PLOS ONE

RESEARCH ARTICLE

TRIO RVEMVS: A Bayesian framework for rare variant association analysis with expectation-maximization variable selection using family trio data

Duo Yu[1], **Matthew Koslovsky**[2], **Margaret C. Steiner**[3], **Kusha Mohammadi**[4], **Chenguang Zhang**[5], **Michael D. Swartz**[6]*

1 Division of Biostatistics, Data Science Institute, Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, **2** Department of Statistics, Colorado State University, Fort Collins, Colorado, United States of America, **3** Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **4** Department of Biostatistics and Data Management, Regeneron Pharmaceuticals, Inc., Tarrytown, New York, United States of America, **5** Biostatistics and Research Decision Sciences, Merck & Co., Inc., North Wales, Pennsylvania, United States of America, **6** Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America

* michael.d.swartz@uth.tmc.edu

OPEN ACCESS

Citation: Yu D, Koslovsky M, Steiner MC, Mohammadi K, Zhang C, Swartz MD (2024) TRIO RVEMVS: A Bayesian framework for rare variant association analysis with expectation-maximization variable selection using family trio data. PLoS ONE 19(12): e0314502. https://doi.org/10.1371/journal.pone.0314502

Editor: Priyadarshini Kachroo, Rutgers The State University of New Jersey, UNITED STATES OF AMERICA

Received: August 1, 2024

Accepted: November 11, 2024

Published: December 4, 2024

Copyright: © 2024 Yu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying this study cannot be shared publicly as they are owned by the National Institutes of Health (NIH) and are available through the Gabriella Miller Kids First Pediatric Research Program (https://commonfund.nih.gov/kidsfirst/overview). Access to the data requires an application process, and users are prohibited from sharing the data with others.

Abstract

It is commonly reported that rare variants may be more functionally related to complex diseases than common variants. However, individual rare variant association tests remain challenging due to low minor allele frequency in the available samples. This paper proposes an expectation maximization variable selection (EMVS) method to simultaneously detect common and rare variants at the individual variant level using family trio data. TRIO_RVEMVS was assessed in both large (1500 families) and small (350 families) datasets based on simulation. The performance of TRIO_RVEMVS was compared with gene-level kernel and burden association tests that use pedigree data (PedGene) and rare-variant extensions of the transmission disequilibrium test (RV-TDT). At the region level, TRIO_RVEMVS outperformed PedGene and RV-TDT when common variants were included. TRIO_RVEMVS performed competitively with PedGene and outperformed RV-TDT when the analysis was only restricted to rare variants. At the individual variants level, with 1,500 trios, the average true positive rate of individual rare variants that were polymorphic across 500 datasets was 12.20%, and the average false positive rate was 0.74%. In the datasets with 350 trios, the average true and false positive rates of individual rare variants were 13.10% and 1.30%, respectively. When applying TRIO_RVEMVS to real data from the Gabriella Miller Kids First Pediatric Research Program, it identified 3 rare variants in q24.21 and q24.22 associated with the risk of orofacial clefts in the Kids First European population.

Funding: This project was fully funded by the NIH/NICHD grant R03HD083674. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist. K.M. contributed to this article as an employee of the University of Texas Health Science Center and the views expressed no not necessarily represent the views of Regeneron Pharmaceuticals, Inc. This does not alter our adherence to PLOS ONE Polices on sharing data and materials.

Introduction

Birth defects are prevalent, occurring in 1 out of every 33 babies born annually in the United States, and are the primary cause of infant mortality, responsible for 20% of all infant deaths [1]. The impact of birth defects may be underestimated in mortality statistics [2], thus understanding the etiology of these major birth defects remains a research priority in birth defects epidemiology. Family data supports the hypothesis that a significant component of the risk for various birth defects stems from genetic variation [3–6]. Recent genome-wide association studies have identified common SNPs associated with varying birth defects, including obstructive heart defects (OHDs) [7], multiple congenital heart defect (CHD) phenotypes [8], conotruncal heart defects (CTD) [9], left-sided lesions (LSL) [10], and tetralogy of Fallot [11]. However, the identified variants only account for a small portion of the heritability. Part of the missing genetic heritability is thought to reside in rare variants, which are largely undetectable through genome-wide association platforms [12–17].

Although next-generation sequencing allows researchers to sequence each variant along the genome, there are statistical challenges inherent in identifying which rare variants are associated with disease. The power of traditional association methods for SNPs depends on allele frequency; the low frequency of minor alleles may reduce the power to analyze rare variants [18–22]. Previous methods are based on global tests that pool rare variants in a region to test the association with diseases. These global tests can be classified into burden tests (such as Cohort Allelic Sums Test (CAST) [23], the Collapsed Multivariate Collapsing Method (CMC) [24], and the Variable Threshold (VT) method [25, 26])), or quadratic tests (like C-alpha test [27] and the Sequence Kernel Association Test (SKAT) [28]).

Some methods have been developed that can identify specific rare variants that drive association within a given region of interest, as well as include common variants [24, 29–31]. However, these methods are based on a case-control design and are therefore subject to bias from population stratification [31–34]. Diseases that affect young children, like birth defects, are good candidates for family-based study designs, such as parent-child trios, which allow for more robust methods to analyze genetic data, including both common and rare variants, in the presence of population substructure [32, 35, 36]. Methods that take advantage of pedigree data for the genetic variant association, such as PedGene [37], and RV-TDT [38] have been developed. PedGene extends kernel and burden statistics for unrelated case-control data to include known pedigree relationships, which can account for the population-structured data [37]. Similarly, to avoid the spurious associations derived from the population-based method when the population substructure and admixture exist, RV-TDT extends commonly used population-based methods to analyze the association of rare variants in population-structured data, including aforementioned CMC [24] and VT [25]. However, none of these methods can detect individual rare variants within the region of interest. In this study, we proposed TRIO_RVEMVS, a Bayesian framework for individual rare variant association analysis with expectation-maximization variable selection using family trio data, which can simultaneously detect common and rare variants at the individual variant level.

The paper is organized as follows: We begin by constructing the likelihood of common and rare variants using case-parent trios. Next, we detail the Bayesian framework of TRIO_RVEMVS, which includes specifying priors for common and rare variants' coefficients, conducting posterior inference using the EM algorithm, and tuning selection parameters. To assess TRIO_RVEMVS, we perform simulations on both large (1500 case-trios) and small (350 case-trios) datasets and compare the results with those obtained using PedGene and RV-TDT. We then apply TRIO_RVEMVS to a real-world trio dataset from the Gabriella Miller Kids First Pediatric Research Program consisting of trios with a child suffering from cleft lip with or without cleft palate (CL+P). The paper concludes with a discussion of our findings.

Constructing the likelihood of common and rare variants using case-parents trios

In the case-parent trio design, small nuclear families are collected, where the child is affected by the disease or phenotype of interest. Then, the affected child and both parents are genotyped. Assuming each family can be phased, we denote the haplotype pair of the child [39]

$$g = (g_m, g_f)$$

where g_m and g_f denote the haplotypes inherited from the mother and father, respectively. Let D^+ represent the child is diseased, Θ denote the transmission parameters, and denote the parental haplotype pairs from mother and father as G_m and G_f, respectively. To model the sampling distribution of the case-trio family data, we propose to use the conditional logistic regression likelihood to model the probability of haplotype transmission from parents to the diseased child, which is motivated by the literature [39–41]. In more detail, the sampling distribution for observing a case trio can first be expressed as

$$P(g, G_m, G_f | \Theta, D^+) = P(g | G_m, G_f, \Theta, D^+) P(G_m, G_f | \Theta, D^+),$$

Due to Mendelian laws of inheritance, we can assume the transmission parameters contained in Θ are conditionally independent of the parents' genotypes G_m and G_f, given that the child is diseased. Since the trios are sampled through the diseased child, there is no information regarding Θ in the sampling distribution of the parent's haplotypes, which implies

$$P(G_m, G_f | \Theta, D^+) = P(G_m, G_f | D^+),$$

and

$$P(g, G_m, G_f | \Theta, D^+) \propto P(g | G_m, G_f, \Theta, D^+). \tag{1}$$

Eq (1) implies the sampling distribution will based on $P(g|G_m, G_f, \Theta, D^+)$, which can be generally modeled by a conditional logistic regression according to previous studies [39–41].

The conditional probability of disease given the haplotypes of parents can be derived similarly as in [40],

$$P(g | G_m, G_f, \Theta, D^+) = \frac{P(D^+ | g, \Theta)}{\sum_{j=1}^{4} P(D^+ | g_j, \Theta)}, \tag{2}$$

where g_j denotes one of the four different haplotype pairs inheritable from parents, $j = 1, 2, 3, 4$. We use a logistic regression modeling framework to include both common and rare variants in the sampling distribution as

$$\log\{\frac{P(D^+ | g, \Theta)}{1 - P(D^+ | g, \Theta)}\} = g_c \beta + g_r \alpha, \tag{3}$$

where g_c is a $1 \times S$ vector which denotes the common SNPs, β is a $S \times 1$ coefficient vector, g_r is a $1 \times L$ vector representing the rare SNPs, and α represents the effect of the selected rare variants on risk. For rare diseases, we can assume $1 - P(D^+|g, \Theta) \simeq 1$. Therefore, Eq (3) simplifies to

$$\log P(D^+|g, \Theta) = g_c\beta + g_r\alpha, \tag{4}$$

Similarly,

$$\log P(D^+|g_j, \Theta) = g_{cj}\beta + g_{rj}\alpha, \tag{5}$$

where g_{cj} and g_{rj} are the common and rare variants of g_j. With Eqs (4) and (5), finally, Eq (2) can be expressed as

$$P(g|G_m, G_f, \Theta, D^+) = \frac{exp(g_c\beta + g_r\alpha)}{\sum_{j=1}^{4} exp(g_{cj}\beta + g_{rj}\alpha)}. \tag{6}$$

Eq (6) can be understood as the likelihood for a 1:3 matched case-control design, where the affected child plays the role as the case and is matched with the 3 other possible genetic configurations of children that could have been offspring from the same parents. These other configurations are commonly called pseudo-siblings. Therefore, the conditional likelihood function for trios is given by

$$L(\mathbf{g}|\mathbf{G}_m, \mathbf{G}_f, \Theta, D^+) = \prod_{n=1}^{N} P(g_n|G_{nm}, G_{nf}, D^+, \Theta), \tag{7}$$

where $\mathbf{g}$ denotes the collection of haplotype pairs of children from the case-trio data, $\mathbf{G_m}$ and $\mathbf{G}_f$ are the collections of haplotype pairs of parents in the case-trio data, $n = 1, 2, \cdots, N$ are the indexes of families, and $P(g_n|G_{nm}, G_{nf}, D^+, \Theta)$ can be calculated through Eq (6).

Trio rare variants EMVS (TRIO_RVEMVS)

The sampling distribution and likelihood for trio data using common variants have been previously discussed [35, 39–42], modeling the probability of transmitting genes to the affected child. In the previous section, we show the probability of transmission can be extended to incorporate rare variants and be written as Eq (6). In summary, trio data are modeled using conditional logistic regression, similar to the model for 1:3 matched case-control data. Here, the affected child is matched with the 3 "pseudo-siblings" who could potentially be offspring of the parents. Using this sampling distribution as our likelihood, we proceed to outline the remaining mathematical details to develop a Bayesian framework for the variable selection for both common and rare variants associated with disease and implement the EM algorithm to compute the posterior quantities of interest [43].

Hierarchical priors

In this section, we construct the prior distributions to model inclusion in the risk model for each common and rare variant. For common variant selection, we used a single binary indicator, γ_s, to denote whether a common variant is included in the model. For rare variant selection, we implemented a dual selection indicator structure, which was first used in genetics for common variants with multiple alleles [42] and later modified for use in a Bayesian sparse group selection framework [44]. Specifically, we used a binary indicator, η_r, to indicate a group of related rare variants (typically those within the same gene) and a second indicator, λ_{rj}, to indicate individual variants within group r. Each selection indicator follows a Bernoulli distribution, with a prior inclusion probability parameter π_i ($i = 1, 2, 3$):

$$\begin{aligned} &\gamma_s|\pi_1 \sim \mathrm{Ber}(\pi_1), \quad s = 1, 2, \cdots, S, \\ &\eta_r|\pi_2 \sim \mathrm{Ber}(\pi_2), \quad r = 1, 2, \cdots, R, \\ &p(\lambda_{rj}|\eta_r, \pi_3) = \eta_r \pi_3^{\lambda_{rj}} (1 - \pi_3)^{1-\lambda_{rj}} + (1 - \eta_r)\delta_0, \quad j = 1, 2, \cdots, J_r, \end{aligned} \tag{8}$$

where δ_0 is point mass at zero, S denotes the total number of common variants, R denote the total number of regions, and J_r denotes the total number of rare variants in region r. If $\gamma_s = 1$, it indicates that a common variant is included in the model; if $\gamma_s = 0$, it indicates otherwise. Similarly, $\eta_r = 1$ indicates that a group of rare variants in a defined region (typically a gene) is included in the model; $\eta_r = 0$ indicates otherwise. $\lambda_{rj} = 1$ indicates an individual rare variant j at group r is included in the model; $\lambda_{rj} = 0$ indicates otherwise. To make the variable selection more flexible, we assume beta priors on π_i,

$$\pi_i \sim \mathrm{Beta}(a_i, b_i),$$

with hyper-parameters a_i and b_i, $i = 1, 2, 3$. This creates Beta-Bernoulli prior on all inclusion indicators. We use the hyper-parameters to balance power and multiplicity correction for the number of variants similar to [29]. Specifically, we use $a_i = 1 (i = 1, 2, 3)$, and set b_1 as the total number of common variants, b_2 as the total number of groups of rare variants, and b_3 as the total number of rare variants.

Conditional on the selection indicator, the prior for the coefficient of a common variant, β_s, is defined as a normal distribution,

$$p(\beta_s|\gamma_s) = N(0, d_s) \tag{9}$$

where d_s is defined by the corresponding γ_s:

$$d_s = \begin{cases} \nu_1 & if \quad \gamma_s = 1, \\ \nu_0 & if \quad \gamma_s = 0. \end{cases} \quad s = 1, 2, \cdots, S.$$

Here, we set ν_0 as a very small positive value that has the effect of restricting the value of β_s to be close to 0 when the SNP is not selected and ν_1 ($\nu_1 > 0$) large to allow the β_s coefficient to be estimated. Defining the prior on β_s in this way results in marginal normal mixture distributions that are defined by inclusion, typical of common Bayesian variable selection paradigms [29, 42, 43]. For each common variant, β_s is distributed as a normal distribution with the following mean and variance

$$\mu_s = 0, \text{ and } \mathrm{var}(\beta_s) = (1 - \gamma_s)\nu_0 + \gamma_s \nu_1.$$

For the rare variant coefficient $\alpha = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{rj}, \cdots)'$, we proposed:

$$p(\alpha_{rj}|\eta_r, \lambda_{rj}) \sim (1 - \eta_r\lambda_{rj})N(0, \nu_2) + \eta_r\lambda_{rj}N(0, \nu_3), \tag{10}$$

where r is the index of region, $r = 1, 2, \cdots, R$; and j is the index of individual rare variant in the region r, $j = 1, 2, \cdots, J_r$; η_r is the binary selection indicator of region r, λ_{rj} is the binary selection indicator of individual rare variant j in region r; ν_2 is the exclusion parameter of individual rare variant when either $\eta_r = 0$ or $\lambda_{rj} = 0$, ν_3 is the inclusion parameter of individual rare variant when both $\eta_r = 1$ and $\lambda_{rj} = 1$.

Posterior inference using the EM algorithm

Let γ, η, and λ denote the collections of binary selection indicators for individual common variants, regions, and individual rare variants, $\gamma = \{\gamma_s, s = 1, 2, \cdots, S\}$, $\eta = \{\eta_r, r = 1, 2, \cdots, R\}$, and $\lambda = \{\lambda_{rj}, r = 1, 2, \cdots, R; j = 1, 2, \cdots, J_r\}$. The full posterior distribution is denoted as $\log P(\beta, \alpha, \pi_1, \pi_2, \pi_3, \gamma, \eta, \lambda | g, G_m, G_f, D^+)$. Given the likelihood of Eq (7), and priors of Eqs (8) to (10) as defined above, the posterior distribution does not have a closed form. Instead of simulating large samples directly from the posterior using Markov Chain Monte Carlo (MCMC) methods, we employ the expectation maximization (EM) algorithm to estimate the posterior modes of interest. In the EM algorithm, we treat the variable selection indicators γ, η, and λ as missing data and alternate between conditional expectation using the current best estimates for the parameters and maximization of the expectation of the complete log-likelihood (Q function) to estimate the posterior modes of β and α [43].

For the E-step, we determine the Q-function which is the conditional expectation of log-likelihood of complete data with respect to the missing indicator variables, γ, η, and λ, given the current estimates of the unknown parameters $\beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}$, where k is the index of current iteration. For the M-step, we maximize the Q-function with respect to the parameters β, α, π_1, π_2, π_3 and iterate both steps until convergence. For iteration k, the Q-function is defined as:

$$\begin{aligned} &Q\left[\beta, \alpha, \pi_1, \pi_2, \pi_3 | \beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] \\ &= E_{\gamma,\eta,\lambda|\cdot}[\log L(\beta, \alpha, \pi_1, \pi_2, \pi_3, \gamma, \eta, \lambda | \mathbf{g}, \mathbf{G}_m, \mathbf{G}_f, D^+)] \end{aligned} \tag{11}$$

where $E_{\gamma,\eta,\lambda|\cdot} = E_{\gamma,\eta,\lambda|\beta^{(k)},\alpha^{(k)},\pi_1^{(k)},\pi_2^{(k)},\pi_3^{(k)},g,G_m,G_f,D^+}$, where the distributions of γ, η, λ are given by Eq (8), and $L(\beta, \alpha, \pi_1, \pi_2, \pi_3, \gamma, \eta, \lambda | \mathbf{g}, \mathbf{G}_m, \mathbf{G}_f, D^+)$ is the likelihood of complete data which is given by Eq (7).

E-step. The objective function Q can be simplified as the sum of conditional functions

$$\begin{aligned} Q(\cdot) &= C + Q_1\left[\beta, \alpha | \beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] + Q_2\left[\pi_1 | \beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}\right] \\ &+ Q_3\left[\pi_2 | \beta^{(k)}, \alpha^{(k)}, \pi_2^{(k)}\right] + Q_4\left[\pi_3 | \beta^{(k)}, \alpha^{(k)}, \pi_3^{(k)}\right], \end{aligned} \tag{12}$$

where C is constant term and each Q_1, Q_2, Q_3, Q_4 can be maximized independently. For convenience, we index the genotype of cases in each family as g_{0n}, and all the genotypes of pseudo-siblings are indexed by $i \in \{1, 2, 3\}$. Then, the common and rare SNPs of the case child from family n are denoted as g_{c0n} and g_{r0n}, respectively; and the common and rare SNPs of pseudo-siblings from family n are denoted as g_{cin} and g_{rin}, $i = 1, 2, 3$, respectively. In total of N families, according to the likelihood function Eqs (6) and (7), the Q-function with respect to β and α

can be written as

$$
\begin{aligned}
& Q_1\left[\beta, \alpha|\beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] \\
& = \sum_{n=1}^{N} \log\left[\frac{exp(g_{c0n}\beta + g_{r0n}\alpha)}{exp(g_{c0n}\beta + g_{r0n}\alpha) + \sum_{i=1}^{3} exp(g_{cin}\beta + g_{rin}\alpha)}\right] \\
& -\frac{1}{2}\sum_{s=1}^{S} \beta_s^2 E_{\gamma_s|\cdot}\left[\frac{1}{\nu_0(1-\gamma_s)+\nu_1\gamma_s}\right] \\
& -\frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{J_r} \alpha_{rj}^2 E_{\eta_r,\lambda_{rj}|\cdot}\left[\frac{1}{\nu_2(1-\eta_r\lambda_{rj})+\nu_3\eta_r\lambda_{rj}}\right] \\
& = -\sum_{n=1}^{N} \log(1+\sum_{i=1}^{3} e^{-x_{cin}\beta - x_{rin}\alpha}) - \frac{1}{2}\beta' P_c^{(k)}\beta - \frac{1}{2}\alpha' P_r^{(k)}\alpha
\end{aligned}
\tag{13}
$$

where $x_{in} = g_{0n} - g_{in}$ which is a $1 \times p$ vector; $P_c^{(k)}$ is a $S{\times}S$ diagonal matrix with elements $(1-p_s^{(k)})\frac{1}{\nu_0} + p_s^{(k)}\frac{1}{\nu_1}, s \in \{1, 2, \ldots, S\}$, and S is the total number of the common variants, $p_s^{(k)}$ is the conditional expectation of inclusion parameter. Based on Bayes' rule, $p_s^{(k)}$ can be calculated as follows.

$$
p_s^{(k)} = E_{\gamma_s|\cdot}[\gamma_s] = P(\gamma_s = 1|\beta^{(k)}, \pi_1^{(k)}) = \frac{a_s}{a_s + b_s}, \tag{14}
$$

where $a_s = P(\beta^{(k)}|\gamma_s = 1)P(\gamma_s = 1|\pi_1^{(k)})$, $b_s = P(\beta_s^{(k)}|\gamma_s = 0)P(\gamma_s = 0|\pi_1^{(k)})$ and $P(\gamma_s = 1|\pi_1^{(k)}) = \pi_1^{(k)}$. $P_r^{(k)}$ is a $L \times L$ diagonal matrix with elements $(1-q_{rj}^{(k)})\frac{1}{\nu_2} + q_{rj}^{(k)}\frac{1}{\nu_3}$, $r = 1, 2, \cdots, R$, $j = 1, 2, \cdots, J_r$, L is the total number of rare variants, R is the number of groups (regions) of variants, J_r is the number of rare variants in the group (region) r, and

$$
q_{rj}^{(k)} = E_{\lambda_{rj=1}|\eta_r=1,\cdots}[\lambda_{rj}] = P(\lambda_{rj} = 1|\eta_r = 1, \alpha_{rj}^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}) = \frac{c_{rj}}{c_{rj} + d_{rj}}, \tag{15}
$$

where $c_{rj} = \pi_3^{(k)} P(\alpha_{rj}^{(k)}|\lambda_{rj} = 1, \eta_r = 1)$, and $d_{rj} = (1-\pi_3^{(k)})P(\alpha_{rj}^{(k)}|\lambda_{rj} = 0, \eta_r = 1)$. The second and third terms of the Q function can be calculated as

$$
Q_2\left[\pi_1|\beta^{(k)}, \pi_1^{(k)}\right] = \sum_{s=1}^{S} E_{\gamma_s|\cdot}[\gamma_s]\log\left[\frac{\pi_1}{1-\pi_1}\right] + (2S-1)\log(1-\pi_1). \tag{16}
$$

$$
Q_3\left[\pi_2|\alpha^{(k)}, \pi_2^{(k)}\right] = \sum_{r=1}^{R} E_{\eta_r|\cdot}[\eta_r]\log\left[\frac{\pi_2}{1-\pi_2}\right] + (2R-1)\log(1-\pi_2), \tag{17}
$$

where

$$
E_{\eta_r|\cdot}[\eta_r] = P(\eta_r = 1|\alpha_{r1}^{(k)}, \cdots, \alpha_{rJ_r}^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}) = \frac{e_r}{e_r + f_r} \doteq g_r^{(k)}, \tag{18}
$$

in which e_r and f_r are defined as

$$
e_r = \pi_2^{(k)} \prod_{j=1}^{J_r}\left[\pi_3^{(k)} P(\alpha_{rj}^{(k)}|\lambda_{rj} = 1, \eta_r = 1) + (1-\pi_3^{(k)})P(\alpha_{rj}^{(k)}|\lambda_{rj} = 0, \eta_r = 1)\right],
$$

and

$$f_r = (1 - \pi_2^{(k)}) \prod_{j=1}^{J_r} P(\alpha_{rj}^{(k)} | \lambda_{rj} = 0, \eta_r = 0).$$

The last term of the Q function is written as

$$\begin{aligned} Q_4\left[\pi_3 | \alpha^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] &= E_{\eta,\lambda|\cdot} \log\left[P(\pi_3) \prod_{r=1}^{R} \prod_{j=1}^{J_r} P(\lambda_{rj} | \eta_r, \pi_3)\right] \\ &= \sum_{r=1}^{R} g_r^{(k)} \sum_{j=1}^{J_r} q_{rj}^{(k)} \log\left(\frac{\pi_3}{1 - \pi_3}\right) + (L + \sum_{r=1}^{R} g_r^{(k)} J_r - 1) \log(1 - \pi_3) \end{aligned} \tag{19}$$

M-step. For the M-step, we maximize Q_1, Q_2, Q_3 and Q_4 separately. There is no closed-form solution for Q_1 function. However, maximizing the Q_1 with respect to β and α is equivalent to a minimization problem with respect to parameter ω ($p \times 1$), where $p = S + L$, S is the total number of common variants, and L is the total number of rare variants. Based on Eq (13),

$$\begin{aligned} &\max_{\beta,\alpha} Q_1\left[\beta, \alpha | \beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] \\ &= \max_{\beta,\alpha}\left(-\sum_{n=1}^{N} \log(1 + \sum_{i=1}^{3} e^{-x_{cin}\beta - x_{rin}\alpha}) - \frac{1}{2}\beta' P_c^{(k)} \beta - \frac{1}{2}\alpha' P_r^{(k)} \alpha\right) \\ &= \min_{\beta,\alpha}\left(\sum_{n=1}^{N} \log(1 + \sum_{i=1}^{3} e^{-x_{cin}\beta - x_{rin}\alpha}) + \frac{1}{2}\beta' P_c^{(k)} \beta + \frac{1}{2}\alpha' P_r^{(k)} \alpha\right) \\ &= \min_{\omega}\left(\sum_{n=1}^{N} \log(1 + \sum_{i=1}^{3} e^{-x_{in}\omega}) + \frac{1}{2}\omega' P^{(k)} \omega\right), \end{aligned} \tag{20}$$

where $x_{in} = (x_{cin}, x_{rin})$, $\omega = (\beta_1, \beta_2, \ldots, \beta_S, \alpha_{11}, \alpha_{12}, \ldots, \alpha_{rj}, \ldots, \alpha_{RJ_R})'$, $P^{(k)}$ is a $p \times p$ ($p = S + L$) diagonal matrix with diagonal elements $(1 - p_s^{(k)})\frac{1}{v_0} + p_s^{(k)}\frac{1}{v_1}, s \in \{1, 2, \ldots, S\})$ for the first S elements, and $(1 - q_{rj}^{(k)})\frac{1}{v_2} + q_{rj}^{(k)}\frac{1}{v_3}$ for the rest of L elements.

Since the likelihood function of conditional logistic regression and $\frac{1}{2}\omega' P^{(k)} \omega$ are vector convex functions [43], we used stochastic dual coordinate ascent (SDCA) [45, 46], an efficient technique for solving regularized loss minimization problems in machine learning, to solve the minimization problem above. Particularly, the accelerated min-batch SDCA was implemented and the details of the algorithm can be found in the Supplemental Materials [46]. Accordingly, we compute the β and α estimates for the next iteration based on the Q_1 function. The remaining components Q_2, Q_3, and Q_4 have closed forms. The details of solving the closed form solutions to the maximization of Q_2, Q_3, and Q_4 are shown in the Supplemental Materials. The closed form solution for Q_2 is:

$$\pi_1^{(k+1)} = \frac{\sum_{s=1}^{S} p_s^{(k)}}{2S - 1} \tag{21}$$

The closed form solution for Q_3 is:

$$\pi_2^{(k+1)} = \frac{\sum_{r=1}^{R} g_r^{(k)}}{2R - 1} \tag{22}$$

The closed form solution for Q_4 is:

$$\pi_3^{(k+1)} = \frac{\sum_{r=1}^{R} g_r^{(k)} \sum_{j=1}^{J_r} q_{rj}^{(k)}}{L + \sum_{r=1}^{R} g_r^{(k)} J_r - 1} \tag{23}$$

Convergence of the algorithm is determined if the difference between two successive observed-data likelihood is less than ϵ, i.e.

$$\begin{aligned} & l(\theta^{(k+1)}|\mathbf{y}_{obs}) - (\theta^{(k)}|\mathbf{y}_{obs}) \\ & = \left[Q(\theta^{(k+1)}, \theta^{(k)}) - Q(\theta^{(k)}, \theta^{(k)})\right] - \left[R(\theta^{(k+1)}, \theta^{(k)}) - R(\theta^{(k)}, \theta^{(k)})\right] < \epsilon, \end{aligned} \tag{24}$$

where ϵ is a pre-specified threshold, $\theta = (\beta, \alpha, \pi_1, \pi_2, \pi_3)$, and

$$\begin{aligned} & R\left[\beta, \alpha, \pi_1, \pi_2, \pi_3 | \beta^{(k)}, \alpha^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}\right] \\ & = E_{\gamma,\eta,\lambda|.}[\log P(\gamma, \eta, \lambda | g, G_m, G_f, D^+, \beta, \alpha, \pi_1, \pi_2, \pi_3)] \\ & = \sum_{s=1}^{S} E_{\gamma_s|.}[\gamma_s] \log\left[\frac{\pi_1}{1-\pi_1}\right] + S\log(1-\pi_1) \\ & + \sum_{r=1}^{R} E_{\eta_r|.} \eta_r \log\left[\frac{\pi_2}{1-\pi_2}\right] + R\log(1-\pi_2) \\ & + \sum_{r=1}^{R} g_r^{(k)} \sum_{j=1}^{J_r} q_{rj}^{(k)} \log\left(\frac{\pi_3}{1-\pi_3}\right) + (\sum_{r=1}^{R} g_r^{(k)} J_r)\log(1-\pi_3). \end{aligned} \tag{25}$$

Deterministic annealing

Though the conventional EM algorithm has attractive features, it can become trapped in local maximums in multimodal posterior distributions. To enhance the chance of discovering a global mode, the Deterministic Annealing variant of the EM algorithm (DAEM) is considered [47]. During each DAEM iteration, the conditional probability of inclusion indicators is parameterized by temperature $1/t$ $(0 < t < 1)$. Therefore, $p_s^{(k)}$, $q_{rj}^{(k)}$ and $g_r^{(k)}$ from Eqs (14), (15) and (18) were substituted by:

$$p_s^{(k)} = \frac{a_s^t}{a_s^t + b_s^t} \tag{26}$$

$$q_{rj}^{(k)} = \frac{c_{rj}^t}{c_{rj}^t + d_{rj}^t} \tag{27}$$

$$g_r^{(k)} = \frac{e_r^t}{e_r^t + f_r^t} \tag{28}$$

where a_s, b_s, c_{rj}, d_{rj}, e_r and f_r are the same as in Eqs (14), (15) and (18). In this study, for both simulation and real-data analysis, the initial value of coefficients of common and rare variants is 0.5; the initial value of t is 0.1 with an incremental value of 0.1.

Selection parameter tuning

Here, we present a recommendation for tuning the selection parameters in TRIO_RVEMVS. Specifically, selection is controlled through the exclusion parameters, v_0 and v_2, and inclusion parameters, v_1 and v_3, as similar to [43]. To determine suitable values for these parameters, we first considered an odds ratio between [0.95,1.05] to be clinically irrelevant. Then, given a 95% prior probability of variable inclusion of an odds ratio that covers [0.29,3.45] and [0.27,4.3] for common and rare variants respectively. Thus, we set $v_1 = 0.4$ and $v_3 = 0.5$. Next, we evaluated the local stability of regularization plots with respect to exclusion parameters. The tuning process proceeded as follows:

- We initially set the exclusion parameters for common and rare variants to be equal, i.e. $v_0 = v_2$, and chose the common exclusion parameter in the local stable regularization plot window.
- Subsequently, with the common exclusion parameter fixed, we evaluated the local stability of the regularization plot with respect to the rare variant exclusion parameter v_2, and chose the v_2 within a stable window defined by at least 3 points in the grid where no shrinkage occurs.

This procedure was applied in both simulated and real data analyses.

Simulation

We simulated two scenarios consisting of 500 datasets each to evaluate the performance of TRIO_RVEMVS in identifying regions/genes of interest relative to existing methods (PedGene [37] and RV-TDT [38]) as well as its ability to identify individual variants. One scenario generates 1500 case-parent trios per data set, while the other involves generating 350 case-parent trios per dataset. These scenarios allowed us to assess the performance of TRIO_RVEMVS under large-sample and small-sample conditions, respectively.

To measure the overall performance of region detection, we calculated the weighted average correct association percentage with

$$\frac{1}{2}\left[\frac{\sum P(\text{selected}|\text{associated})}{\text{total number of associated regions}}+\frac{\sum P(\text{unselected}|\text{unassociated})}{\text{total number of unassociated regions}}\right] \tag{29}$$

Considering that most of the rare variants were not polymorphic across all data sets, we defined the Average True and False Positive Rate (ATPR and AFPR) for individual variants as follows:

$$\begin{aligned}\text{ATPR} &= \frac{1}{\#\text{ of data sets}}\sum_{\text{data set } d}\frac{N_d(\text{selected}|\text{associated})}{N_d(\#\text{ of polymorphic associated variants})}\\ \text{AFPR} &= \frac{1}{\#\text{ of data sets}}\sum_{\text{data set } d}\frac{N_d(\text{selected}|\text{unassociated})}{N_d(\#\text{ of polymorphic unassociated variants})}\end{aligned} \tag{30}$$

where $N_d(\text{selected}|\cdot)$ denotes the number of detected variants given the variants are associated or unassociated in data set d.

Data simulation

We simulated the population of haplotypes using Cosi2 [48], which is a forward-time genetic simulator. We used the 1000 Genomes Project [49] haplotypes as the reference population for Cosi2 and simulated a 30kb region of chromosome 1, consisting of 45965 SNPs. We simulated populations of 80,000 African haplotypes and 80,000 European haplotypes. Then we constructed 500 samples, each consisting of 60,000 haplotypes (15,000 African and 45,000 European, reflecting 25% and 75%, respectively) from the simulated population (with replacement). For each of the 500 samples of haplotypes, we randomly selected and paired haplotypes within the race to construct individuals, and then randomly paired individuals within the race to construct parents. Then from each set of parents, we randomly selected one haplotype from each parent to be transmitted to the child to form 15,000 trios. The full simulation algorithm is described in the Supplemental Materials.

We defined a gene region as 2700 base pairs, resulting in 12 simulated gene regions. Within gene regions, we used population allele frequencies to determine rare and common variants. Since we simulated admixed samples, we computed a weighted minor allele frequency (MAF) for each SNP. This was based on the frequency estimates from our reference genome (the 1000 Genomes Project) for each population (African and European), weighted by the proportion of each population admixed for our simulation (25% African and 75% European). Rare variants were defined as variants with a weighted MAF < 0.05. For simplicity, we simulated our causal SNPs on genes 1–6. We modeled disease based on two causal common SNPs (risk-increasing allele on gene 3 and risk-decreasing allele on gene 6) and 5% of any variant with weighted MAF less than 3% were randomly chosen as causal rare variants. In total, 1212 rare variants were simulated as causal to the simulated disease, with 606 simulated as risk increasing and 606 as risk decreasing. The distribution of associated rare variants across the 6 genes is reported in Table 1.

We simulated the disease status of each child according to the logistic model:

$$\text{logit}(P(y = 1)) = \alpha_0 + \beta_1 G_1^c + \beta_2 G_2^c + \cdots + \beta_p G_p^c, \tag{31}$$

where $G_1^c, G_2^c, \ldots, G_p^c$ were the children's genotypes for the p causal variants (consisting of the 2 common variants, and the rest are rare variants), and we set $\alpha_0 = -2.2$ to control the disease incidence to be low. For the simulation, we set the magnitude of the coefficients for our causal common variants to be 0.9, and the magnitude of the coefficients for causal rare variants was computed as in [28]: $c|log_{10}MAF_i|$, where $c = 0.4$ for causal risk rare variants, and $c = -0.4$ for causal protective rare variants, and MAF_i is the weighted MAF of locus i. Assigning the coefficients in this way results in rarer variants having a stronger effect on disease.

Table 1. Number of associated rare variants with different range of weighted MAF by region in the haplotype pool. Across all simulated data sets, most of the variants with weighted MAF less than 0.0001 were those non-polymorphic, singleton, doubletons, or triptons.

Region	(0, 0.0001)	[0.0001, 0.001)	[0.001, 0.01)	[0.01, 0.05)
1	204	4	3	1
2	172	2	0	0
3	204	3	2	0
4	220	5	0	0
5	197	2	1	0
6	185	6	1	0
Total	1182	22	7	1

https://doi.org/10.1371/journal.pone.0314502.t001

Table 2. Summary statistics of data sets with 1,500 and 350 case-trios.

Summary statistics	1500 case-trios	350 case-trios
Average # polymorphic variants	4008	1442
Average # polymorphic common variants	48	47
Average # polymorphic causal common variants	2	2
Average # polymorphic rare variants	3960	1395
Average # polymorphic causal rare variants	132	45
# polymorphic variants across 500 data sets	431	133
# polymorphic causal variants across 500 data sets	12	4

https://doi.org/10.1371/journal.pone.0314502.t002

The distribution of associated rare variants across 6 regions is shown in Table 1. After simulating diseased probands in each set of 15,000 trios, we generated 500 replicates of 1,500 case-parent trios, and another 500 replicates of 350 case-parent trios to assess performance for large and small sample sizes, respectively.

Due to sampling, some loci that were polymorphic in the population with rare variants became non-polymorphic in the 500 samples, and the number of polymorphic loci varied across the 500 data sets. Some of the polymorphic loci with rare variants that were in the disease-generating model were not polymorphic in one or more samples. The distribution of polymorphic variants across the 500 data sets for each sample size is depicted in Table 2.

Analysis of simulated data

We used the same analysis strategy across all simulated datasets. First, we classified variants as either common or rare in each simulated data set. We estimated the MAF of each variant in each dataset for each sample size. For each sample size, we defined rare variants based on the median MAF across the 500 data sets, with the threshold of rare variants being median MAF $< 5\%$. To assess performance in identifying regions of interest, we applied all three methods (TRIO_RVEMVS, PedGene, and RV-TDT) to the 500 data sets, summarizing the performance of each method using the weighted average correct association metric. Since only TRIO_RVEMVS identifies specific rare variants, we also report a similar weighted correct association metric for individual rare variants. For TRIO_RVEMVS, the detailed exclusion parameter tuning using regularization plot can be found in the Supplemental Materials.

Simulation results

In this section, we first compare selection at the region level with PedGene [37] and RV-TDT [38] with the weighted average correct association percentage. Specifically, we compared TRIO_RVEMVS with PedGene kernel and burden methods [37]. For RV-TDT, we evaluated different variants, such as BRV-Haplo, VT-BRV-Haplo, WSS-Haplo, CMC-Analytical, CMC-Haplo, and VT-CMC-Haplo [38]. TRIO_RVEMVS outperformed both PedGene and RV-TDT when jointly considering common and rare variants. Our simulation analyses also confirmed that PedGene showed improved performance in detecting rare variants compared to RV-TDT [50]. We conclude our simulation analysis by discussing the capacity of TRIO_RVEMVS to detect individual rare variants, which is not accomplished by either PedGene or RV-TDT.

We compared the performance of all methods in two ways. First, we compared each method's ability to select risk regions based on rare variants only. Second, we compared each method's ability to identify risk regions when jointly analyzing rare and common variants. For all

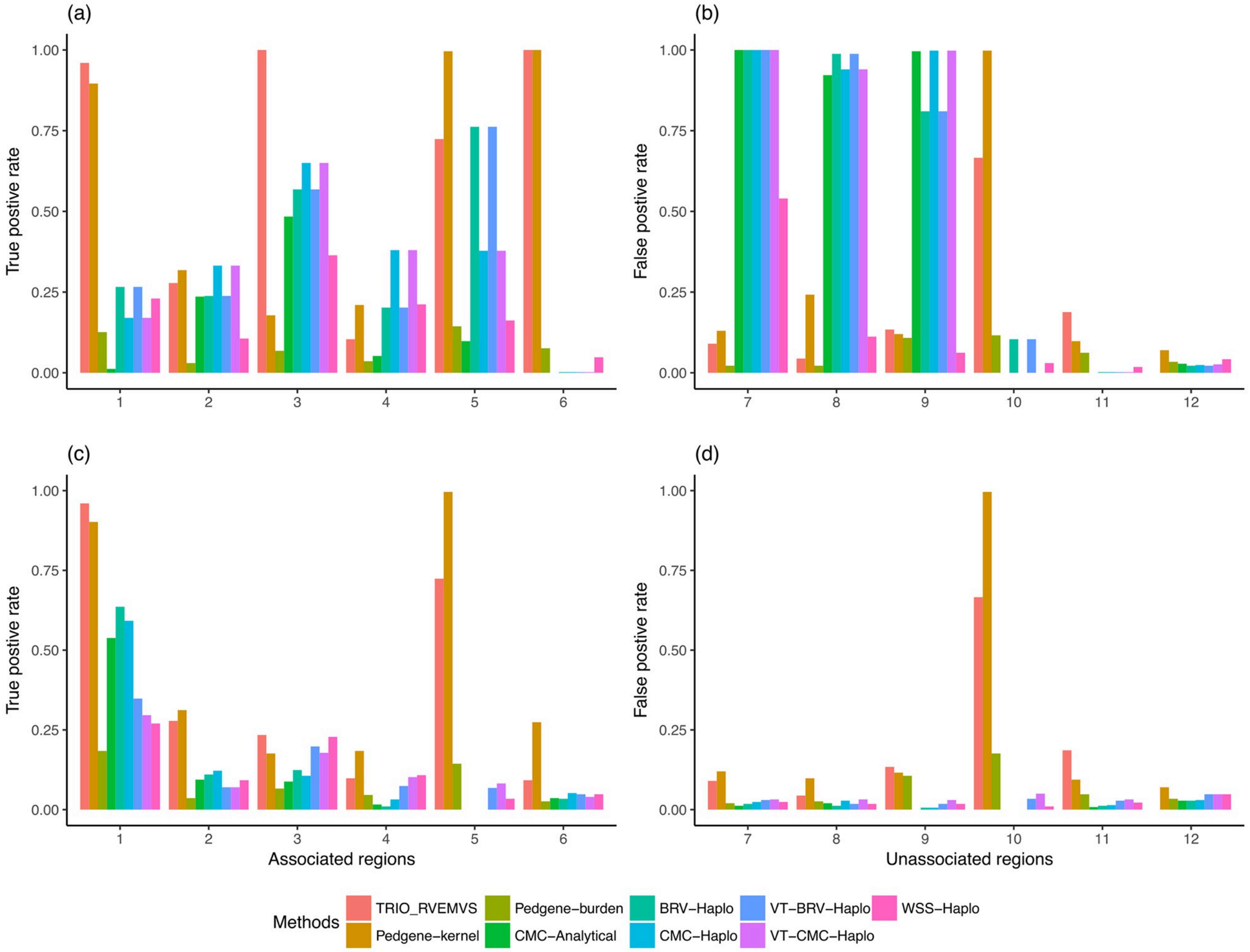

Fig 1. For the data sets with 1,500 case-trios, panels (a) and (b) showed the true and false positive rates of analyzing regions using both common and rare variants; panels (c) and (d) showed the true and false positive rates of regions with rare variants only.

https://doi.org/10.1371/journal.pone.0314502.g001

analyses, we defined rare variants as SNPs whose MAF < 5%. Panels a) and b) of Fig 1 show the true and false positive rate of region selection when jointly analyzing common and rare variants using datasets with 1500 case-trios. In panel a), it shows that TRIO_RVEMVS outperformed both PedGene and RV_TDT with higher true positive rates across the 6 simulated causal regions except for simulated region 5, where PedGene shows a better true positive rate of detection. Considering the false positive rates across regions 7–12, PedGene and Trio_RVEMVS were competitive, TRIO_RVEMVS outperformed PedGene's false positive rate across regions 7, 8,10, and 12, while PedGene had slightly lower false positive rates for regions 9 and 11. Panels c) and d) of Fig 1 show the true and false positive rates when focusing solely on rare variants detection for data sets with 1500 case-trios. TRIO_RVEMVS outperformed PedGene in regions 1 and 3 in terms of true positive rate but performed just behind PedGene in regions 2, 4, 5, and 6.

Table 3 summarizes the weighted average correct association percentage (WACAP), Eq (29), for TRIO_RVEMVS, RV-TDT, and PedGene. TRIO_RVEMVS shows the highest WACAP when selecting both common and rare variants. When focusing on using rare

Table 3. The comparison of weighted average correct association percentage between TRIO_RVEMVS, PedGene and RV-TDT with and without common variants.

Methods	1500 case-trios		350 case-trios	
	common and rare	rare only	common and rare	rare only
TRIO_RVEMVS	74.53	60.55	66.37	52.07
PedGene-kernel	66.17	61.25	55.73	52.65
PedGene-burden	50.97	50.77	49.90	49.70
CMC-Analytical	32.80	55.86	36.32	52.35
BRV-Haplo	42.60	56.98	40.58	53.08
CMC-Haplo	41.23	56.68	39.00	52.95
VT-BRV-Haplo	42.60	55.25	52.73	52.65
VT-CMC-Haplo	41.21	54.53	57.33	52.10
WSS-Haplo	52.65	55.33	50.18	52.75

https://doi.org/10.1371/journal.pone.0314502.t003

variants only, TRIO_RVEMVS was competitive with PedGene-Kernel, but did not always have the highest WACAP.

In the 350 case-trio data sets, TRIO_RVEMVS outperformed both PedGene and RV_TDT with respect to the weighted average correct association percentage, shown in Table 3. TRIO_RVEMVS achieved the highest average correct association percentage among all methods at 66.37%. VT-CMC-Haplo had the second-highest WACAP at 57.33%. In each region, we observed a consistent pattern of true and false positive rates when analyzing both common and rare variants, Fig 2. Specifically, TRIO_RVEMVS exhibited superior performance in terms of the true positive rate in regions 1, 3, and 6, shown in panel a) in Fig 2. With respect to the false positive rate, TRIO_RVEMVS performed similarly to PedGene; and both have smaller false positive rates compared to RV_TDT. See panel b) in Fig 2. When analyzing only rare variants, all methods demonstrated similar average correct association percentages, as indicated in Table 3. Because of the smaller sample size, all methods showed low power to detect the causal rare variants in general, shown in panels c) and d) in Fig 2. Therefore, compared to the 1500 case-trios, we did not observe notably higher true positive rates or false positive rates at the region level when analyzing only rare variants in the 350 case-trios data analysis.

At the individual variants level, TRIO_RVEMVS detected 94 variants including 8 causal in 500 datasets with 1500 case-trios, and 57 variants with 4 causal in 500 datasets with 350 case-trios. The true positive rate (TPR) and false positive rate (FPR) of detected variants are shown in Figs 3 and 4. We primarily focus on reporting the individual-level selection results for variants that were polymorphic across all datasets, due to the low MAF of rare variants, Table 4. When considering the variants that were not all polymorphic across datasets, individual-level selection results were summarized in the Supplemental Materials. For datasets with 1500 case-trios, the ATPRs were 26.83% and 12.20% with and without common variants. The two causal common variants were constantly detected with ATPR of 100%. The AFPRs were 0.67% and 0.74% with and without common variants. For datasets with 350 case-trios, when the common and rare variants were jointly analyzed the ATPR was 48.45%, and the AFPR was 1.38%; when only rare variants were analyzed the ATPR was 13.10% and AFPR was 1.30%. ATPR and AFPR for variants in different ranges of MAF were summarized in Table 4.

We observe that the true positive rate is lower and the false positive rate is higher for variants that have lower MAF. We illustrate this using the dataset of 1500 case-trios. For variants with a median MAF less than 0.01, the ATPR and AFPR were 3.08% and 0.09%, respectively. The highest FPR in this group was 8.8%, and the variant with such a high FPR had an equal median MAF to one of the associated variants (0.006) in region 1. In addition, the associated

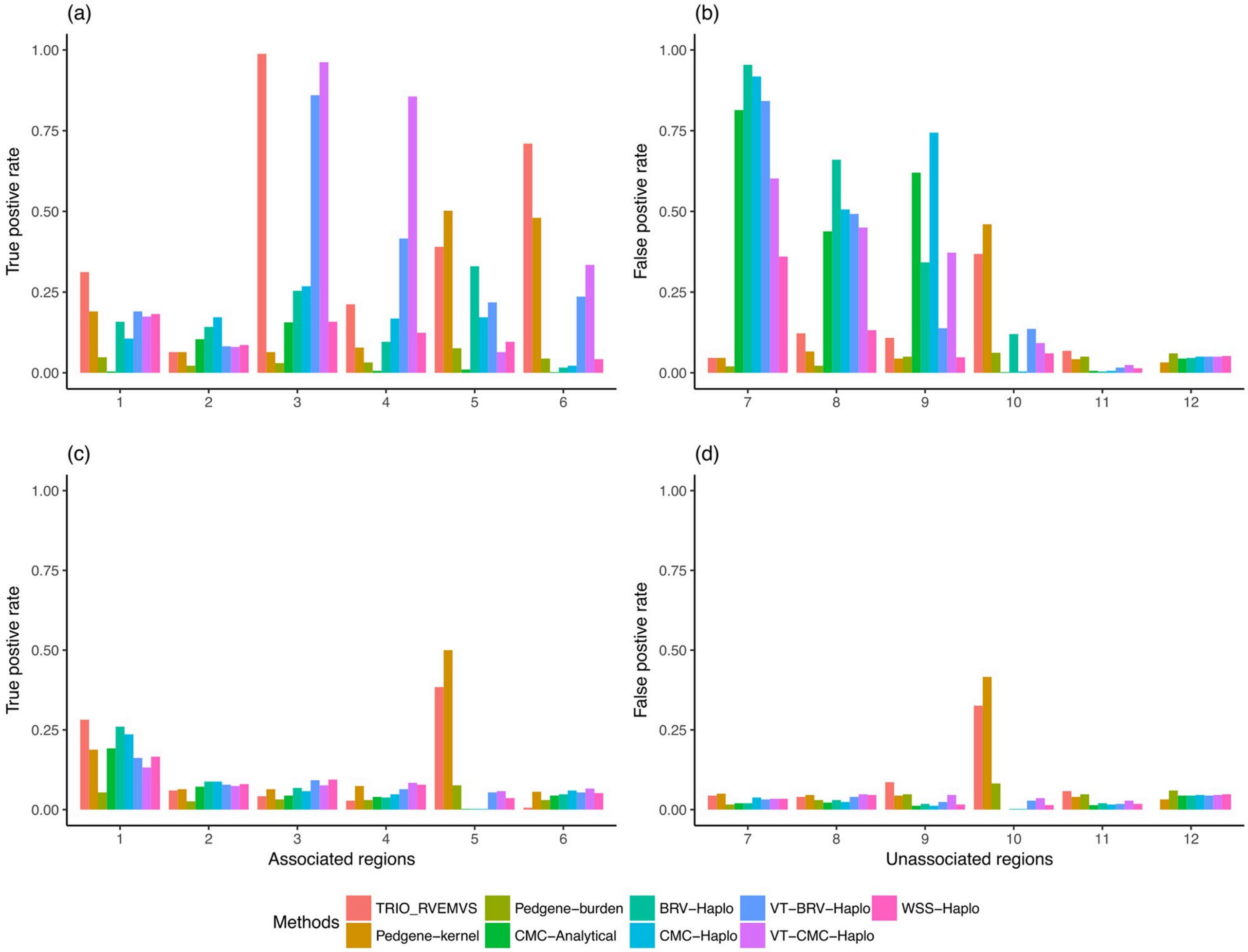

Fig 2. For the data sets with 350 case-trios, panels (a) and (b) showed the true and false positive rates of regions with common variants; panels (c) and (d) showed the true and false positive rates of regions with rare variants only.

https://doi.org/10.1371/journal.pone.0314502.g002

variant in region 1 and this falsely detected unassociated variant was only separated by 53 base-pairs, and the linkage disequilibrium between them was 1 (both Dprime and rSquare), Fig 5. The average FPR in the group of variants that had MAF in the range of [0.01, 0.05) was 6.08%. The TPR of the only associated variant (median MAF: 0.026) was 94.2%. Two variants in the same group have relatively high FPR: one rare variant from region 5 had an FPR of 71.8%; another rare variant from region 10 had an FPR of 62.8%. (Those three variants may have contributed the high true and false positive rates, respectively, at the region level for both methods TRIO_RVEMVS and PedGene, Fig 1).

Real data application

We applied TRIO_RVEMVS to a trio data set from the Gabriella Miller Kids First Pediatric Research Program (https://commonfund.nih.gov/kidsfirst/overview). Access to the data and analysis was exempted by the UTHealth IRB under protocol HSC-SPH-18–1127. On November 11, 2019, we obtained a total of 380 trios afflicted with cleft lip with or without cleft palate from the Gabriella Miller Kids First Data Resource Center (DRC). All authors confirmed that

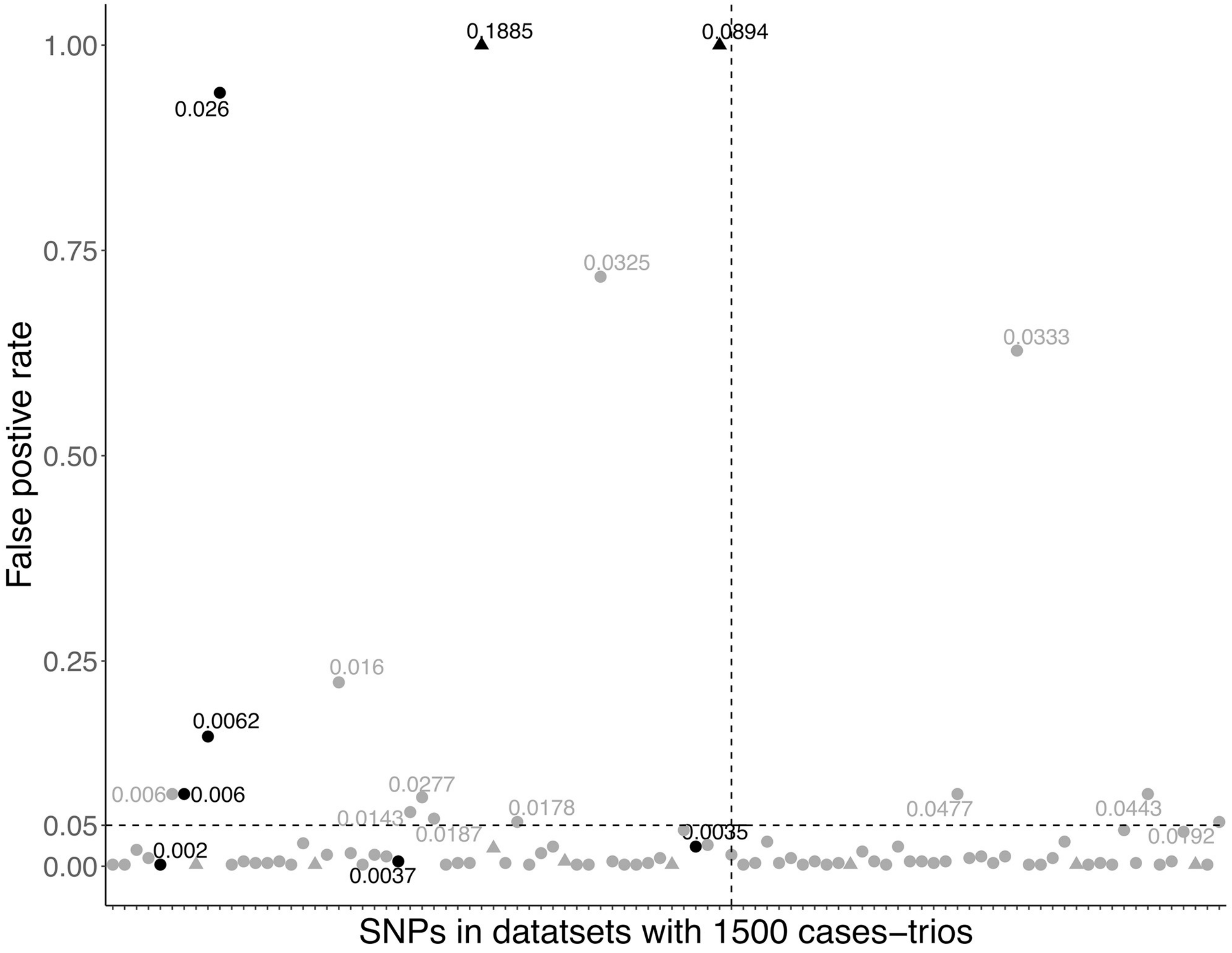

Fig 3. The true and false positive rate of SNPs with corresponding median MAF in 500 data sets with 1500 case-trios respectively. The horizontal dash line represents a threshold of rate 0.05; the vertical line separates the causal and non-causal variants. Variants with black color are true positives; grey illustrates the false positive variants; dot denotes rare variants; triangle denotes common variants.

https://doi.org/10.1371/journal.pone.0314502.g003

we did not have access to any information that could identify individual participants during and after data collection. We applied TRIO_RVEMVS to analyze chromosome 8 sequencing data for association with the risk of orofacial clefts in the European population. Quality control was performed using PLINK [51] according to the guidance from [52], including sample and marker genotyping efficiency/call rate, Mendelian inconsistency, and Hardy-Weinberg equilibrium. Subsequently, SHAPEIT2 [53] was utilized to phase the genotypes and obtain haplotypes for each trio of individuals. For TRIO_RVEMVS testing, our focus was on the region around 8q24 where the SNPs have been identified to be associated with the risk of orofacial clefts in the previous literature [54]. First, we identified the LD blocks in Chromosome 8 using Big-LD [55]. The LD block covering the region previously associated with orofacial clefts consists of 10401 SNPs after omitting singletons, doubletons, and tripletons from our analysis across the 380 trios (1140 individuals). We used the same procedure as described above for the simulated data to determine the exclusion parameters of the priors, which incorporates the regularization plot, Fig 6. The final selected SNPs were shown in Table 5. In total, we identified

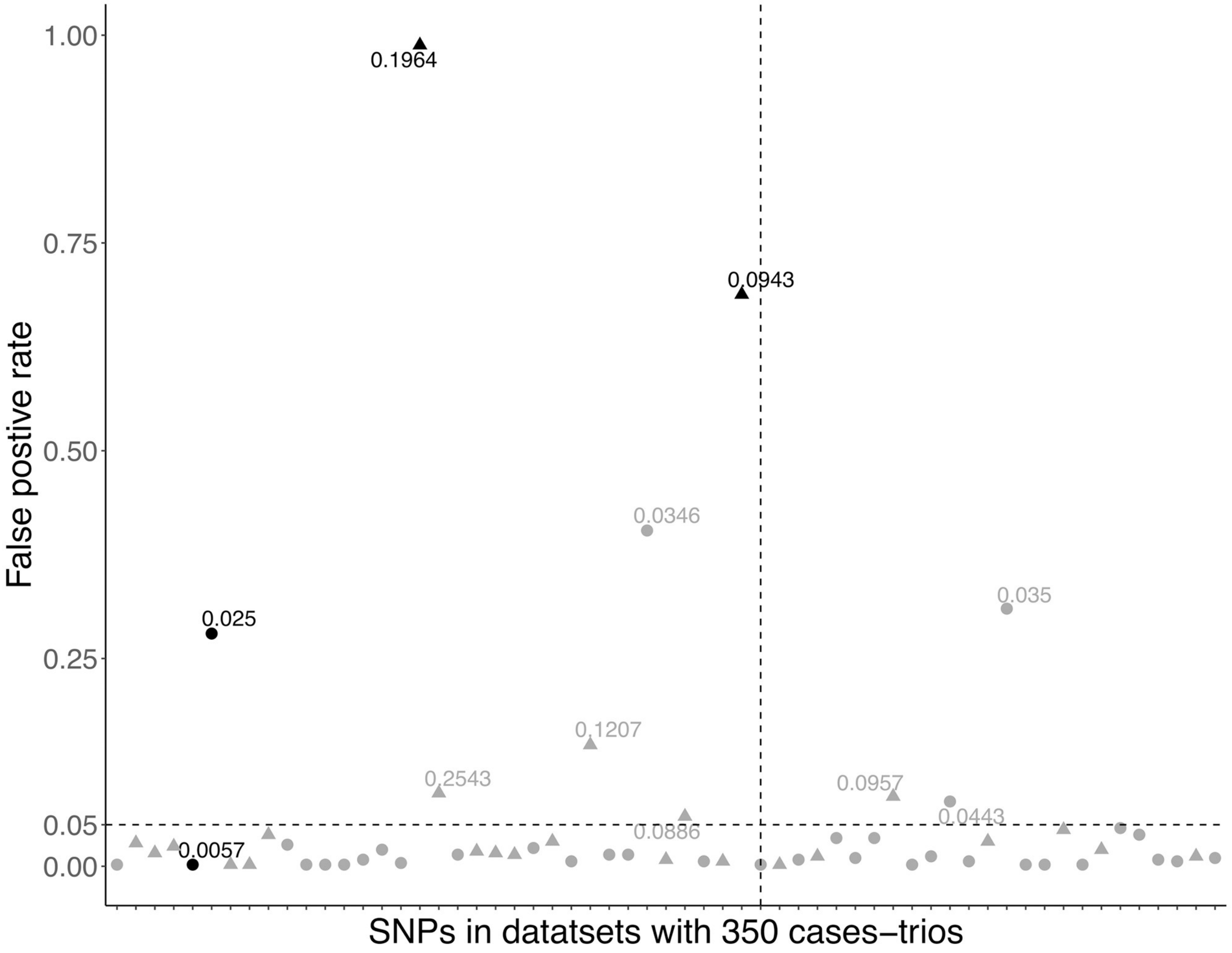

Fig 4. The true and false positive rate of SNPs with corresponding median MAF in 500 data sets with 350 case-trios respectively. The horizontal dash line represents a threshold of rate 0.05; the vertical line separates the causal and non-causal variants. Variants with black color are true positives; grey illustrates the false positive variants; dot denotes rare variants; triangle denotes common variants.

https://doi.org/10.1371/journal.pone.0314502.g004

Table 4. The average true and false positive rate of individual variants detection in different median MAF ranges for variants that were polymorphic across 500 datasets with different sample sizes.

	Sample size	MAF<0.01	0.01≤ MAF<0.05	MAF≥0.05	Total
Number of associated	1500	9	1	2	12
	350	1	1	2	4
ATPR (%)	1500	3.08	94.2	100	26.83
	350	0	26.2	83.8	48.45
Number of unassociated	1500	332	41	46	419
	350	43	40	46	129
AFPR (%)	1500	0.09	6.08	0.09	0.67
	350	0	2.70	1.51	1.38

https://doi.org/10.1371/journal.pone.0314502.t004

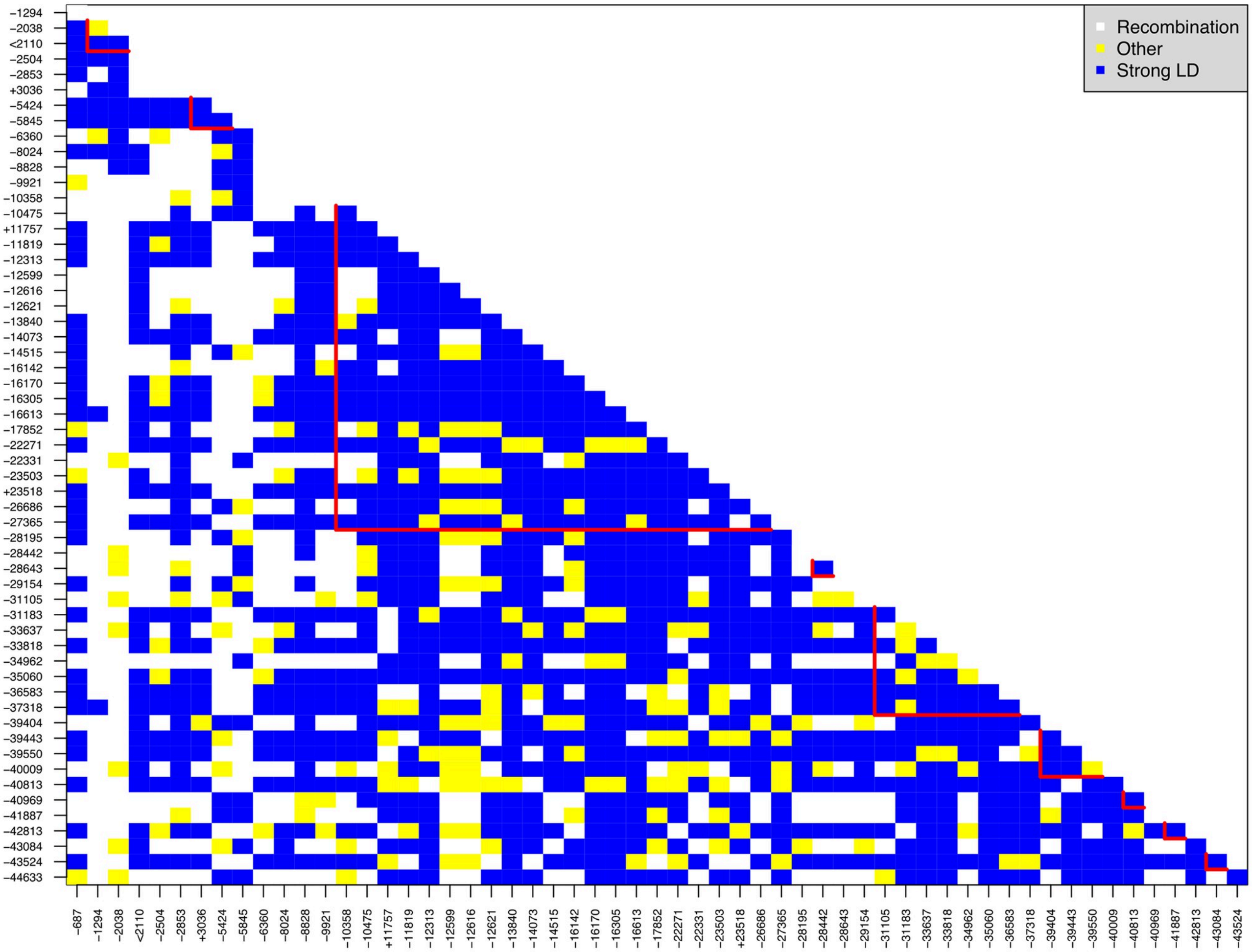

Fig 5. Linkage-disequilibrium (LD) for variants that were polymorphic across 500 data sets with 350 case-trios. The symbol '+' before variant names denote detected associated rare variants; '-' denotes detected non-associated variants; '<' denotes never detected associated variants. V16613 from region 5, and V37318 from region 10 display a high false positive rate, potentially due to LD. They both have strong LD with all the causal rare variants that are polymorphic across 500 data sets.

https://doi.org/10.1371/journal.pone.0314502.g005

8 SNPs in q24.21 and q24.22 associated with the risk of orofacial clefts in the Kids First European population.

Discussion

Although new sequencing technologies and statistical methods have accelerated genome-wide association studies, a large portion of the genetic variability associated with birth defects remains to be discovered [6, 56–58]. These missing inheritances may reside in rare variants. Most existing genetic association methods, such as SKAT [28], PedGene [37], and RV-TDT [38], aggregate the burden of risks of rare variants within a region to test for association between that region and diseases. These methods often experience reduced power when a large number of unassociated rare variants are present within the region pooled, or when the rare variants are antagonistic within the same region (i.e. some promote risk while some offer protection against the disease) [29]. Additionally, it is well known that trio data is more robust to population stratification, and trio methods are well suited to help identify the risk of birth

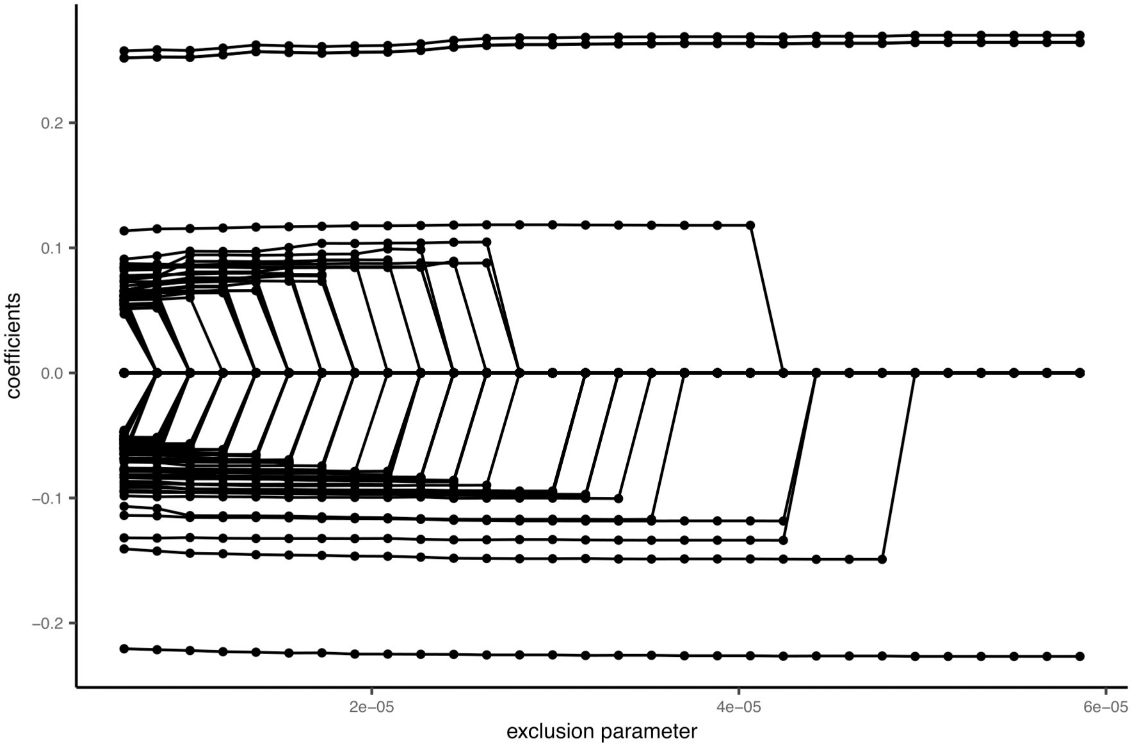

Fig 6. Regularization plot for rare variants based on the trio data from the Gabriella Miller Kids First Data Resource Center (DRC).

https://doi.org/10.1371/journal.pone.0314502.g006

defects stemming from genetic variation [6, 56–59]. We developed a statistical tool based on trio family data, TRIO_RVEMVS, that jointly models common and rare variants to identify the rare variants driving the association of a genetic region with the disease rather than simply assessing genetic regions. One of the advantages of the proposed method is that the common and rare variants are detected simultaneously. The selection of rare variants does not need to be restricted to the region where common variants have been detected previously. Additionally, TRIO_RVEMVS can be potentially applied in fine-mapping studies, particularly following genome-wide association studies (GWAS) that have identified broad regions associated with certain phenotypes or diseases.

Using simulated data, we assessed the performance of TRIO_RVEMVS by comparing its performance at the region level with PedGene and RV-TDT using a weighted average

Table 5. Final selected SNPs in the trio data from the Gabriella Miller Kids First Data Resource Center (DRC).

dbSNP	Position ref	MAF	Locus	Coefficients
rs1474668949	128825584	0.01	q24.21	-0.12
rs7017665	128946138	0.29	q24.21	0.26
rs17242358	128952627	0.29	q24.21	0.26
rs55658222	128963890	0.29	q24.21	0.27
rs1472381856	129156395	0.07	q24.21	-0.23
–	129243536	0.04	–	-0.15
rs1192270083	129364943	0.02	q24.21	-0.13
rs78061696	130619334	0.03	q24.22	0.12

https://doi.org/10.1371/journal.pone.0314502.t005

correct association metric. We also examined the average true positive rate (ATPR) and average false positive rates (AFPR) when identifying individual variants. TRIO_RVEMVS outperformed PedGene when common variants were included whereas both methods were competitive when considering only rare variants. In this study, we also confirmed the result that PedGene outperformed RV-TDT whether common variants were included or not at the region level [50]. For 500 datasets with 1,500 trios, the ATPR was 2.45% and AFPR was 0.07% when both common and rare were considered at the individual level; the ATPR was 0.94%, and AFPR was 0.07% when the rare variants were considered. For 500 data sets with 350 trios, the ATPR was 4.33% and AFPR was 0.13% with common variants; ATPR was 0.62% and AFPR was 0.08% without common variants at the individual level.

When applying TRIO_RVEMVS to real data from the Gabriella Miller Kids First Data Resource Center (DRC), it identified 8 SNPs in q24.21 and q24.22 that were associated with the risk of orofacial clefts in the Kids First European population. Three SNPs were previously reported as common variants in locus 8q24. SNP rs7017665 has been reported in literature [60] and is highly correlated with another generally reported SNP, rs987525, with LD ($r^2 = 0.847$, $D' = 0.983$) [54, 60–62]. SNP rs55658222 and rs17242358 have both been previously reported in the literature [63, 64], respectively. One SNP we identified, rs78061696 has not yet been identified in the literature as associated with orofacial clefts.

We admit that one limitation of the proposed method is modeling the haplotype data which needs the assumption of accurate phasing. Therefore, accurate phasing is crucial before applying the proposed TRIO_EVEMVS. Fortunately, many phasing methods and software have been developed and different methods can be applied in different situations to achieve better accuracy [65]. For example, given the trio data, one can apply MERLIN [66], BEAGLE [67], and SHAPE-IT2 [53] et al. These methods work well for trios and parent-offspring pairs. In this study, we applied SHAPE-IT2 for phasing the real-data analysis. However, comparing different phasing methods and exploring their effect on the downstream analysis is beyond the scope of this study. Incorporating the uncertainty of phasing in the TRIO_EVEMVS is considered future research. Some challenges remain for TRIO_RVEMVS to detect rare variants: 1) TRIO_RVEMVS may fail to detect associated rare variants if their MAF is too small (less than 0.0027), and this threshold varies by sample size. 2) If there are too few rare variants associated with a disease in a given region, TRIO_RVEMVS may only detect the region and not detect the individual variant. 3) TRIO_RVEMVS may falsely detect some variants due to high LD. Despite these challenges, TRIO_RVEMVS is pioneering in its ability to identify individual rare variants alongside gene regions. Extending TRIO_RVEMVS to genome-wide data is considered as future research.

Supporting information

S1 File. Supplemental materials.
(PDF)

Acknowledgments

We would like to acknowledge the Gabriella Miller Kids First Pediatric Research Program for providing the data.

Author Contributions

Conceptualization: Matthew Koslovsky, Michael D. Swartz.

Data curation: Margaret C. Steiner, Kusha Mohammadi.

Formal analysis: Duo Yu.

Funding acquisition: Michael D. Swartz.

Investigation: Duo Yu, Margaret C. Steiner, Kusha Mohammadi, Chenguang Zhang, Michael D. Swartz.

Methodology: Duo Yu, Matthew Koslovsky, Michael D. Swartz.

Resources: Michael D. Swartz.

Software: Duo Yu.

Supervision: Michael D. Swartz.

Writing – original draft: Duo Yu, Michael D. Swartz.

Writing – review & editing: Duo Yu, Matthew Koslovsky, Margaret C. Steiner, Kusha Mohammadi, Chenguang Zhang, Michael D. Swartz.

References

1. CDC. About Birth Defects, May162024; 2024. https://www.cdc.gov/birth-defects.
2. Sattolo ML, Arbour L, Bilodeau-Bertrand M, Lee GE, Nelson C, Auger N. Association of birth defects with child mortality before age 14 years. JAMA Network Open. 2022; 5(4):e226739–e226739. https://doi.org/10.1001/jamanetworkopen.2022.6739 PMID: 35404459
3. Beames TG, Lipinski RJ. Gene-environment interactions: aligning birth defects research with complex etiology. Development. 2020; 147(21):dev191064. https://doi.org/10.1242/dev.191064 PMID: 32680836
4. Qiao F, Wang Y, Zhang C, Zhou R, Wu Y, Wang C, et al. Comprehensive evaluation of genetic variants using chromosomal microarray analysis and exome sequencing in fetuses with congenital heart defect. Ultrasound in Obstetrics & Gynecology. 2021; 58(3):377–387. https://doi.org/10.1002/uog.23532 PMID: 33142350
5. Yang X, Li Q, Wang F, Yan L, Zhuang D, Qiu H, et al. Newborn screening and genetic analysis identify six novel genetic variants for primary carnitine deficiency in Ningbo Area, China. Frontiers in Genetics. 2021; 12:686137. https://doi.org/10.3389/fgene.2021.686137 PMID: 34249102
6. Yuan S, Zaidi S, Brueckner M. Congenital heart disease: emerging themes linking genetics and development. Current opinion in genetics & development. 2013; 23(3):352–359. https://doi.org/10.1016/j.gde.2013.05.004 PMID: 23790954
7. Rashkin SR, Cleves M, Shaw GM, Nembhard WN, Nestoridi E, Jenkins MM, et al. A genome-wide association study of obstructive heart defects among participants in the National Birth Defects Prevention Study. American Journal of Medical Genetics Part A. 2022; 188(8):2303–2314. https://doi.org/10.1002/ajmg.a.62759 PMID: 35451555
8. Shabana N, Shahid SU, Irfan U. Genetic contribution to congenital heart disease (CHD). Pediatric cardiology. 2020; 41(1):12–23. https://doi.org/10.1007/s00246-019-02271-4 PMID: 31872283
9. Lyu C, Webber DM, MacLeod SL, Hobbs CA, Li M, Study NBDP. Gene-by-gene interactions associated with the risk of conotruncal heart defects. Molecular Genetics & Genomic Medicine. 2020; 8(1):e1010. https://doi.org/10.1002/mgg3.1010 PMID: 31851787
10. Sun H, Yi T, Hao X, Yan H, Wang J, Li Q, et al. Contribution of single-gene defects to congenital cardiac left-sided lesions in the prenatal setting. Ultrasound in Obstetrics & Gynecology. 2020; 56(2):225–232. https://doi.org/10.1002/uog.21883 PMID: 31633846
11. Cordell HJ, Töpf A, Mamasoula C, Postma AV, Bentham J, Zelenika D, et al. Genome-wide association study identifies loci on 12q24 and 13q32 associated with Tetralogy of Fallot. Human Molecular Genetics. 2013; 22(7):1473–1481. https://doi.org/10.1093/hmg/dds552 PMID: 23297363
12. Case AP, Mitchell LE. Prevalence and patterns of choanal atresia and choanal stenosis among pregnancies in Texas, 1999-2004. American Journal of Medical Genetics, Part A. 2011; 155(4):786–791. https://doi.org/10.1002/ajmg.a.33882 PMID: 21416593

13. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. American Journal of Human Genetics. 2008; 82 (1):100–112. https://doi.org/10.1016/j.ajhg.2007.09.006 PMID: 18179889

14. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods in Molecular Biology (Clifton, NJ). 2007; 376:71–84. https://doi.org/10.1007/978-1-59745-389-9_6 PMID: 17984539

15. Smith DJ. The allelic structure of common disease. Human Molecular Genetics. 2002; 11(20):2455–2461. https://doi.org/10.1093/hmg/11.20.2455 PMID: 12351581

16. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, Dickson SP, et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American Journal of Human Genetics. 2011; 88(4):458–468. https://doi.org/10.1016/j.ajhg.2011.03.008 PMID: 21457907

17. Young AI. Solving the missing heritability problem. PLoS Genetics. 2019; 15(6):e1008222. https://doi.org/10.1371/journal.pgen.1008222 PMID: 31233496

18. Asimit J, Zeggini E. Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics. 2010; 44(1):293–308. https://doi.org/10.1146/annurev-genet-102209-163421 PMID: 21047260

19. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics. 2010; 11(11):773–785. https://doi.org/10.1038/nrg2867 PMID: 20940738

20. Basu S, Pan W. Comparison of Statistical Tests for Disease Association with Rare Variants. Bone. 2011; 23(1):1–7. https://doi.org/10.1002/gepi.20609 PMID: 21769936

21. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology. 2010; 34(2):188–193. https://doi.org/10.1002/gepi.20450 PMID: 19810025

22. Hoogmartens J, Cacace R, Van Broeckhoven C. Insight into the genetic etiology of Alzheimer's disease: A comprehensive review of the role of rare variants. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2021; 13(1):e12155. https://doi.org/10.1002/dad2.12155 PMID: 33665345

23. Andrés AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2007; 31(7):659–671. https://doi.org/10.1002/gepi.20185 PMID: 17922479

24. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. American Journal of Human Genetics. 2008; 83(3):311–321. https://doi.org/10.1016/j.ajhg.2008.06.024 PMID: 18691683

25. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. American Journal of Human Genetics. 2010; 86 (6):832–838. https://doi.org/10.1016/j.ajhg.2010.04.005 PMID: 20471002

26. Pan W, Shen X. Adaptive Tests for Association Analysis of Rare Variants. Genetic Epidemiology. 2011; 35(5):381–388. https://doi.org/10.1002/gepi.20586 PMID: 21520272

27. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS genetics. 2011; 7(3):e1001322. https://doi.org/10.1371/journal.pgen.1001322 PMID: 21408211

28. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics. 2011; 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059

29. Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating Model Uncertainty in Detecting Rare Variants: The Bayesian Risk Index. Genetic Epidemiology. 2011; 35(7):638–649. https://doi.org/10.1002/gepi.20613 PMID: 22009789

30. Liang F, Xiong M. Bayesian Detection of Causal Rare Variants under Posterior Consistency. PLoS ONE. 2013; 8(7):1–16. https://doi.org/10.1371/journal.pone.0069633 PMID: 23922764

31. Boutry S, Helaers R, Lenaerts T, Vikkula M. Rare variant association on unrelated individuals in case–control studies using aggregation tests: existing methods and current limitations. Briefings in Bioinformatics. 2023; 24(6):bbad412. https://doi.org/10.1093/bib/bbad412 PMID: 37974506

32. Liu J, Lewinger JP, Gilliland FD, Gauderman WJ, Conti DV. Confounding and heterogeneity in genetic association studies with admixed populations. American Journal of Epidemiology. 2013; 177(4):351–360. https://doi.org/10.1093/aje/kws234 PMID: 23334005

33. Li Y, Lee S. Integrating external controls in case–control studies improves power for rare-variant tests. Genetic Epidemiology. 2022; 46(3-4):145–158. https://doi.org/10.1002/gepi.22444 PMID: 35170803

34. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. Biostatistics. 2016; 17(1):1–15. https://doi.org/10.1093/biostatistics/kxv033 PMID: 26363037

35. Schaid DJ, Rowland C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. American Journal of Human Genetics. 1998; 63(5):1492–1506. https://doi.org/10.1086/302094 PMID: 9792877

36. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics. 2012; 44(3):243–246. https://doi.org/10.1038/ng.1074 PMID: 22306651

37. Schaid DJ, Mcdonnell SK, Sinnwell JP, Thibodeau SN. Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. Genetic Epidemiology. 2013; 37(5):409–418. https://doi.org/10.1002/gepi.21727 PMID: 23650101

38. He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RLP, et al. Rare-variant extensions of the transmission disequilibrium test: Application to autism exome sequence data. American Journal of Human Genetics. 2014; 94(1):33–46. https://doi.org/10.1016/j.ajhg.2013.11.021 PMID: 24360806

39. Thomas D, Pitkaeniemi J, Langholz B, Tuomilehto-Wolf E, Tuomilehto J, the DiMe Study Group. Variation in HLA-Associated Risks of Childhood Insulin-Dependent Diabetes in the Finnish Population: II. Haplotype Effects. Genetic Epidemiology. 1995; 12:455–466. https://doi.org/10.1002/gepi.1370120503

40. Schaid DJ. General Score Tests for Associations of Genetic Markers with Disease Using Cases and Their Parents. Genetic Epidemiology. 1996; 13:423–449. https://doi.org/10.1002/(SICI)1098-2272(1996)13:5%3C423::AID-GEPI1%3E3.0.CO;2-3 PMID: 8905391

41. Schaid D. Relative-risk Regression Models Using Cases and Their Parents. Genetic Epidemiology. 1995; 12:813–818. https://doi.org/10.1002/gepi.1370120647 PMID: 8788014

42. Swartz MD, Kimmel M, Mueller P, Amos CI. Stochastic search gene suggestion: a Bayesian hierarchical model for gene mapping. Biometrics. 2006; 62(2):495–503. https://doi.org/10.1111/j.1541-0420.2005.00451.x PMID: 16918914

43. Ročková V, George EI. EMVS: The EM approach to Bayesian variable selection. Journal of the American Statistical Association. 2014; 109(506):828–846. https://doi.org/10.1080/01621459.2013.869223

44. Chen RB, Chu CH, Yuan S, Wu YN. Bayesian Sparse Group Selection. Journal of Computational and Graphical Statistics. 2016; 25(3):665–683. https://doi.org/10.1080/10618600.2015.1041636

45. Shalev-Shwartz S, Zhang T. Stochastic Dual Coordinate Ascent methods for regularized loss minimization. Journal of Machine Learning Research. 2013; 14(1):567–599.

46. Shalev-Shwartz S, Zhang T. Accelerated mini-batch stochastic dual coordinate ascent. Advances in Neural Information Processing Systems. 2013; p. 1–17.

47. Ueda N, Nakano R. Deterministic annealing variant of the EM algorithm. Advances in Neural Information Processing Systems. 1995; p. 545–552.

48. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Research. 2005; 15(11):1576–1583. https://doi.org/10.1101/gr.3709305 PMID: 16251467

49. Siva N. 1000 Genomes project. Nature Biotechnology. 2008; 26(3):256–257. https://doi.org/10.1038/nbt0308-256b PMID: 18327223

50. Wang L, Choi S, Lee S, Park T, Won S. Comparing family-based rare variant association tests for dichotomous phenotypes. BMC Proceedings. 2016; 10(Suppl 7). https://doi.org/10.1186/s12919-016-0027-8 PMID: 27980633

51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. https://doi.org/10.1086/519795 PMID: 17701901

52. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Current Protocols in Human Genetics. 2011; 68(1):1–19. https://doi.org/10.1002/0471142905.hg0119s68 PMID: 21234875

53. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature Methods. 2013; 10(1):5. https://doi.org/10.1038/nmeth.2307 PMID: 23269371

54. Assis Machado R, de Toledo IP, Martelli-Júnior H, Reis SR, Neves Silva Guerra E, Coletta RD. Potential genetic markers for nonsyndromic oral clefts in the Brazilian population: A systematic review and meta-analysis. Birth Defects Research. 2018; 110(10):827–839. https://doi.org/10.1002/bdr2.1208 PMID: 29446255

55. Kim SA, Cho CS, Kim SR, Bull SB, Yoo YJ. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. Bioinformatics. 2018; 34(3):388–397. https://doi.org/10.1093/bioinformatics/btx609 PMID: 29028986

56. Copp AJ, Stanier P, Greene NDE. Neural tube defects: Recent advances, unsolved questions, and controversies. The Lancet Neurology. 2013; 12(8):799–810. https://doi.org/10.1016/S1474-4422(13)70110-8 PMID: 23790957

57. Greene NDE, Stanier P, Copp AJ. Genetics of human neural tube defects. Human Molecular Genetics. 2009; 18(R2). https://doi.org/10.1093/hmg/ddp347 PMID: 19808787

58. Shkoukani MA, Chen M, Vong A. Cleft lip—A comprehensive review. Frontiers in Pediatrics. 2013; 1 (DEC):1–10. https://doi.org/10.3389/fped.2013.00053 PMID: 24400297

59. Carroll NC. Clubfoot in the twentieth century: Where we were and where we may be going in the twenty-first century. Journal of Pediatric Orthopaedics Part B. 2012; 21(1):1–6. https://doi.org/10.1097/BPB.0b013e32834a99f2 PMID: 21946867

60. Leslie EJ, Taub MA, Liu H, Steinberg KM, Koboldt DC, Zhang Q, et al. Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. The American Journal of Human Genetics. 2015; 96(3):397–411. https://doi.org/10.1016/j.ajhg.2015.01.004 PMID: 25704602

61. Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, et al. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. The Journal of Pediatrics. 2009; 155(6):909–913. https://doi.org/10.1016/j.jpeds.2009.06.020 PMID: 19656524

62. Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nature Genetics. 2010; 42(6):525–529. https://doi.org/10.1038/ng.580 PMID: 20436469

63. Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, et al. A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24. 2, 17q23 and 19q13. Human Molecular Genetics. 2016; 25(13):2862–2872. https://doi.org/10.1093/hmg/ddw104 PMID: 27033726

64. Zhang W. Identification and comparison of imputed and genotyped variants for genome-wide association study of orofacial cleft case-parent trios. Johns Hopkins University; 2019.

65. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011; 12(10):703–714. https://doi.org/10.1038/nrg3054 PMID: 21921926

66. Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. The American Journal of Human Genetics. 2005; 77(5):754–767. https://doi.org/10.1086/497345 PMID: 16252236

67. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics. 2009; 84(2):210–223. https://doi.org/10.1016/j.ajhg.2009.01.005 PMID: 19200528