size estimates, such as the log odds ratios or regression coefficients, in n independent studies. If the sample sizes in each study are sufficiently large, then

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}, \Sigma), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$, with β_i being the true effect size in the i_{th} study; and $\Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2)$, with σ_i^2 being the variance of $\widehat{\beta}_i$.

FE assumes that all the studies share a common effect-size μ (i.e. $\beta_1 = \cdots = \beta_n = \mu$). FE is powerful at detecting genetic effects that are present in most, if not all, of the studies

with homogeneous effect sizes. The RE model assumes that the true effect size β_i for the i_{th} study is generated from a normal distribution with mean μ and variance τ_1,

$$\beta_i = \mu + \eta_i, \quad \eta_i \sim N(0, \tau_1). \tag{2}$$

RE typically assumes that even under the null hypothesis of no association, β_is can be different across studies, since τ_1 is not assumed to be zero under the null hypothesis. Due to this conservative assumption, RE has far more limited power at detecting association signals than FE. Han and Eskin (2011) developed a new RE approach (RE-HE) that assumes no genetic effect heterogeneity under the null hypothesis. Specifically, they assumed that β_is are zero among all the studies under the null hypothesis (i.e. $\mu = 0$ and $\tau_1 = 0$), and they allowed varying effect sizes among studies under the alternative hypothesis. The likelihood ratio test was used to evaluate the null hypothesis of $\mu = 0$ and $\tau_1 = 0$. Since asymptotic p-values of RE-HE are only accurate when the number of studies (n) is very large, they provide tabulated p-values precomputed with an assumption of equal sample sizes across studies. In the presence of inter-study effect-size heterogeneity, RE-HE yields higher power than FE.

The aforementioned three frequentist meta-analysis methods can all be summarized under model (2) with certain assumptions on τ_1. With $\tau_1 = 0$ under both the null and the alternative hypotheses, model (2) is exactly the same as FE. RE assumes that τ_1 is non-zero under both the null and the alternative hypotheses, and tests whether $\mu = 0$ or not, while accounting for the between-study variance τ_1. RE-HE assumes that $\tau_1 = 0$ under the null hypothesis, and tests whether both μ and τ_1 are zero.

Unlike the frequentist approaches, the Bayesian meta-analysis approach, MANTRA, assigns studies into ethnic clusters under model (1). It assumes that studies that are grouped into the same ethnic cluster share the same underlying genetic effect. If we fix the number of clusters as one, all the studies are grouped into one ethnic cluster with homogeneous

genetic effects; in this case, MANTRA can be viewed as a Bayesian implementation of the fixed-effects method. If the number of cluster is fixed to be the same as the number of studies (n), each study is assigned to be its own cluster; in this case, MANTRA can be viewed as a Bayesian implementation of the random-effects method. MANTRA uses the Bayesian partition model to adaptively determine the number of ethnic clusters and the cluster membership and assesses the association evidence by means of the Bayes factor.

2.2 *New Model Framework for GWAS Meta-Analysis*

The existing frequentist meta-analysis methods based on (2) are not optimal when the effect sizes exhibit certain structures across studies. In multi-ethnic meta-analysis, for example, the studies can be grouped by their ethnicities. Genetically similar groups may have more homogeneous genetic effects compared to genetically diverse groups. In response, we propose a statistical framework that can accommodate prior assumptions on genetic effect distributions. Specifically, we adopt the kernel machine framework (Liu et al., 2007) to flexibly model the genetic effect distributions. Instead of assuming η_is are i.i.d normal samples, we assume that η_is jointly follow a mean zero Gaussian process with kernel function $\tau_1 K(\cdot,\cdot)$, where $K(\cdot,\cdot)$ is a bivariate function to represent genetic similarity between two groups. This kernel regression framework has been successfully applied in many areas of genetic studies, including rare variant association analysis (Wu et al., 2011) and pathway analysis (Liu et al., 2007). In Section 2.3, we will discuss choices of kernels for trans-ethnic meta-analysis.

We first propose to extend (2) to a hierarchical model by modeling μ as a random variable with distribution $N(0,\tau_2)$. From this extension, our proposed model framework can be summarized as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}|\boldsymbol{\beta} &\sim MVN(\boldsymbol{\beta},\Sigma) \\ \boldsymbol{\beta}|\tau_1,\tau_2 &\sim MVN(\mathbf{0},\tau_1\mathbf{K}+\tau_2\mathbf{1}\mathbf{1}^{'}), \end{aligned} \tag{3}$$

where $\mathbf{K}$ is an $n \times n$ kernel matrix and $\mathbf{1} = (1, \cdots, 1)'$. We then apply a re-parameterization $\tau_1 = \tau(1-\rho)$ and $\tau_2 = \tau\rho$, where ρ reflects whether genetic effects are homogeneous ($\rho = 1$) or heterogeneous ($\rho = 0$) across ancestry groups, and τ represents the size of the regression coefficients $\boldsymbol{\beta}$. From this re-parameterization, testing for both μ and τ_1 being zero becomes testing for the common variance component τ being zero. Our final model framework is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}|\boldsymbol{\beta} &\sim MVN(\boldsymbol{\beta}, \Sigma) \\ \boldsymbol{\beta}|\tau &\sim MVN(\mathbf{0}, \tau V_\rho) \\ V_\rho &= (1-\rho)\mathbf{K} + \rho\mathbf{1}\mathbf{1}', \quad 0 \leqslant \rho \leqslant 1 \end{aligned} \tag{4}$$

where V_ρ is an $n \times n$ (scaled) covariance matrix of $\boldsymbol{\beta}$. We note that V_ρ is a linear combination of two matrices, $\mathbf{1}\mathbf{1}'$ and $\mathbf{K}$, with coefficient ρ that determines the degree of heterogeneity. $\rho = 0$ indicates that the covariance structure of β_is is the same as the kernel matrix $\mathbf{K}$, and $\rho = 1$ indicates that β_is are perfectly correlated (and hence homogeneous).

Our proposed model includes the three frequentist meta-analysis approaches as special cases. For example, if $\rho = 1$ (i.e $V_\rho = \mathbf{1}\mathbf{1}'$), the model is effectively the same as FE since all β_is should be the same under the alternative hypothesis. We show in Section 2.4 that the variance component score test for $\tau = 0$ with $\rho = 1$ is exactly the same as the inverse-variance weighted meta-analysis test, the most popular test for the FE approach. As a result, one of the important features of our model is that it includes FE regardless of the choice of $\mathbf{K}$. We believe this is a desirable feature since numerous disease-associated SNPs in various meta-analysis scenarios including trans-ethnic meta-analysis exhibit homogeneous genetic effects across studies (Marigorta and Navarro, 2013). RE and RE-HE are equivalent to testing for $\tau_2 = 0$ and $\tau_1 = \tau_2 = 0$ under (3), respectively, with $\mathbf{K} = \mathbf{I}$. This indicates that RE is equivalent to testing for $\rho = 0$, and RE-HE is equivalent to testing for $\tau = 0$ while adaptively selecting ρ under the re-parameterized model (4) with $\mathbf{K} = \mathbf{I}$.

2.3 *Choice of the Kernel Matrix* $\mathbf{K}$ *for Trans-Ethnic Meta-Analysis*

Suppose the GWAS meta-analysis has p ancestry groups, and the t_{th} group is denoted by t $(t = 1, \ldots, p)$. Using those notations, we propose two choices for the kernel structure $\mathbf{K}$:

Choice 1. Group-wise independent kernel structure:

We consider a simple assumption in which genetic effect sizes are independently distributed across ancestry groups, but homogeneous within the same ancestry group. In particular, $K_{ij} = 1$ if and only if study i and j belong to the same ancestry group t for some $t \in \{1, \ldots, p\}$. In Web Appendix Section 1, we provide the general form of $\mathbf{K}$ under this group-wise independent structure.

Choice 2. Genetic similarity (F_{st}) kernel structure:

The fixation index (F_{st}) is a widely used measure of population differentiation due to genetic structure (Wright, 1949). $F_{st} = 0$ indicates there is no allele frequency differentiation between populations, whereas a large value of F_{st} indicates that populations are genetically very different. F_{st} has been used as a genetic distance among populations. For example, MANTRA uses F_{st} to group studies to ethnic clusters. For each cluster, it is assumed that studies share the same genetic effect. We adopt the strategy of using F_{st} in constructing the kernel matrix $\mathbf{K}$ to incorporate genetic similarity into genetic effect similarity. In particular,

$$K_{ij} = 1 - \frac{F_{st_{tt'}}}{D}, \quad \text{with } D = \max_{t,t' \in \{1,\ldots,p\}} \{F_{st_{tt'}}\},$$

where study i and j belong to ancestry group t and t' respectively, and $F_{st_{tt'}}$ is the pairwise F_{st} between the corresponding ancestry groups. In Web Appendix Section 1, we provide the general form of $\mathbf{K}$ under this genetic similarity (F_{st}) kernel structure. Unlike MANTRA, which adaptively groups studies based on the prior model of relatedness and observed effect sizes via the Bayesian partition model, our method constructs the genetic similarity (F_{st}) kernel using only the genotype data and fixes it prior to carrying out the data analysis.

2.4 *Hypothesis Test*

Under the proposed model (4), testing for $H_0 : \beta_1 = \cdots = \beta_n = 0$ is the same as testing for the variance component $\tau = 0$ (i.e. $H_0 : \tau = 0$). We first consider a situation in which ρ is given before carrying out the test. Following Zhang and Lin (2003), the score test statistics of the variance component τ with a given ρ is

$$S_\rho = \widehat{\boldsymbol{\beta}}^{'}\widehat{\Sigma}^{-1}V_\rho\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{'}\widehat{\Sigma}^{-1}[(1-\rho)\mathbf{K} + \rho\mathbf{1}\mathbf{1}^{'}]\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\beta}}, \tag{5}$$

where $\widehat{\Sigma} = diag(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_n^2)$, and $\widehat{\sigma}_i^2$ is an estimate of σ_i^2. When $\rho = 1$, the test statistic S_ρ becomes $\left(\sum_{i=1}^n \widehat{\beta}_i/\widehat{\sigma}_i^2\right)^2$, which is the test statistics of the inverse variance weighting.

For any given ρ, S_ρ asymptotically follows a mixture of χ^2 distributions under the null hypothesis. Specifically, if $(\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of $\widehat{\Sigma}^{-1/2}V_\rho\widehat{\Sigma}^{-1/2}$, the null distribution of S_ρ can be closely approximated by

$$\sum_{j=1}^{n} \lambda_j \chi_{1,j}^2, \tag{6}$$

where $\{\chi_{1,j}^2\}$ are independent χ_1^2 random variables. Several methods exist to obtain tail probabilities of the mixture of χ^2 distributions. Among them, the method to invert a characteristic function (Davies, 1980) provides very accurate estimates of tail probabilities and is widely used in many recently developed genetic association tests (Wu et al., 2011). We use this method when ρ is given.

In practice, however, we rarely have prior information on which ρ is optimal in terms of power. Lee et al. (2012) have studied a similar problem within a context of rare variant association analysis; they proposed to use the minimum p-value over a grid of ρ as a test statistics. We adopt the same approach here. Specifically, the test statistic is $T = \inf_{0 \leqslant \rho \leqslant 1} p_\rho$, where p_ρ is the p-value based on S_ρ. T can be obtained by a simple grid search across a range of ρ: set a grid $0 \leqslant \rho_1 \leqslant \rho_2 \leqslant \ldots \leqslant \rho_b \leqslant 1$, then the test statistic becomes

$$T = \min\{p_{\rho_1}, \ldots, p_{\rho_b}\},$$

and the optimal ρ is set as the one whose corresponding p-value (p_ρ) equals to T. We observed that a dense grid of ρ does not necessarily improve power. Therefore, we suggest using $\rho = (0, 0.3^2, 0.5^2, 1)$ for simulations and real data analysis. Once the test statistic T is calculated, the next step is to obtain the corresponding p-value for assessing the association evidence. If we had just used the minimum p-value (which is denoted as our test statistic T) to assess significance, we would ignore the multiple comparisons between different p_ρ values, which would result in inflated type I error control. Thus, we derived the asymptotic distribution of T to obtain its p-value, details provided in Web Appendix Section 2.

2.5 *Using Z-scores instead of Effect-size Estimates*

In previous sections, we constructed our methods based on estimates of effect sizes and their standard errors. However, Z-score based approaches are also very popular in GWAS. Z-score based approaches use p-values (p_i), sample sizes (n_i) and direction of effects (Δ_i) to construct Z-scores for each study, and then calculate a weighted sum of Z-scores to carry out meta-analysis. A major advantage of the Z-score based approach is that it allows meta-analysis of data when effect size estimates are not available or measurements of traits are difficult to standardize, ex. tobacco or alcohol use (Evangelou and Ioannidis, 2013). In this section, we extend our method to Z-score based approaches.

Based on input summary statistics (p_i, n_i, Δ_i), a signed Z-score is constructed as $Z_i = \Phi^{-1}(1 - p_i/2) * sign(\Delta_i)$ for each study, where $\Phi(\cdot)$ is the standard normal distribution function. For continuous traits, the effect size estimate $\widehat{\beta}_i$ is asymptotically equivalent to $Z_i/\sqrt{n_i q_i(1-q_i)}$ (up to a scalar factor), where q_i is a minor allele frequency (MAF) of the SNP (Web Appendix Section 3). For binary traits, the log odds ratio estimate $\widehat{\beta}_i$ is asymptotically equivalent to $Z_i/\sqrt{n_i r_i(1-r_i)q_i(1-q_i)}$, where $r_i = n_{case,i}/n_i$ is a proportion of case samples (Web Appendix Section 3). If all studies have similar ratios of cases and

controls, $r_i(1-r_i)$ term can be ignored. Therefore, $\tilde{\beta}_i = Z_i/\sqrt{n_i q_i(1-q_i)}$ and its standard error $\tilde{\sigma}_i = 1/\sqrt{n_i q_i(1-q_i)}$ can be used as inputs for both continuous and binary traits.

3. Results

3.1 *Simulation Studies*

To investigate the performance of TransMeta, we ran a series of simulations with varying assumptions on genetic effect heterogeneity across multiple ancestry groups. To generate SNPs with realistic MAF spectrums across different ancestry groups, we used Phase III of the HapMap Project (HMP3) data (Consortium et al., 2010). HMP3 consists of approximately 1.6 million SNPs, obtained from 1,184 subjects from 11 populations. We excluded the admixed African American population, combined the Japanese and Chinese as one population, and used the resulting 9 populations as seed populations to generate SNP genotypes.

The retrospective binary phenotype Y_{ik} of the k_{th} individual in the i_{th} study was generated using the following logistic regression model

$$\text{logit } Pr(Y_{ik}=1) = \beta_0 + \beta_i g_{ik}, \tag{7}$$

where g_{ik} is a genotype of the selected SNP, and β_i is a log odds ratio parameter. The intercept β_0 was chosen to have disease prevalence 0.05. In each replication, we randomly chose a SNP with a MAF of at least 1% in all populations, and generated SNP genotypes as $g_{ik} \sim Binomial(2, q_i)$, where q_i denotes the MAF of the selected SNP. We also used model (7) to estimate log odds ratio $\widehat{\beta}_i$ and its standard error $\widehat{\sigma}_i$ as the input data. In addition, we recorded $\Delta_i = sign(\widehat{\beta}_i)$, the direction of effect and the p-value p_i for testing $H_0 : \beta_i = 0$. We generated 500 cases and 500 controls for each of the 9 ancestry groups in triplicate, which resulted in a total of 27 studies with a total sample size of 13,500 cases and 13,500 controls.

3.1.1 *Type I error simulations.* To estimate type I error rates at stringent α levels, we generated 20 million replicates from model (7) with $\beta_i = 0$. Table 1 shows that the proposed methods control type I error rates at a very stringent significance level ($\alpha = 10^{-6}$) with the F_{st} kernel (denoted as TransMeta.Fst), although slightly conservative with the independent kernel (denoted as TransMeta.Indep). We also considered a setting where there is only one study per ancestry group. Each study now has 1500 cases and 1500 controls. We again used model (7) with $\beta_i = 0$ to simulate a total of 100 million replicates, and found that empirical type I error rates were well controlled (Supplementary Table 1).

[Table 1 about here.]

3.1.2 *Power Simulations.* Recently, Wang et al. (2013) carried out comparisons of trans-ethnic meta-analysis methods under five different scenarios, which cover a wide range of possible scenarios of genetic effect heterogeneity. We adopted these five scenarios:

(a) 'Trans-ethnic fixed-effect', where no heterogeneity exists in genetic effects at the causal SNP between populations, specifically that, each of the 27 studies carries a genetic relative risk of 1.12 at the causal SNP.

(b) 'Out-of-Africa effect', where each of the 18 studies from the non-African populations carries a genetic relative risk of 1.08, whereas the 9 studies from the African populations (LWK, MKK and YRI) present no genetic effects.

(c) 'Europe and south Asia effect', where the 12 studies from the European and south Asian populations (CEU, GIH, MEX and TSI) share the same genetic relative risk of 1.2, whereas the 15 studies from the remaining populations present no genetic effects.

(d) 'Heterogeneous Out-of-Africa effect', where the causal variant has genetic effects only in non-African populations, with the 6 studies from the east Asian populations (CHB+JPT and CHD) each carrying a genetic relative risk of 1.15 while the European and south Asian populations carry a genetic relative risk of 1.12.

(e) 'Environment modifying effect', where the causal variant has a genetic effect only in the populations living in Europe and USA, with the 9 studies from CHD, CEU and TSI each carry a genetic relative risk of 1.2.

In all scenarios, causal SNPs have the same direction of associations across ancestry groups. For each scenario, we generated 2,000 replicates to obtain empirical power. To perform a fair comparison between the frequentist and Bayesian methods, we generated 20 million SNPs under the null hypothesis and calculated Bayes factors using MANTRA. We observed that a log10 Bayes factor threshold larger than 5 corresponds to a p-value threshold less than $\alpha = 1.8 \times 10^{-6}$. To find a log10 Bayes factor threshold corresponding to the genome-wide significance level, we carried out a simple regression analysis between empirical type I error rates and log10 Bayes factors, and found that log10 Bayes factor= 6.34 corresponds to $\alpha = 5 \times 10^{-8}$ (see Web Appendix Section 4 for details).

Figure 1 shows the empirical power of the five methods under all five scenarios. TransMeta.Fst yields the highest or near highest power among the five methods, except in scenario (e). In scenario (a) where no heterogeneity exists, all five methods performed similarly, with FE having the highest power, as expected. In the remaining three scenarios with heterogeneous genetic effects that are not due to the environment modification, TransMeta.Fst outperformed the four existing meta-analysis methods. Unsurprisingly, RE yielded the lowest power across all five approaches. In scenario (e) where the genetic effect is influenced by environmental exposures, populations that are closely related do not necessarily share similar genetic effects. This violates the assumption of using the F_{st} to take account of the variability in genetic effects, and in this case, TransMeta.Indep yielded the highest power.

[Figure 1 about here.]

Figure 2 shows the empirical power of the five methods with one integrated study per ancestry group. The patterns of empirical power in this setting are very similar to what we

observed in Figure 1, where we had 3 substudies per ancestry group, except for RE-HE, which had slightly higher power than that of TransMeta.Indep. Since TransMeta.Indep used the identity matrix as the kernel matrix (i.e $\mathbf{K} = \mathbf{I}$) under this setting, the similar performance of TransMeta.Indep and RE-HE is not surprising. Overall, TransMeta.Fst attained similar or higher power over competing methods except in scenario (e).

[Figure 2 about here.]

The barplots in Supplementary Figure 2 and 3 summarize the power of the five methods at the more stringent level $\alpha = 5 \times 10^{-8}$; the results were quantitatively similar to the patterns we observed in Figure 1 and 2.

3.1.3 *Comparison Between Effect-size Based and Z-score Based TransMeta.* To demonstrate that Z-scores can be used for TransMeta as input summary statistics without loss of efficiency, we compared the power of the effect-size based and Z-score based TransMeta. The proportion of case samples was one (i.e $r_i = 1$) for all studies, so we ignored r_i in the transformation. We also obtained transformed Z-scores with sample sizes only, assuming that MAFs of SNPs are the same across all studies (i.e $\tilde{\beta}_i = Z_i/\sqrt{n_i}$). We included this setting because Z-scores are typically obtained without MAFs.

Figure 3 compares the power of the effect-size based and the Z-score based TransMeta under the same five scenarios used for Figure 1. The power of these two approaches was nearly identical when we used both sample sizes and MAFs for the Z-score transformations, and the power of the Z-score based TransMeta was slightly lower than the effect-size based TransMeta when only sample sizes were used for the Z-score transformations. For the one integrated study per ancestry group setting, the results were quantitatively similar to the patterns in Figure 3 (Supplementary Figure 4). At the genome-wide significance level, we again observed similar patterns as in Figure 3 and Supplementary Figure 4 (data not shown).

[Figure 3 about here.]

3.1.4 *Computation Time.* TransMeta provides scalable computation time for genome-wide datasets. To analyze 2,000 SNPs in the power simulations, both TransMeta.Fst and TransMeta.Indep took 20 seconds on average on a Linux cluster node with 2.80 GHz CPU. To analyze one million SNPs in a genome-wide dataset, TransMeta would require less than 3 hours. Among the competing methods, MANTRA was computationally expensive and took 45 and 95 minutes on average to analyze 2,000 SNPs with 9 and 27 studies, respectively. An R package 'TransMeta' has been developed to implement our proposed method and can be downloaded at the authors' website (https://sites.google.com/a/umich.edu/leeshawn/software).

3.2 *Application to Type 2 Diabetes (T2D) GWAS*

We re-analyzed the published T2D GWAS meta-analysis (Consortium et al., 2014). The aggregated data include 69 lead SNPs from the previously established T2D susceptibility loci, with 26,488 cases and 83,964 controls from four major ancestry groups of Europeans (12,171 cases and 56,862 controls), east Asians (6,952 cases and 11,865 controls), south Asians (5,561 cases and 14,458 controls), and Mexican and Mexican-Americans (1,804 cases and 779 controls). Association summary statistics – such as MAFs, effect size estimates, and standard errors – of the lead 69 SNPs were available for all four ancestry groups (Supplementary Table 3 of Consortium et al. (2014)). FE was employed to conduct the meta-analysis.

We applied TransMeta to the aggregated data along with the other four approaches. Due to the small number of SNPs in the aggregated dataset, estimates of F_{st} may be unreliable. Instead, we used the pairwise F_{st} from HMP3 (Supplementary Table 2). Supplementary Tables 3 and 4 list p-values (or Bayes factors) of the 69 SNPs with selected optimal ρs of TransMeta. Among those 69 SNPs, 37 had optimal $\rho < 1$ under TransMeta.Fst. Figure 4 compares p-values of TransMeta.Fst and FE for different selected optimal ρs. When the selected optimal $\rho = 0$, our method yields a smaller p-value than FE, which indicates that

TransMeta can be more powerful than FE. When the selected $\rho = 1$, and hence FE is the optimal test, FE yields a smaller p-value than TransMeta, but the difference is minimal.

[Figure 4 about here.]

At the significance level $\alpha = 1.8 \times 10^{-6}$ or a log10 Bayes factor > 5, TransMeta.Fst, TransMeta.Indep, FE and RE-HE all identified 31 SNPs, while RE and MANTRA identified 18 and 28 SNPs, respectively. At the genome-wide significance level of $\alpha = 5 \times 10^{-8}$ or a log10 Bayes factor > 6.34, both TransMeta.Fst and TransMeta.Indep identified 24 SNPs, while FE, RE, RE-HE and MANTRA identified 23, 12, 22 and 19 SNPs respectively.

At the genome-wide significance level, TransMeta was able to identify one more SNP, rs10830963, with TransMeta.Fst p-value$= 2.98 \times 10^{-8}$ (selected optimal $\rho = 0.25$) and TransMeta.Indep p-value$=3.76 \times 10^{-8}$ (selected optimal $\rho = 0.25$), respectively. In contrast, p-values of FE, RE and RE-HE were all larger than 10^{-7}, and MANTRA log10 Bayes factor was 5.6. The SNP rs10830963 is located in Melatonin receptor 1-B, which belongs to the seven transmembrane G protein-coupled receptor superfamily, and a previous study has shown that this SNP is associated with fasting glycemia and T2D (Sparsø et al., 2009).

Figure 5 displays a forest plot of odds ratios and their corresponding confidence intervals for this SNP (extracted from Supplementary Table 3 in Consortium et al. (2014)). The odds ratios of Europeans, south Asians and Mexicans were all close to 1.1, although the odds ratio for Mexicans was non-significant due to small sample size. In contrast, the odds ratio in east Asians was close to one. Since east Asians are genetically more distant than other populations (Supplementary Table 2), this result indicates that our approach to modeling genetic effect heterogeneity using genetic distance can increase power.

[Figure 5 about here.]

4. Discussion

We have proposed a novel trans-ethnic meta-analysis framework that flexibly models the genetic effect heterogeneity across ancestry groups. The framework incorporates the genetic distances to model the genetic effect heterogeneity and adaptively uses variance component test to achieve robust power. Simulations and the trans-ethnic T2D GWAS application suggest that our approach can improve power when genetic effect-size heterogeneity exists.

Since TransMeta.Fst accommodates genetic similarity to model the effect size similarity, we recommend TransMeta.Fst as the primary test. However, if there is evidence suggesting that the genetic effects are modified by non-genetic exposures (such as environmental or lifestyle factors), TransMeta.Indep may be a better choice. To avoid data fishing, the choice of using TransMeta.Fst or TransMeta.Indep needs to be made prior to data analysis. For the sequence of ρ values in the grid search, we found that using a wide range of ρs does not necessarily improve power. In fact, in Supplementary Figure 5, we applied a denser grid with eleven points of $\rho = (0, 0.1, \ldots, 0.9, 1)$ in the power simulations and found that the resulting power is very similar or even identical to the power based on $\rho = (0, 0.09, 0.25, 1)$. So we suggest using $\rho = (0, 0.09, 0.25, 1)$ as the default sequence of ρ values. We note that it is not required to select ρ from the grid prior to perform the analysis, since TransMeta automatically select the optimal ρ, and calculate p-values while accounting for the selection.

Unlike the I^2 statistic (Higgins et al., 2002), which was developed to measure the extent of heterogeneity, the optimal ρ is set as the value (over a pre-specified grid) whose score statistic has the smallest p-value among all. As a result, the optimal ρ should not be interpreted as a metric of heterogeneity. For example, we counted the number of optimal ρ values in each of the five scenarios in the power simulations (Supplementary Table 6) and observed that in the homogeneous effect size scenario, only less than half of the optimal ρ values in TransMeta.Fst are determined to be 1. (Please recall that ρ equals to 1 models homogeneous effect sizes;

the closer ρ is to 0, the stronger the indication of heterogeneity.) However, the optimal ρ does provide some insights into the extent of heterogeneity. For example, in our power simulations, we observed that the I^2 statistic tends to decrease as the optimal ρ increases, as shown in Supplementary Table 5. (Please recall that $I^2 = 0$ means homogeneity; and the level of heterogeneity increases as I^2 approaches to 1.) In addition, we observed in Supplementary Table 6 that when heterogeneity does exist, such as scenarios (b) - (e) in the power simulations, the majority of the optimal ρ values in TransMeta.Fst are selected to be 0. Similar trends are observed in TransMeta.Indep, data not shown.

Our score statistics S_ρ is a linear combination of two components, each models the genetic effect homogeneity and the genetic effect heterogeneity, respectively. As a result, although TransMeta is designed to tackle heterogeneous effect sizes situations, it can also handle homogeneity scenarios. In fact, the right panel of Figure 4 demonstrates that under genetic effect homogeneity, our approach achieves almost the same statistical significance as FE.

We note that the empirical power of MANTRA is similar or lower than that of TransMeta.Fst in scenarios (b)-(d), but is higher in scenario (e)(Figure 1 and 2). This occurred because under scenario (e), the genetic distance does not provide guidance to the genetic effect similarity, which violates the key assumption in both MANTRA and TransMeta.Fst. Since MANTRA groups studies into clusters data-adaptively, it is more robust than TransMeta.Fst under this situation. As a result, MANTRA had higher power than TransMeta.Fst.

RE-HE is equivalent to testing for $\tau = 0$ while adaptively selecting ρ under model (4) with $\mathbf{K} = \mathbf{I}$. When we have one integrated study per ancestry group, the $\mathbf{K}$ matrix in TransMeta.Indep is exactly equal to $\mathbf{I}$, which makes RE-HE equivalent to TransMeta.Indep in terms of testing (although they use different approaches to obtain p-values). As a result, RE-HE and TransMeta.Indep have similar power in all five scenarios in Figure 2. When we have three studies per ancestry group (Figure 1), RE-HE treats each study as its own cluster. In

contrast, TransMeta.Indep groups studies in the same ancestry. As a result, TransMeta.Indep yields higher power than RE-HE in nearly all scenarios in Figure 1.

The advancement of high-throughput sequencing technologies enables us to study associations of rare variants. Since the statistical power of single rare variant test is low, gene/region based tests are commonly used for rare variant association analysis. Several groups have developed gene/region-based rare variant meta-analysis methods (Lee et al., 2013; Tang and Lin, 2014; Liu et al., 2014). In the future, we will extend the framework of using genetic distance for modeling the genetic effect heterogeneity to gene/region based tests.

5. Supplementary Materials

Web Appendices, Tables and Figures referenced in Section 2 and 3, along with the R code which implements our proposed methods on the T2D data are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was supported by NIH grant R00 HL113164. The authors wish to thank for Dr. Andrew P. Morris for the MANTRA software.

References

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10,** 101–129.

Consortium, D. S., Consortium, D. M., Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics* **46,** 234–244.

Consortium, I. H. . et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58.

Cooper, R. S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: implications for multiethnic samples. *Human molecular genetics* **17,** R151–R155.

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics* pages 323–333.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11,** 446–450.

Evangelou, E. and Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics* **14,** 379–389.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics* **88,** 586–598.

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics* **93,** 42–53.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63,** 1079–1088.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature genetics* **46,** 200–204.

Marigorta, U. M. and Navarro, A. (2013). High trans-ethnic replicability of gwas results implies common causal variants. *PLoS genetics* **9,** e1003566.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9,** 356–369.

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology* **35,** 809–822.

Sparsø, T., Bonnefond, A., Andersson, E., Bouatia-Naji, N., Holmkvist, J., Wegner, L., Grarup, N., Gjesing, A. P., Banasik, K., Cavalcanti-Proença, C., et al. (2009). G-allele of intronic rs10830963 in mtnr1b confers increased risk of impaired fasting glycemia and type 2 diabetes through an impaired glucose-stimulated insulin release studies involving 19,605 europeans. *Diabetes* **58,** 1450–1456.

Tang, Z.-Z. and Lin, D.-Y. (2014). Meta-analysis of sequencing studies with heterogeneous genetic associations. *Genetic epidemiology* **38,** 389–401.

Wang, X., Chua, H.-X., Chen, P., Ong, R. T.-H., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-C., Tay, W.-T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Human molecular genetics* page ddt064.

Wright, S. (1949). The genetical structure of populations. *Annals of eugenics* **15,** 323–354.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89,** 82–93.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

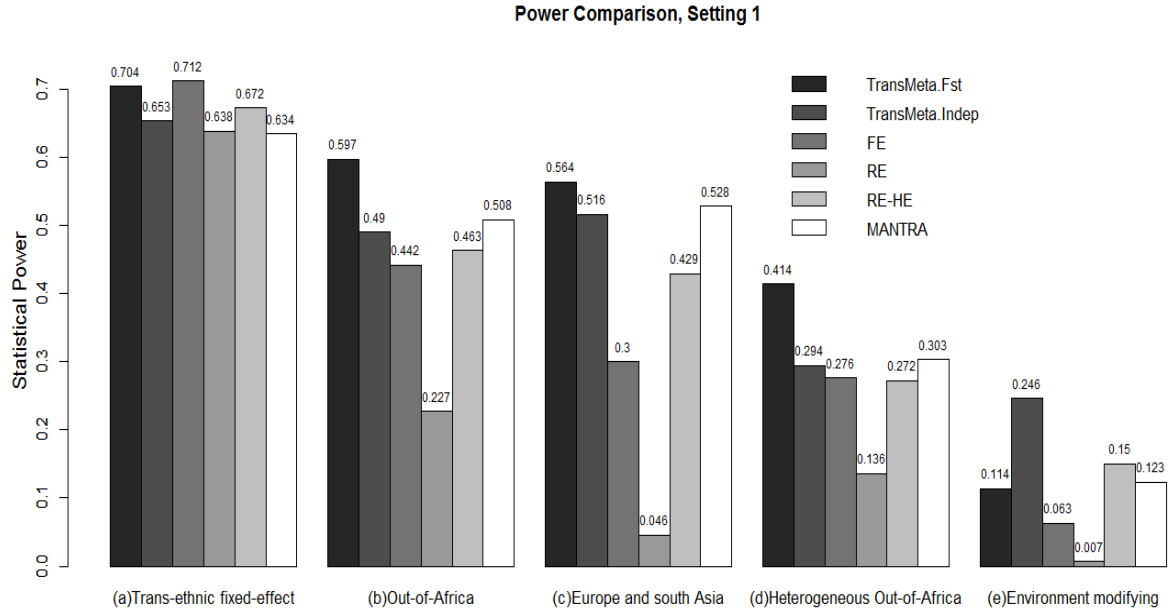

Figure 1: Empirical power for TransMeta and existing methods under the five effect-size scenarios. Three studies were simulated per ancestry group, each with 500 cases and 500 controls. The empirical power was obtained based on 2000 replicates with the level of significance defined as a p-value less than 1.8×10^{-6} or as a log10 Bayes factor larger than 5. The five-effect size scenarios are (a) 'Trans-ethnic fixed-effect', where no heterogeneity exists in allelic effects at the causal SNP between populations; (b) 'Out-of-Africa effect', where only studies from the non-African populations carry the causal variant; (c) 'Europe and south Asia effect', where only studies from the European and south Asian populations carry the causal variant; (d) 'Heterogeneous Out-of-Africa effect', where the causal variant has genetic effects only in non-African populations, but the effect size in the east Asian populations is different from that in the European and south Asian populations; (e) 'Environment modifying effect', where the causal variant has genetic effect only in the populations living in Europe and USA.

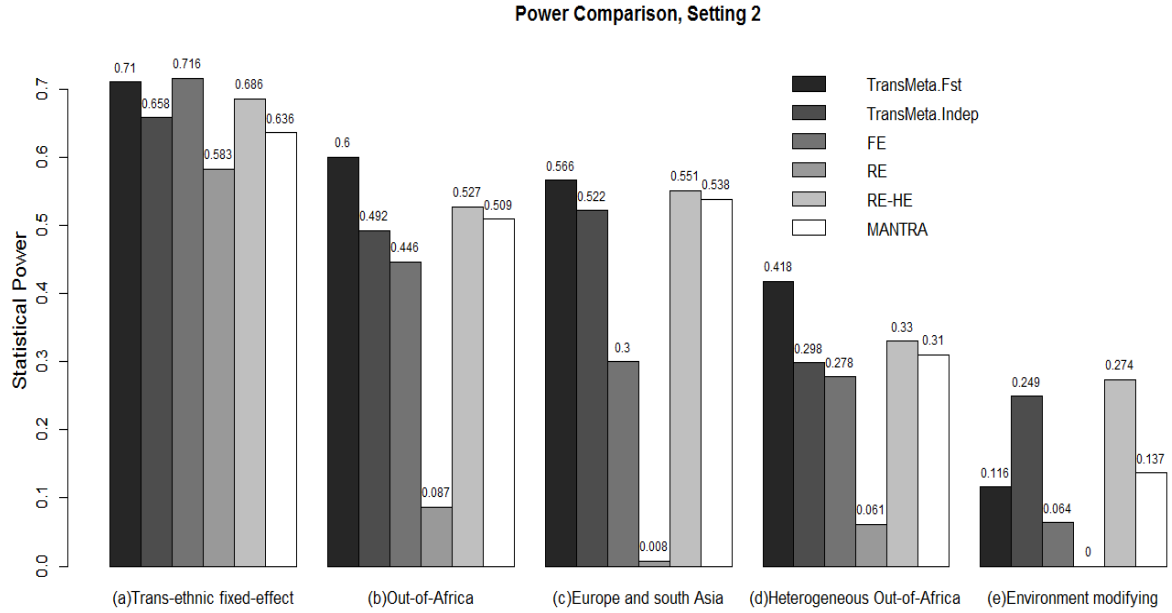

Figure 2: Empirical power for TransMeta and existing methods under the five effect-size scenarios. One integrated study was simulated per ancestry group, each with 1500 cases and 1500 controls. The empirical power was obtained based on 2000 replicates with the level of significance defined as a p-value less than 1.8×10^{-6} or as a log10 Bayes factor larger than 5. The five effect-size scenarios are (a) 'Trans-ethnic fixed-effect', where no heterogeneity exists in allelic effects at the causal SNP between populations; (b) 'Out-of-Africa effect', where only studies from the non-African populations carry the causal variant; (c) 'Europe and south Asia effect', where only studies from the European and south Asian populations carry the causal variant; (d) 'Heterogeneous Out-of-Africa effect', where the causal variant has genetic effects only in non-African populations, but the effect size in the east Asian populations is different from that in the European and south Asian populations; (e) 'Environment modifying effect', where the causal variant has genetic effect only in the populations living in Europe and USA.

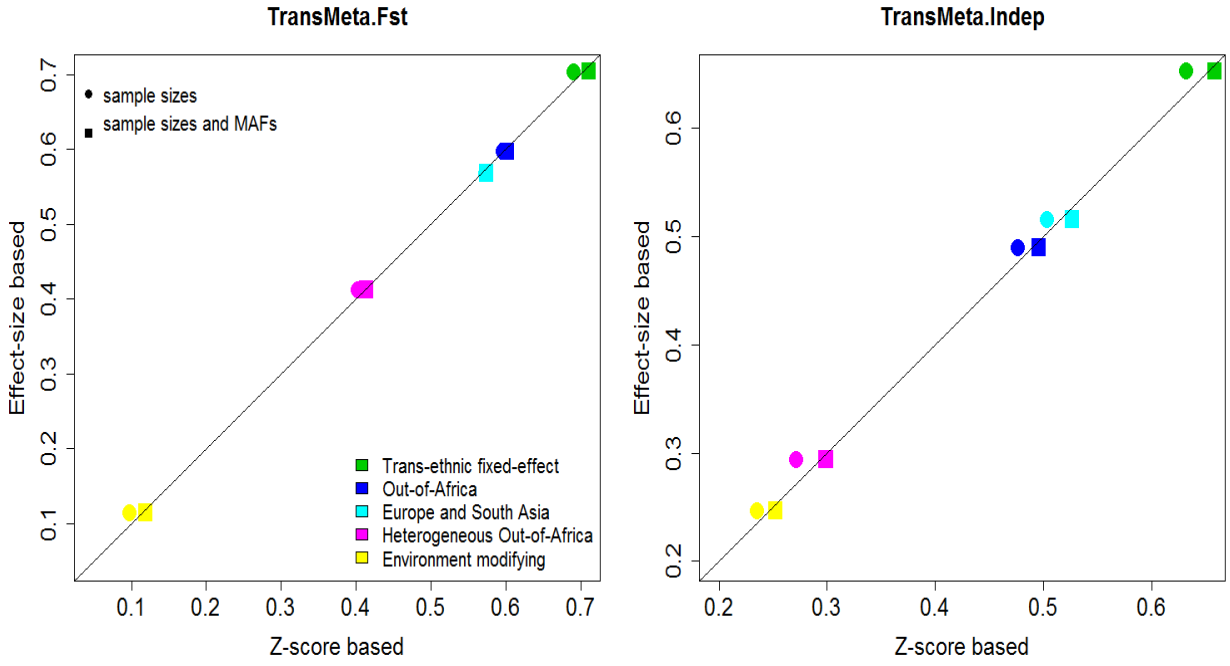

Figure 3: Power comparison of the effect-size and Z-score based TransMeta under the five effect size scenarios. Three studies were simulated per ancestry group, each with 500 cases and 500 controls. The empirical power was obtained based on 2000 replicates with the level of significance defined as a p-value less than 1.8×10^{-6}. The left panel is based on TransMeta.Fst and the right panel is based on TransMeta.Indep. In each plot, the x-axis denotes empirical power of the the Z-score based TransMeta and the y-axis denotes empirical power of effect-size based TransMeta. The solid dots represent the power of transformed Z-scores using only sample sizes, and the solid squares represent transformed Z-scores using both sample sizes and MAFs.

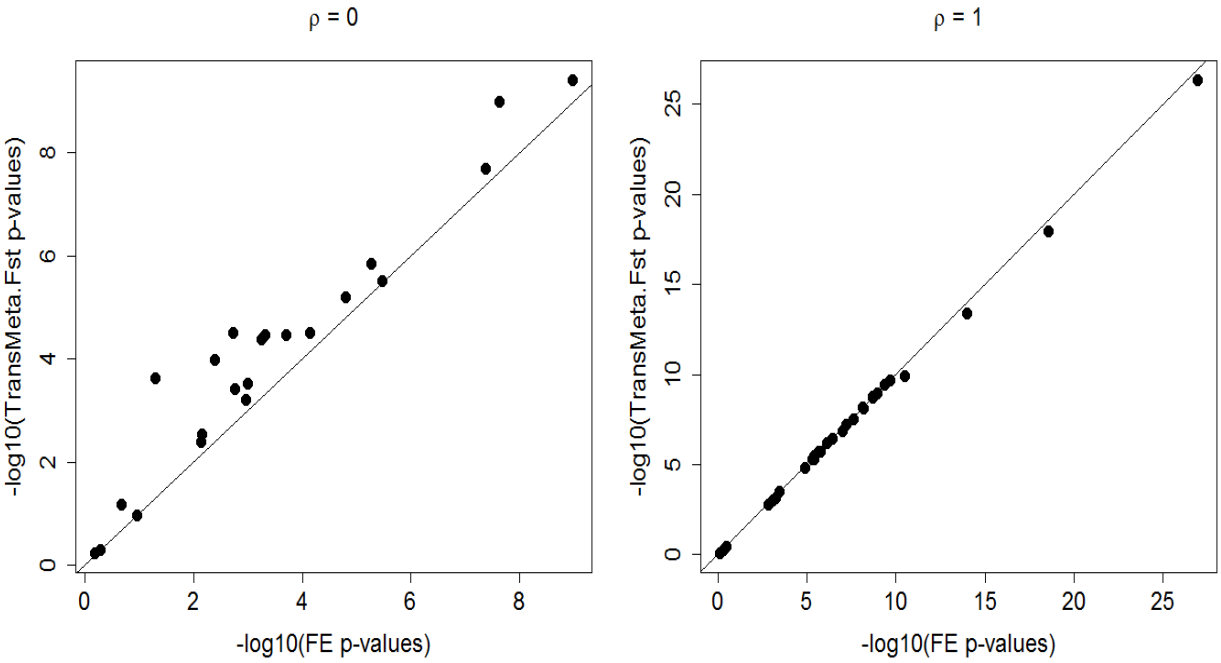

Figure 4: Comparison of p-values of TransMeta.Fst and FE for 69 lead SNPs in T2D meta-analysis data. The left panel displays p-values of SNPs whose TransMeta.Fst ρ is zero; the right panel displays p-values of SNPs whose TransMeta.Fst ρ is one. In each plot, the x-axis denotes $-\log_{10}$(FE p-values), and the y-axis denotes $-\log_{10}$(TransMeta.Fst p-values).

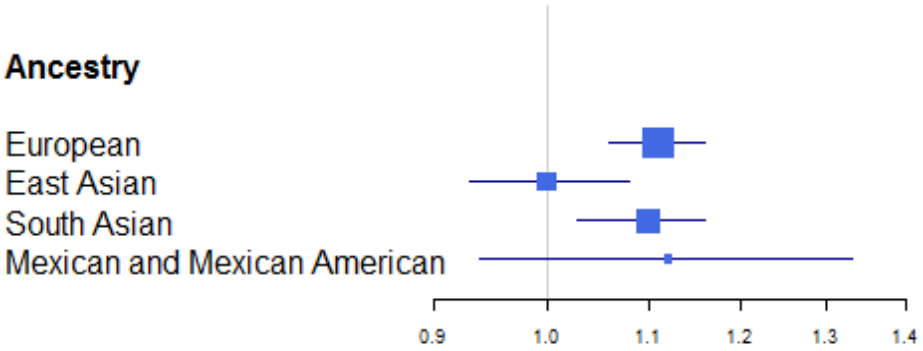

Figure 5: Forest plot of the estimated OR and 95 % CI for rs10830963 in each ancestry group. The association signal of rs10830963 is detected by TransMeta only.

Table 1: Type-I error rate estimates at different α levels based on 20 million replicates. Each entry represents an estimated type I error rate calculated using the proportion of empirical p-values smaller than the given level α. Three studies were simulated per ancestry group, and each study had 500 cases and 500 controls.

	$\alpha = 10^{-2}$	10^{-3}	10^{-4}	10^{-5}	10^{-6}
TransMeta.Fst	9.7×10^{-3}	1.1×10^{-3}	9.6×10^{-5}	1.0×10^{-5}	9.5×10^{-7}
TransMeta.Indep	9.8×10^{-3}	0.9×10^{-3}	7.6×10^{-5}	5.8×10^{-6}	4.0×10^{-7}