General Framework for Meta-analysis of Haplotype Association Tests

Shuai Wang[1], Jing Hua Zhao[2], Ping An[3], Xiuqing Guo [4], Richard A. Jensen [5,16], Jonathan Marten [6], Jennifer E. Huffman [6], Karina Meidtner [7], Heiner Boeing [8], Archie Campbell [9], Kenneth M Rice[15], Robert A Scott[2], Jie Yao [4], Matthias B Schulze [7,10], Nicholas J Wareham[2], Ingrid B. Borecki[3], Michael A. Province[3], Jerome I. Rotter [4], Caroline Hayward [6 9], Mark O. Goodarzi[11], James B. Meigs [12 13], Josée Dupuis [1∗ 14],

1 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
2 MRC Epidemiology Unit, University of Cambridge, School of Clinical Medicine, Box 285 Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, United Kingdom
3 Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, USA
4 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, LABioMed at Harbor-UCLA Medical Center, Torrance, California, USA
5 Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA
6 MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, United Kingdom
7 Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany
8 Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany
9 Generation Scotland, Centre for Genomic and Experimental Medicine, University of Edinburgh Institute of Genetic and Molecular Medicine, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, United Kindom
10 German Center for Diabetes Research (DZD), Germany
11 Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, USA
12 Massachusetts General Hospital, General Medicine Division, Boston, Massachusetts 02114, USA
13 Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA
14 National Heart, Lung, and Blood Institute (NHLBI) Framingham Heart Study, Framingham, Massachusetts 01702, USA
15 Department of Biostatistics, University of Washington, Seattle, WA, USA
16 Department of Medicine, University of Washington, Seattle, Washington, USA
∗ Correspondence: Josée Dupuis, 801 Massachusetts Ave, 3rd Floor, Boston, MA 02118 (617)638-5880, dupuis@bu.edu

Abstract

For complex traits, most associated single nucleotide variants (SNV) discovered to date have a small effect, and detection of association is only possible with large sample sizes. Because of patient confidentiality concerns, it is often not possible to pool genetic data from multiple cohorts, and meta-analysis has emerged as the method of choice to combine results from multiple studies. Many meta-analysis methods are available for single SNV analyses. As new approaches allow the capture of low frequency and rare genetic variation, it is of interest to jointly consider multiple variants to improve power. However, for the analysis of haplotypes

formed by multiple SNVs, meta-analysis remains a challenge, because different haplotypes may be observed across studies. We propose a two-stage meta-analysis approach to combine haplotype analysis results. In the first stage, each cohort estimate haplotype effect sizes in a regression framework, accounting for relatedness among observations if appropriate. For the second stage, we use a multivariate generalized least square meta-analysis approach to combine haplotype effect estimates from multiple cohorts. Haplotype-specific association tests and a global test of independence between haplotypes and traits are obtained within our framework. We demonstrate through simulation studies that we control the type-I error rate, and our approach is more powerful than inverse-variance-weighted meta-anaysis of single SNV analysis when haplotype effects are present. We replicate a published haplotype association between fasting glucose-associated locus (G6PC2) and fasting glucose in 7 studies from the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium and we provide more precise haplotype effect estimates.
keywords: meta-analysis, haplotype association tests, family samples, linear mixed effects model

Introduction

In recent years, genome-wide association studies (GWAS) have identified multiple common variants associated with disease and disease-related traits. In a typical GWAS, association between a trait and genetic variants is tested one variant at-a-time, and variants with weak association routinely fail to be detected, especially in small cohorts. Therefore, meta-analysis is often used by large consortia to increase statistical power [Stram, 1996; Zeggini et al., 2008; Scott et al., 2012; Dupuis et al., 2010] to detect variants with a moderate to weak association with the trait of interest. Even with large meta-analysis, variants identified to date only explain a small proportion of the total heritability. In order to identify the source of the unexplained heritability, emerging approaches have attempted to account for multiple variants at once when evaluating association with a trait. Such approaches include penalized regression methods [Wu et al., 2009; Li et al., 2011], pathway analysis [Holden et al., 2008], gene-based tests such as burden [Madsen and Browning, 2009] and SKAT [Wu et al., 2010], and haplotype analysis [Schaid et al., 2002; Tregouet et al., 2004; Liu et al., 2008]. Power of these approaches can be enhanced by increasing sample size or combining multiple studies. Methods for meta-analysis of gene-based tests are well established and widely used [Hu et al., 2013; Liu et al., 2014], but there are no widely-used methods for the meta-analysis of haplotype association tests.

In this article, we propose a meta-analysis approach to combine haplotype association results from multiple studies. In the first step of our method, each study provides regression estimates and covariance matrix of haplotype effects, with adjustment for familial correlation to accommodate familial samples or cryptic relatedness. In our second step, cohort-specific haplotype effect estimates are pooled using a multivariate generalized least square meta-analysis approach. A global association test and evaluation of the effect of each haplotype can be obtained within our framework. We perform a simulation study to evaluate our approach, comparing results with more traditional meta-analysis of single-variant association tests and gene-based tests. Finally, we replicate a published haplotype association between a fasting glucose-associated locus (G6PC2) and fasting glucose in 7 studies from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium and are able to provide more precise haplotype effect estimates than the prior report involving haplotype estimates from a single cohort [Mahajan et al., 2015]. Code implementing the novel approach, along with a tutorial, is available at http://sites.bu.edu/fhspl/publications/metahaplo.

Methods

Haplotype association test at cohort level

Our approach is based on Zaykin et al.'s haplotype analysis method [Zaykin et al., 2002] for unrelated samples. We incorporate random effects to account for family structure, making the approach applicable to both family-based cohorts or unrelated samples (or a mix of the two). Assuming in a cohort a total of n subjects are sequenced in a region with q SNVs and as a result, K haplotypes are observed. A general linear (mixed-effect) model can be written as:

$$\mathbf{Y} = \mathbf{X}\alpha + \beta_1\mathbf{h_1} + ... + \beta_K\mathbf{h_K} + \mathbf{b} + \epsilon, \tag{1}$$

where
$\mathbf{Y}$ is a $n \times 1$ quantitative trait vector, $\mathbf{X}$ is a $n \times p$ matrix of covariates (without intercept) including for example, age, sex and associated genetic principal components controlling for potential population stratification, α is a $p \times 1$ coefficient vector for the p covariates adjusted, each $n \times 1$ vector $\mathbf{h_m}$ $(m = 1, \cdots, K)$ is the expected haplotype dosage, $\mathbf{b}$ is a $n \times 1$ random effect vector to account for the relatedness within families, and ϵ is a $n \times 1$ vector of the random error terms. When haplotype m of the j-th $(j = 1, \cdots, n)$ subject is observed, h_{mj}, the j-th entry in $\mathbf{h_m}$ is either 0, 1 or 2, i.e. the number of copies of haplotype m the j-th subject carries. Otherwise, expected haplotype dosages $E[h_{mj}|\mathbf{G_j}]$ are inferred from $\mathbf{G_j}$, the $q \times 1$ genotype vector of the j-th subject using statistical algorithms such as the Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. For the j-th subject, the sum of the K haplotype dosage $\sum_{m=1}^{m=K} h_{mj}$ is always equal to 2. The $n \times 1$ random effect vector $\mathbf{b}$ is assumed to follow a normal distribution $N(\mathbf{0}, \sigma_a^2\mathbf{\Phi})$, where σ_a^2 is the additive variance and $\mathbf{\Phi}$ is the relationship matrix (with entries equal to twice the kinship coefficient for related pairs and 0 for unrelated pairs) derived from pedigree structure or genome-wide information; in unrelated samples, the matrix $\mathbf{\Phi}$ reduces to $\mathbf{I}$, the $n \times n$ identify matrix.. Finally, the vector of error terms ϵ follows a normal distribution $N(0, \sigma_e^2\mathbf{I})$, where σ_e^2 is the variance of the error term.

Let $\mathbf{X_o} = (\mathbf{X}, \mathbf{h_1}, \cdots, \mathbf{h_K})$ denote the overall design matrix of size $n \times (p + K)$, and define the overall variance matrix as $\mathbf{\Omega} = \sigma_a^2\mathbf{\Phi} + \sigma_e^2\mathbf{I}$. The parameters α and β_k $(k = 1, \cdots, K)$ are estimated as $(\mathbf{X_o^T\hat{\Omega}^{-1}X_o})^{-1}\mathbf{X_o^T\hat{\Omega}^{-1}Y}$, where $\mathbf{\hat{\Omega}}$ is evaluated at the maximum likelihood estimates $\hat{\sigma_a^2}$ and $\hat{\sigma_e^2}$ which can be obtained using the lmekin function in R package coxme [Therneau, 2012]. Then the estimated variance of the effect estimates is obtained from $(\mathbf{X_o^T\hat{\Omega}^{-1}X_o})^{-1}$. It reduces to an ordinary linear regression when applied to unrelated samples.

Meta-Analysis

We assume a total of N cohorts participate in the meta-analysis and the i-th $(i = 1, \cdots, N)$ cohort provides the estimates $\hat{\boldsymbol{\beta}^i}$ and the covariance matrix $\hat{var}(\hat{\boldsymbol{\beta}^i})$ of the haplotype effects for K^i haplotypes, and a total of K' haplotypes are observed in at least one cohort. We propose a multivariate meta-analysis approach [Becker and Wu, 2007] based on generalized weighted least squares to combine the length K^i haplotype effect estimates from each cohort, denoted by $\hat{\boldsymbol{\beta}^i}$ for studies $i = 1, \cdots, N$, into a single effect estimate vector $\tilde{\boldsymbol{\beta}}$ of length K'. The generalized weighted least square approach is formulated as:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}^1} \\ \vdots \\ \hat{\boldsymbol{\beta}^N} \end{pmatrix} = \mathbf{W}\boldsymbol{\beta} + \mathbf{e} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta^1 \\ \vdots \\ \beta^{K'} \end{pmatrix} + \mathbf{e}, \tag{2}$$

where

$\hat{\boldsymbol{\beta}}^{\mathbf{i}}$ $(i = 1, \cdots, N)$ is the $K^i \times 1$ haplotype coefficient vector for cohort i;
$\hat{\boldsymbol{\beta}}$ is the stacked haplotype coefficient vector from $\hat{\boldsymbol{\beta}}^{\mathbf{i}}$ $(i = 1, \cdots, N)$;
$\boldsymbol{\beta}$ is the $K' \times 1$ coefficient vector of the haplotype effects;
$\mathbf{W} = \begin{pmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_N \end{pmatrix}$ is a $\sum_i K^i \times K'$ design matrix stacked from the N cohorts, where $\boldsymbol{W}_i$ $(i = 1, \cdots, N)$ is a $K^i \times K'$ matrix, with zeros and one in each row indicating which haplotype effect is observed by cohort i;
$\mathbf{e}$ is the error term which is assumed to have a multivariate normal distribution with a mean of $\mathbf{0}$ and a covariance matrix of $\Sigma = \begin{pmatrix} var(\hat{\boldsymbol{\beta}}^1) & \cdots & 0 \\ \vdots & var(\hat{\boldsymbol{\beta}}^k) & \vdots \\ 0 & \cdots & var(\hat{\boldsymbol{\beta}}^N) \end{pmatrix}$.
Note that in the meta-analysis stage, cohort haplotypes are reordered to match the order assigned to the K' haplotypes observed in at least one cohort, and the design matrix $\mathbf{W}$ reflects this re-ordering. Furthermore, because $\boldsymbol{\Sigma}$ is unknown, we substitute the sample estimate $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{var}(\hat{\boldsymbol{\beta}}^1) & \cdots & 0 \\ \vdots & \hat{var}(\hat{\boldsymbol{\beta}}^k) & \vdots \\ 0 & \cdots & \hat{var}(\hat{\boldsymbol{\beta}}^N) \end{pmatrix}$, hence the weighted least square estimator of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = (\mathbf{W'}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{W})^{-1}\mathbf{W'}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\beta}}$ and $\mathbf{V} = \mathbf{Var}(\tilde{\boldsymbol{\beta}}) = (\mathbf{W'}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{W})^{-1}$.

Hypothesis Testing

The global null hypothesis of no association of any haplotype with the trait is expressed as

$$H_0 : \beta^1 = \beta^2 = ... = \beta^{K'} \tag{3}$$

To construct a test statistic to test for haplotype association, we reparameterize it into the equivalent null hypothesis, where β^1 is chosen from commonly observed haplotypes:

$$H_0 : \boldsymbol{\gamma} = \begin{pmatrix} \gamma^1 \\ \vdots \\ \gamma^{K'-1} \end{pmatrix} = \begin{pmatrix} \beta^2 - \beta^1 \\ \vdots \\ \beta^{K'} - \beta^1 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4}$$

The null hypothesis can be tested using a Wald test statistic of the form

$$\chi^2 = \hat{\boldsymbol{\gamma}}^{\mathbf{T}}[\mathbf{V}^*]^{-1}\hat{\boldsymbol{\gamma}}. \tag{5}$$

where $\hat{\boldsymbol{\gamma}}$ is estimated from $\tilde{\boldsymbol{\beta}}$ and $\mathbf{V}^*$ is the covariance matrix of $\hat{\boldsymbol{\gamma}}$, with a dimension of $(K' - 1) \times (K' - 1)$ and the jj'-th element having the form $V^*_{jj'} = V_{jj'} - V_{j1} - V_{j'1} + V_{11}$. Under the null hypothesis, the wald-test statistic follows a $\chi^2_{K'-1}$ distribution asymptotically.

Cohorts for Heart and Aging Research in Genetic Epidemiology Consortium

The Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium comprises multiple studies with the common goal of identifying genes and loci associated with cardiovascular-related traits. Seven CHARGE cohorts contributed to a meta-analysis evaluating the association between genetic variants and fasting glucose in 25,305 non-diabetic participants (Table 1). Fasting glucose levels in mmol/l were analyzed in participants free of type-2 diabetes. Type-2 diabetes was defined by cohorts referring to at least one of the following criteria: a physician diagnosis of type-2 diabetes, on the anti-diabetic treatment of type-2 diabetes, fasting plasma glucose$\geq 7mmol/l$, random plasma glucose$\geq 11.1mmol/l$ or hemoglobin A1C $\geq 6.5\%$. Study-specific sample exclusions were detailed in [Wessel et al., 2015].

Table 1. CHARGE cohorts

Cohort	Sample size
Generation Scotland: Scottish Family Health Study* (GS)	7678
Framingham Heart Study* (FHS)	6561
Cardiovascular Health Study (CHS)	3525
Family Heart Study* (FamHS)	3393
Multi-Ethnic Study of Atherosclerosis (MESA)	2507
FENLAND (FLD)	1341
European Prospective Investigation into Cancer and Nutrition, Potsdam (EPIC-Potsdam)	300
Total	25305
*: Family-based cohort	

Genotypes were obtained from the Illumina HumanExome BeadChip [Grove et al., 2013], a genotyping array containing 247,870 variants discovered through exome sequencing in $\sim$ 12,000 individuals, in which $\sim$ 75% of the variants are low-frequency variants (MAF$< 0.5\%$). The main content of the chip comprises protein-altering variants (nonsynonymous coding, splice-site and stop gain or loss codons) seen at least three times in a study and in at least two studies providing information to the chip design. We selected 4 *G6PC2* variants previously studied for their haplotype association with fasting glucose [Mahajan et al., 2015].

Simulation Studies

To evaluate the validity and power of our approach, we perform a simulation study varying the number of cohorts included in the meta-analysis (5 or 10), and the type of samples (unrelated, family-based, mix of the two). We also vary the sample size from 100 up to 1600 subjects per cohort. See Table 2 for a description of the various study designs investigated in type-I error rate and power.

Simulated trait values are dependent on sex, age and haplotypes/genetic variants (power evaluation only). Sex of mothers and fathers (founders) are fixed in a heterosexual marriage but are randomly assigned to offspring, with equal probability. The age for unrelated individuals and the first offspring in a family are generated from a uniform distribution over the range 30 to 50. Additional offspring's ages are set to be within 5 years of the first offspring with at least a one year gap (no twins), using a uniform distribution. For family samples, the age of the mother is restricted to be 20 to 45 years older than her offspring, and the father's age to be within 5 years of the mother's age, with a restriction that the age be at least 20 years older than the older offspring.

We select the known T2D associated genes *G6PC2* (chromosome 2) (Table 3-4) and *JAZF1* (chromosome 7) (Table 5-6) to generate the reference panel haplotypes (Tables 3 and 4). We use the observed haplotypes and frequencies estimated by EM-algorithm from 6561 participants from the Framingham Heart Study. For example, in *JAZF1* no single haplotype has a frequency greater than 25% (Table 6) and 8 haplotypes have frequency greater than 1%.

Genotypes are simulated by randomly assigning a pair of haplotypes to founders, and by dropping randomly-selected haplotypes to offspring assuming no recombination within haplotypes. Although phasing information is available in our simulation setting, we do not use the phase information when implementing our approach because such information is not typically available in real datasets. We use the EM algorithm to infer expected haplotype dosage conditional on genotypes, via R package haplo.stats [Sinnwell and Schaid, 2013].

When estimating haplotype effects at the cohort-level, rare haplotypes (frequency$< 0.1\%$) are collapsed, to stabilize the computation and to avoid potential singularities due to high LD among SNVs.

Type-I error rate

For evaluating the type-I error rate of our new approach, a trait unassociated with the haplotypes is simulated using a multivariate normal distribution with mean $\hat{\mu} = 0.02 \times \mathbf{age} + 0.5 \times \mathbf{sex}$

(sex is set to 1 for males and to 2 for females) and a covariance matrix $\sigma_a^2\Phi + \sigma_e^2 I$, with $\sigma_a^2 = \sigma_e^2 = 0.5$. Age and sex explained about 10% and 5% of the trait variance, respectively, resulting in a trait with moderate heritability ($h^2 = \frac{\sigma_a^2}{Var(Y)} \approx 0.42$).

Cohort-specific analyses are performed by first estimating haplotypes using the EM algorithm implemented in the R package haplo.stats, followed by regression analysis using haplotype dosages and covariates as independent variables. Cohort results are then meta-analyzed using the novel approach previously described, and the global association test p-values are recorded. Ten thousand simulations are performed to assess the type-I error rate in all scenarios (Table 2).

Table 2. Study Designs for Type-I error evaluation

Study Design	# cohort	Sample Sizes	Type-I error (G6PC2)	Type-I error (JAZF1)
1	5	250 NF2 (×5)	0.010	0.010
2	5	250 NFv (×5)	0.010	0.012
3	5	100 NF2, 175 NF2, 400 U, 700 U, 1000 U	0.013	0.010
4	5	100 NFv, 175 NFv, 400 U, 700 U, 1000 U	0.011	0.011
5	5	100 NFv, 175 NFv, 250 NFv, 325 NFv, 400 NFv	0.011	0.012
6	10	250 NF2 (×5); 1000 U (×5)	0.010	0.011
7	10	400 U, 700 U, 1000 U, 1300 U, 1600 U	0.008	0.012
8	5	100 NF2, 175 NF2, 250 NF2, 325 NF2, 400 NF2	0.012	0.011
9	5	250 NF2, 125 NF2 (×2), 375 NF2 (×2) 1000 U, 500 U (×2), 1500 U (×2)	0.011	0.011
10	10	250 NFv (×7), 1000 U (× 3)	0.012	0.011

NF2: Nuclear family with 2 offspring;

NFv: Nuclear family with the number of offspring randomly selected to be between 1 and 4

U: Unrelated subjects;

Table 3. *G6PC2* variants

Name	Chr	MapInfo	dbSNPID	Minor	Major	FHS MAF
exm-rs560887	2	169763148	rs560887	A	G	0.293
exm239664	2	169763262	rs138726309	T	C	0.0036
exm239667	2	169764141	rs2232323	C	A	0.0078
exm239672	2	169764176	rs492594	C	G	0.4553

Table 4. *G6PC2* haplotype frequencies

rs560887	rs138726309	rs2232323	rs492594	FHS Frequency
C	C	A	C	0.46
T	C	A	G	0.29
C	C	A	G	0.24
T	C	C	G	0.006
C	T	A	C	< 0.001
T	C	A	C	< 0.001
C	T	A	G	< 0.001
C	C	C	G	< 0.001

Power Evaluation

The power of our novel haplotype meta-analysis approach is evaluated in a total of 16 scenarios (phenotype datasets) divided into 4 study designs (Study Design 1 to 4 from Table 2), with varying haplotype or SNV effects. For each scenario, we first compute the meta-analysis haplotype global test statistic, and then compare to meta-analysis of both single variant tests and gene-based tests. For single variant tests, we compute the meta-analysis test statsitic using inverse-variance-weighted method which has been shown to be the most powerful when the effect size is constant across cohorts [Zhou et al., 2011]. We then select the SNP with the minimum meta-analysis p-value $p_{min} = min_{\{1 \leq i \leq K\}} p_i$ ($K = 4$ for *G6PC2*; $K = 5$ for

JAZF1) and adjust the meta-analysis p-value for multiple testing using a Bonferroni correction for the effective number of independent variants [Gao et al., 2008]. We denote the result for the best SNP in the single variant analysis by "min P". For gene-based tests, we choose SKAT and Burden test with Wu weights and perform the analysis using R package seqMeta [Voorman et al., 2014]. We use $\alpha = 0.001$ to evaluate the power of all four approaches.

For each scenario, the phenotype is simulated using a multivariate normal distribution with mean $\hat{\mu}$ and a covariance matrix $\sigma_a^2\Phi + \sigma_e^2 I$, with $\sigma_a^2 = \sigma_e^2 = 0.5$, but unlike the type-I error scenarios, the value of $\hat{\mu}$ depends on genotypes/haplotypes in addition to the covariates of age and sex. We investigate 4 genetic effect scenarios: one or two causal genetic variants (gene-based scenarios), or one or two causal haplotypes. For the causal variant scenario, $\hat{\mu} = 0.02 \times \mathbf{age} + 0.5 \times \mathbf{sex} + b_{g_1} \times \mathbf{g_1} + b_{g_2} \times \mathbf{g_2}$ where $\mathbf{g_j}$ ($j = 1, 2$) is a vector containing the number of minor alleles (0, 1 or 2) carried by individuals in the sample, and b_{g_j} is the effect of variant j, set to $\sqrt{\frac{R^2}{2MAF_j(1-MAF_j)}}$ where MAF_j is the minor allele frequency of variant j and $R^2 = 0.01$ is the proportion of variance explained by this specific variant (haplotype). When only one causal variant is included in the model, $b_{g2} = 0$. For the causal haplotype models, $\hat{\mu} = 0.02 \times \mathbf{age} + 0.5 \times \mathbf{sex} + b_{h1} \times \mathbf{h_1} + b_{h2} \times \mathbf{h_2}$ where h_j is a vector containing the number of haplotypes j (0, 1 or 2) carried by individuals in the sample, and b_{hj} is the effect of haplotype j, set to $\sqrt{\frac{R^2}{\bar{h}_j(1-\bar{h}_j/2)}}$ where $\bar{h}_j$ is mean haplotype dosage of haplotype j and $R^2 = 0.01$. When only one causal haplotype is included in the model, $b_{h2} = 0$. For the *JAZF1* gene, we select two haplotypes, GTATA (the most frequent haplotype) and GCGCG (the third most frequent haplotype), to have an effect on the phenotype while all other haplotypes have no effect on the phenotype. For models with single variant effects, we select rs849134 and rs38523 to have non-zero effect on the trait while all other genetic variants have no effect. For the *G6PC2* gene, we select CCAC and TCAG, the two most frequent haplotypes to have an effect on the phenotype. For models with single variant effects, we select rs560887 and rs2232323 to have non-zero effect on the trait.

A thousand simulations with 5 independent cohorts are performed to compare the power of our approach to the single variant method adjusted for multiple testing and gene-based methods.

Table 5. *JAZF1* variants (chromosome 7)

Name	Position	dbSNPID	Minor	Major	MAF
exm-rs10486567	27976563	rs10486567	A	G	0.2415
exm2270592	28039797	rs38523	C	T	0.3683
exm-rs864745	28180556	rs864745	G	A	0.4965
exm-rs1635852	28189411	rs1635852	C	T	0.4973
exm-rs849134	28196222	rs849134	G	A	0.4917

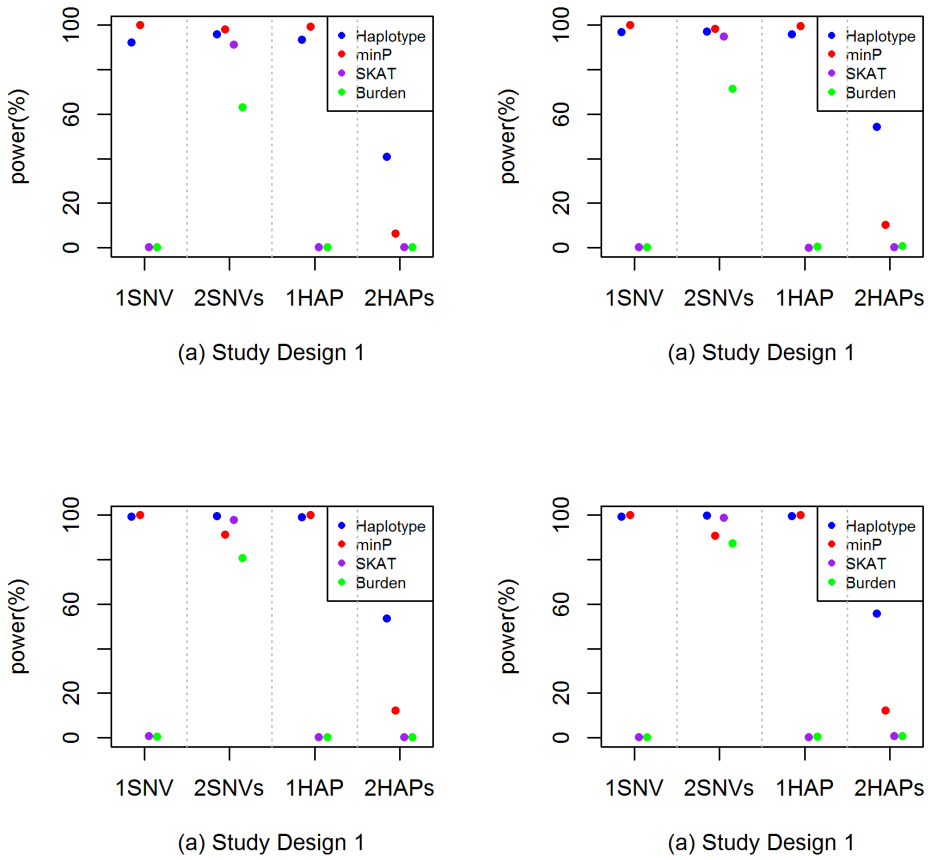

Figure 1. Power of the haplotype meta-analysis approach compared to gene-based methods and single SNV meta-analysis (min P) adjusted for multiple testing in the *G6PC2* region, evaluated at $\alpha = 0.001$ in four study designs. Description of the 4 study designs used in the simulation can be found in Table 2 (Study Design 1-4). The labels on the x axes denote that when 1 (SNV) or 2 (2SNVs) are influencing the phenotypes, or 1 (1HAP) or 2 (2HAPs) have an effect on the phenotypes.

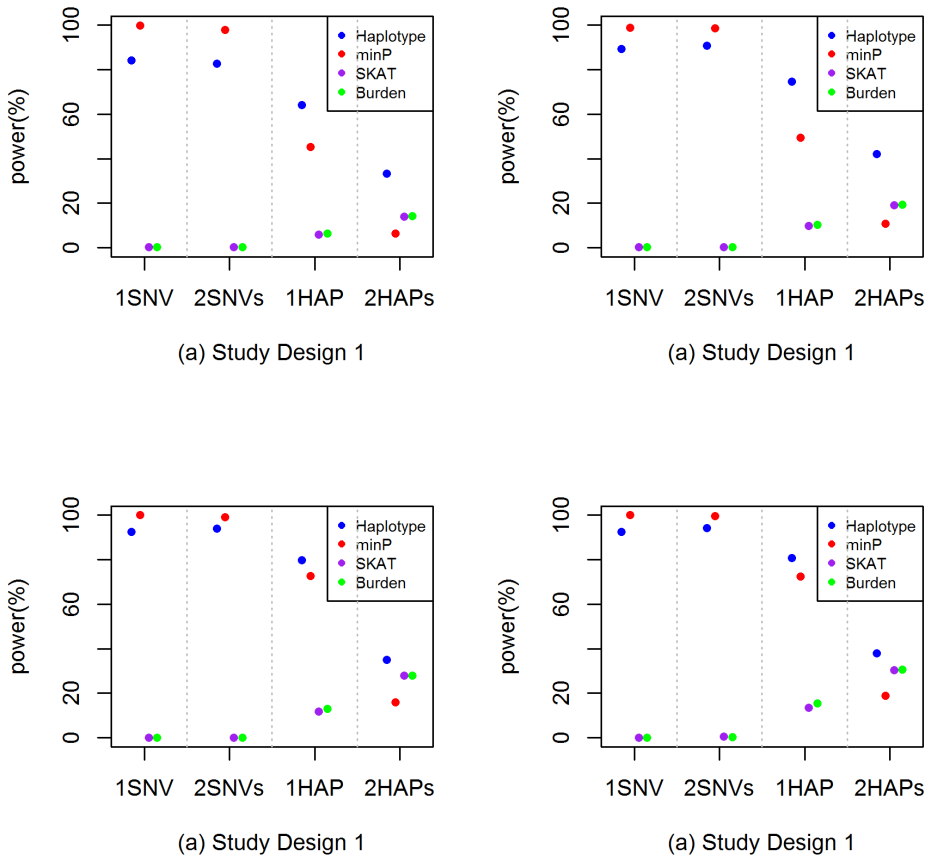

Figure 2. Power of the haplotype meta-analysis approach compared to gene-based methods and single SNV meta-analysis (min P) adjusted for multiple testing in the *JAZF1* region, evaluated at $\alpha = 0.001$ in four study designs. Description of the 4 study designs used in the simulation can be found in Table 2 (Study Design 1-4). The labels on the x axes denote that when 1 (SNV) or 2 (2SNVs) are influencing the phenotypes, or 1 (1HAP) or 2 (2HAPs) have an effect on the phenotypes.

Table 6. *JAZF1* haplotype frequencies

Haplotype	rs10486567	rs38523	rs864745	rs1635852	rs849134	Frequency
1	G	T	A	T	A	0.2327
2	G	T	G	C	G	0.2295
3	G	C	G	C	G	0.1608
4	G	C	A	T	A	0.1295
5	A	T	A	T	A	0.0866
6	A	T	G	C	G	0.0793
7	A	C	A	T	A	0.0434
8	A	C	G	C	G	0.0259
9	A	T	G	T	A	0.0029
10	A	T	A	C	A	0.0029
11	A	C	A	C	A	0.0023
12	G	T	A	C	A	0.0019
13	G	T	G	T	A	0.0017
14	G	C	G	T	A	0.0005

Results

Meta-analysis of 4 coding variants on G6PC2 region

G6PC2 is a known locus to affect FG level. Among the 17 exonic variants on the exome chip, 15 are rare variants (MAF<1%) and 2 are common variants (rs560887 with MAF=25.4%; rs492594 with MAF =43.7%). Previous GWAS have identified the A allele of rs560887, one of the two common variants to be associated with lower FG level ([Bouatia-Naji et al., 2008]: $\beta = -0.07 mmol/l$, $p = 6 \times 10^{-16}$; [Dupuis et al., 2010]: $\beta = 0.075 \pm 0.003 mmol/l$, $p = 8.5 \times 10^{-122}$). A recent large-scale exome-chip analysis indicated that these 15 rare variants also had a joint effect on FG [Wessel et al., 2015]. Our approach is applied to study the association between the haplotype structure of four coding variants rs560887, rs138726309, rs2232323 and rs492594 and FG, using CHARGE exome-chip data. We perform a meta-analysis of 7 studies comprising of 3 family-based and 4 population-based cohorts with up to 25305 non-diabetic European participants, to better understand how the overall haplotype structure as well as how the single haplotype affect FG level.

With a meta-analysis sample size of 25305, we have successfully replicated a previous reported haplotype analysis of 4 coding variants on G6PC2 region [Mahajan et al., 2015], but with higher precision (Table 7). Our effect size estimates are consistent with previously published estimates, in terms of both direction and magnitude. However, prior results were based on a single population-based cohort with 4442 participants. In contrast, our analysis is based on seven cohorts with over 25000 participants. Consequently we are able to estimate haplotype effect sizes with more precision, as reflected in the smaller standard errors. Among the 5 haplotypes shared by all 7 studies, one copy of the most significant haplotype, TCAG, decreases FG levels by 0.074 (95%CI: [0.063,0.085])mmol/l, on average; one copy of the second most significant haplotype, CCAG, increases the average FG levels by 0.039 (95% confidence interval (CI): [0.028,0.050])mmol/l; and one copy of the third most significant haplotype, TCCG, decreasing FG levels by an average of 0.12 (95% CI: [0.065,0.18])mmol/l. Most haplotype effect estimates reported in Mahajan et al.[Mahajan et al., 2015] fall within our 95 % CI, with the exception of estimates for TCCG (Mahajan et al's estimates = 0.205), which fall just outside our reported CI.

Simulations

Ten scenarios with increasing diversity in the study designs of the cohorts included in the meta-analysis are simulated to evaluate type-I error rate of our approach. The type-1 error is well controlled in all scenarios investigated (Table 1).

In the simulations to evaluate power, our approach is shown to be almost as powerful as the single SNV approach when SNVs are influencing the trait, but much more powerful to detect

Table 7. Single haplotype association test using 4SNVs on G6PC2 region

rs560887	rs138726309	rs2232323	rs492594	β (SE)	P-value	Frequency	$\beta_M(SE_M)^*$
C	C	A	C			0.4394	
T	C	A	G	-0.073 (0.0055)	4.56×10^{-41}	0.2671	-0.065(0.011)
C	C	A	G	0.039 (0.0056)	5.98×10^{-12}	0.2645	0.034(0.012)
T	C	C	G	-0.12 (0.029)	2.82×10^{-5}	0.0065	-0.205(0.057)
C	T	A	C	-0.022 (0.056)	0.70	0.0021	-0.202(0.077)
T	C	A	C	-0.031 (0.020)	0.12	0.0195	NA

The haplotypes are observed in all cohorts except that the last one is observed only in FHS, CHS, GS and FamHS.

*: β_M and SE_M denote the estimates from the paper of Mahajan et al.

true haplotype effects. For example, in the family based design scenarios, our approach is 40% more powerful than single SNV analyses when 2 haplotypes have non-zero effect on the phenotypes (Figure 1, 2). A similar pattern is observed for designs with a mix of unrelated and related samples. The gain in power is smaller when a single haplotype is influencing the trait, but present for all scenarios evaluated. When compared to the gene-based tests, our approach is uniformly more powerful in all scenarios across all study designs (Figure 3, 4) because of the Wu (default) weighing scheme that downweights common variants.

Discussion

We proposed a general meta-analysis approach to combine the haplotype association results from multiple cohorts. Our approach imposes no restrictions on the haplotypes observed across cohorts. Instead, our approach can incorporate information from haplotypes observed in a single cohort in addition to haplotypes observed in multiple cohorts. In the first stage of our approach, haplotype association analysis is performed at the cohort-level. Information about the haplotype structure, frequencies, effect estimates and covariance of effect estimates is collected, and meta-analyzed in a second stage using a generalized weighted least square approach. The association between a trait and any single or multiple haplotypes can be easily evaluated within our framework.

We evaluated the type-I error rate in a variety of scenarios with different cohort designs that included a mix of unrelated and family samples. Type-I error rate was controlled in all scenarios investigated. We also compared the power of our approach with single variant tests corrected for multiple testing (min P approach), and demonstrated that our approach had equivalent power when variants, but not haplotypes, influenced the trait, but was more powerful in the presence of true haplotype effects. Our haplotype approach also provided more evidence for association compared to gene-based tests applied with the default weighting scheme, as exemplified in a recent large-scale exome-chip project [Wessel et al., 2015] applied to the G6PC2 region comprising 15 rare variants (MAF$<$1%). Our simulations also illustrated that the haplotype effect size estimates obtained from meta-analysis were unbiased, even when family-based cohorts were included.

While our approach can not serve as the only tool for the discovery of associated variants and regions, it is a complementary tool to single-variant and gene-based tests. Mahajan et al. [Mahajan et al., 2015] demonstrated the usefulness of haplotype analysis in their investigation of the effect of *G6PC2* variants on fasting glucose. In 4442 non-diabetic subjects from the Oxford Biobank, the G allele from the coding variant rs492594 appears to significantly decrease fasting glucose levels. However, when conditioning on the variant with largest effect (rs560887) on fasting glucose, the effect estimates of the G-allele from rs492594 is reversed, and the G allele appears to decrease fasting glucose, an apparent paradox. However, looking at the haplotype estimates elucidates the mystery: the rs492594 G allele is most frequently observed on the same haplotype as the glucose raising allele (T) from the strongest associated variant (rs560887), giving the impression that the G allele also increases fasting glucose. Our analysis supports this conclusion, and refines the effect estimates provided by Mahajan et al. by increasing the number of samples used to obtain effect estimates via meta-analysis.

Our approach has some limitations. The variants included in the haplotype analysis must be genotyped or imputed in all cohorts. In other words, all cohorts must include the same set of variants in their analysis. Moreover, when using imputed genotypes, best-guess genotypes must be used because the approach does not currently handle genotypes in the form of dosage. The EM algorithm currently employed for inferring haplotypes works best for a moderate number of variants (< 15), and very rare haplotypes (frequency$< 0.1\%$) cannot be evaluated and must be collapsed to ensure computation stability. Despite these limitations, our approach has the potential to shed some light on the relationship between traits and multiple associated SNVs in a region.

Supporting Information

supporting figures and tables

Acknowledgement

Generation Scotland: Generation Scotland received core funding from the Chief Scientist Office of the Scottish Government Health Directorate CZD/16/6 and the Scottish Funding Council HR03006. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the UK's Medical Research Council. Ethics approval for the study was given by the NHS Tayside committee on research ethics (reference 05/S1401/89). We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses.
FamHS: Family Heart Study was supported by NIH grants RO1-HL-087700 and RO1-HL-088215 (Michael A. Province, PI) from NHLBI, and RO1-DK-8925601 and RO1-DK-075681 (Ingrid B. Borecki, PI) from NIDDK.
MESA: MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, and UL1-TR-000040. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Funding for MESA Family was provided by grants R01-HL-071051, R01-HL-071205, R01-HL-071250, R01-HL-071251, R01-HL-071252, R01-HL-071258, R01-HL-071259, and UL1-RR-025005. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.
FHS: Framingham Heart Study: Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). A portion of this research was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resources at Boston University Medical Campus. Also supported by National Institute for Diabetes and Digestive and Kidney Diseases (NIDDK) R01 DK078616, NIDDK K24 DK080140 and American Diabetes Association Mentor-Based Postdoctoral Fellowship Award #7-09-MN-32, all to Dr Meigs.
FENLAND: The Fenland Study is funded by the Medical Research Council (MC_U106179471) and Wellcome Trust. We are grateful to all the volunteers for their time and help, and to the General Practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and

Laboratory teams. The Fenland Study is funded by the Medical Research Council (MC_U106179471) and Wellcome Trust.
EPIC-Potsdam: We thank all EPIC-Potsdam participants for their invaluable contribution to the study. The study was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). The recruitment phase of the EPIC-Potsdam study was supported by the Federal Ministry of Science, Germany (01 EA 9401) and the European Union (SOC 95201408 05 F02). The follow-up of the EPIC-Potsdam study was supported by German Cancer Aid (70-2488-Ha I) and the European Community (SOC 98200769 05 F02). Furthermore, we thank Dr. Manuela Bergmann who was responsible for the methodological and organizational work of data collections of exposures and outcomes and Wolfgang Fleischhauer for his medical expertise that was employed in case ascertainment and contacts with the physicians and Ellen Kohlsdorf for data management.
CHS: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants HL080295, HL087652, HL103612, HL068986 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institute on Aging (NIA). A full list of CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.htm. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. A total of 3,525 / 5,888 subjects provided informed consent for participation in DNA studies, had available phenotype and genotype data and met the inclusion (European Ancestry)/exclusion criteria for this study.

Author Contribution

JD & SW conceived and designed the experiments. SW, JHZ, PA, JY, RAJ, JM, JEH and KM performed the experiments. SW, JHZ, PA, JY, RAJ, JM, JEH, KM analyzed the data. SW, JD, HB, AC, KMR, RAS, XG, MBS, NJW, IBB, MAP, JIR, CH, MOG, and JBM wrote the paper and provided critical input.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Stram, D. O. (1996). Meta-analysis of published data using a linear mixed-effects model. *Biometrics*, pages 536–544.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics*, 70(2):425–434.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J., and Ehm, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human heredity*, 53(2):79–91.

Tregouet, D., Escolano, S., Tiret, L., Mallet, A., and Golmard, J. (2004). A new algorithm for haplotype-based association analysis: the stochastic-em algorithm. *Annals of human genetics*, 68(2):165–177.

Becker, B. J. and Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, pages 414–429.

Bouatia-Naji, N., Rocheleau, G., Van Lommel, L., Lemaire, K., Schuit, F., Cavalcanti-Proença, C., Marchand, M., Hartikainen, A.-L., Sovio, U., De Graeve, F., et al. (2008). A polymorphism within the g6pc2 gene is associated with fasting plasma glucose levels. *Science*, 320(5879):1085–1088.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology*, 32(4):361–369.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785.

Liu, N., Zhang, K., and Zhao, H. (2008). Haplotype-association analysis. *Advances in genetics*, 60:335–405.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638–645.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523.

Zhou, B., Shi, J., and Whittemore, A. S. (2011). Optimal methods for meta-analysis of genome-wide association studies. *Genetic epidemiology*, 35(7):581–591.

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., Mägi, R., Strawbridge, R. J., Rehnberg, E., Gustafsson, S., et al. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*, 44(9):991–1005.

Therneau, T. (2012). *coxme: Mixed Effects Cox Models.* R package version 2.2-3.

Grove, M. L., Yu, B., Cochran, B. J., Haritunians, T., Bis, J. C., Taylor, K. D., Hansen, M., Borecki, I. B., Cupples, L. A., Fornage, M., et al. (2013). Best practices and joint calling of the humanexome beadchip: the charge consortium. *PLoS One*, 8(7):e68095.

Hu, Y.-J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E., and Lin, D.-Y. (2013). Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *The American Journal of Human Genetics*, 93(2):236–248.

Sinnwell, J. P. and Schaid, D. J. (2013). *haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous*. R package version 1.6.8.

Liu, D., 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(2):200–204.

Voorman, A., Brody, J., Chen, H., and Lumley, T. (2014). *seqMeta: Meta-Analysis of Region-Based Tests of Rare DNA Variants*. R package version 1.5.

Mahajan, A., Sim, X., Ng, H. J., Manning, A., Rivas, M. A., Highland, H. M., Locke, A. E., Grarup, N., Im, H. K., Cingolani, P., et al. (2015). Identification and functional characterization of g6pc2 coding variants influencing glycemic traits define an effector transcript at the g6pc2-abcb11 locus. *PLoS genetics*, 11(1):e1004876–e1004876.

Wessel, J., Chu, A. Y., Willems, S. M., Wang, S., Yaghootkar, H., Brody, J. A., Dauriz, M., Hivert, M.-F., Raghavan, S., Lipovich, L., et al. (2015). Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications*, 6.