

Article

Gene-Based Methods for Estimating the Degree of the Skewness of X Chromosome Inactivation

Meng-Kai Li ^{1,2,†} , Yu-Xin Yuan ^{1,2,†} , Bin Zhu ^{1,2} , Kai-Wen Wang ^{1,2} , Wing Kam Fung ³ 
and Ji-Yuan Zhou ^{1,2,*} 

- ¹ Department of Biostatistics, State Key Laboratory of Organ Failure Research, Ministry of Education, and Guangdong Provincial Key Laboratory of Tropical Disease Research, School of Public Health, Southern Medical University, Guangzhou 510515, China; lmksum@163.com (M.-K.L.); yuxinyuan2000@163.com (Y.-X.Y.); labenjy@163.com (B.Z.); wkw18927270041@hotmail.com (K.-W.W.)
- ² Guangdong-Hong Kong-Macao Joint Laboratory for Contaminants Exposure and Health, Guangzhou 510006, China
- ³ Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China; wingfung@hku.hk
- * Correspondence: zhoujiyuan5460@hotmail.com
- † These authors contributed equally to this work.

Abstract: Skewed X chromosome inactivation (XCI-S) has been reported to be associated with some X-linked diseases, and currently several methods have been proposed to estimate the degree of the XCI-S (denoted as γ) for a single locus. However, no method has been available to estimate γ for genes. Therefore, in this paper, we first propose the point estimate and the penalized point estimate of γ for genes, and then derive its confidence intervals based on the Fieller's and penalized Fieller's methods, respectively. Further, we consider the constraint condition of $\gamma \in [0, 2]$ and propose the Bayesian methods to obtain the point estimates and the credible intervals of γ , where a truncated normal prior and a uniform prior are respectively used (denoted as GBN and GBU). The simulation results show that the Bayesian methods can avoid the extreme point estimates (0 or 2), the empty sets, the noninformative intervals ($[0, 2]$) and the discontinuous intervals to occur. GBN performs best in both the point estimation and the interval estimation. Finally, we apply the proposed methods to the Minnesota Center for Twin and Family Research data for their practical use. In summary, in practical applications, we recommend using GBN to estimate γ of genes.

Keywords: skewed X chromosome inactivation; Fieller's method; penalized Fieller's method; Bayesian method; Minnesota Center for Twin and Family Research data



Citation: Li, M.-K.; Yuan, Y.-X.; Zhu, B.; Wang, K.-W.; Fung, W.K.; Zhou, J.-Y. Gene-Based Methods for Estimating the Degree of the Skewness of X Chromosome Inactivation. *Genes* **2022**, *13*, 827. <https://doi.org/10.3390/genes13050827>

Academic Editor: Zhongxue Chen

Received: 14 April 2022

Accepted: 2 May 2022

Published: 6 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

X chromosome inactivation (XCI) is an important epigenetic phenomenon. Under the XCI, one of two X chromosomes in females is silenced in the early stage of embryonic development to ensure that the transcriptional dosage of X chromosomes in females and that in males are balanced [1]. Generally, there are three patterns of the XCI [2], random X chromosome inactivation (XCI-R), skewed X chromosome inactivation (XCI-S) [3–6], and escape from X chromosome inactivation (XCI-E) [7,8]. The XCI-R means that the paternal and maternal X chromosomes in females have the same probabilities to be inactive, i.e., for a locus on the X chromosome, approximately 50% of the cells inactivate one of the alleles, while the remaining 50% of the cells keep the other allele inactive. Under the XCI-E, the alleles on both the X chromosomes in females are expressed, which are similar to those at an autosomal locus. For humans, about 15–30% of the X-linked genes have been reported to undergo the XCI-E [7]. Finally, the XCI-S is defined as more than 75% of the cells in females inactivating the same allele [9]. For some extreme skewed cases, it is possible that more than 90% of the cells keep the same allele silenced [9,10]. As such, the difference

in the number of the X chromosomes in females and males and the complexity of the XCI make the association tests for the X chromosomes more complicated than those for the autosomes.

The skewness of the XCI can reflect, or cause, biological consequences for females [9]. The clonal expansion of a somatic cell in females may lead to a cell population with extremely skewed XCI [9]. For some X-linked disorders, there is strong selection of the cells which keep the mutant allele inactive in the heterozygous carriers and, hence, assessing the degree of the skewness of the XCI is helpful in terms of being indicative of the carrier's disease status [11]. Further, the degree of the skewness of the XCI can determine the severity of certain X-linked diseases, such as haemophilia B [12,13]. On the other hand, even for the same mutant allele, the XCI-S in different tissues or cells may result in different clinical consequences. For example, in heterozygous females with a mutant FoxP3 allele, the XCI-S against the mutant allele in specific tissues can prevent autoimmune disease, while the XCI-S skewed towards the mutant allele in breast epithelial cells can cause breast cancer [14]. Besides this, studies have shown that some diseases, such as ovarian cancer, Rett syndrome, Duchenne muscular dystrophy and recurrent miscarriage, are also related to the XCI-S [15–18]. Therefore, in recent years, researchers have proposed some methods to test the association between the alleles at an X-chromosomal single nucleotide polymorphism (SNP) locus and traits [19–26]. For example, Wang et al. [23] developed a permutation-based test statistic which considers all the XCI patterns. For the XCI-R and the XCI-S, this method respectively codes three female genotypes (dd , Dd and DD) as 0, γ and 2 at an X-chromosomal SNP, with the major allele d and the minor allele D , where $\gamma \in [0, 2]$ is an unknown genotypic value for heterozygous females, and respectively codes two male genotypes (d and D) as 0 and 2. Here, γ can be used to measure the degree of the XCI skewing. For instance, $\gamma \in [0, 1)$ is indicative of the XCI-S skewed towards the minor allele D , $\gamma = 1$ means that the XCI pattern is the XCI-R, and $\gamma \in (1, 2]$ indicates the XCI-S skewed towards the major allele d . For the XCI-E, three female genotypes are coded as 0, 1 and 2, and two male genotypes are coded as 0 and 1, respectively. However, the X-chromosomal association tests mentioned above are only applicable to a single SNP and common variants, and are not suitable for genetic regions or genes with multiple SNPs and rare variants. Rare variants refer to the variants with a minor allele frequency (MAF) less than 1%, and those with $MAF \geq 1\%$ are called common variants [27,28]. Over the past few years, genome-wide association studies have identified many common variants associated with complex traits, but these variants usually explain only a small part of the estimated heritability for a given trait. On the other hand, it has been shown that rare variants play a key role in influencing traits [29]. Single-variant tests often have low test power when applied to the rare variants. Therefore, many statistical methods had been presented, which focus on testing the cumulative effect of rare variants in genetic regions or SNP sets (such as genes), including the burden test and the variance-component tests [27,30–33]. The burden test collapses all the rare variants in a genetic region into a single burden variable, and then regresses the trait on the burden variable to test the cumulative effect of the rare variants in that region [27]. The variance-component tests, such as the sequence kernel association test (SKAT), do not directly aggregate the variants in the modeling process, but aggregate the association between the variants and the trait through a kernel matrix [33]. Another method, SKAT-O, proposed by Lee et al. [34], has the advantages of both the burden and SKAT tests, but the time cost is higher than the previous two methods. All these methods have one thing in common, i.e., increasing the weights of rare variants' contributions and decreasing the weights of common variants' contributions. However, for a trait-related gene, the relative influence of rare and common variants is not known [35]. Therefore, Iuliana et al. [35] put forward several multi-locus association tests, such as the adaptive sum test, which consider the effects of both common and rare variants on the trait, and these methods are more powerful when the genes simultaneously contain rare and common variants. Note that these multi-locus association tests are all based on genetic regions or genes on autosomes, and may not be directly applied to the X chromosomes. Therefore,

Clement et al. [36] improved the traditional burden test, SKAT and SKAT-O methods and suggested three gene-based X-chromosomal association tests. However, these methods only take account of the XCI-R and XCI-E patterns. What is more, the FxSKAT method, proposed by Asuman et al. [37], is not only applicable to pedigree data, but also takes the XCI-E into account during the analysis process.

Except for testing the association between the genes on the X chromosome and the traits under study, it is also important to develop methods to measure the corresponding degree of the skewness of the XCI (denoted as γ). At present, researchers have put forward several methods to estimate γ for a single SNP, which can simultaneously get the point estimates and the confidence intervals (CIs) of γ . Specifically, Xu et al. [38] proposed a statistical index for estimating γ based on family trios (both parents and their daughter), which can be represented as the ratio of two relative risks in females, and derived the corresponding CI with the likelihood ratio (LR) test. Wang et al. [39] used the ratio of two regression coefficients of a logistic regression to estimate γ , and obtained the CIs with the LR, Fieller's and delta methods, respectively. Li et al. [40] further extended the methods of Wang et al. so that they can accommodate quantitative traits. However, the above-mentioned methods are all constructed for a single SNP, and are not suitable for genetic regions or genes containing multiple SNPs. Furthermore, when applied to rare variants, they perform poorly. In addition, it should be noted that the delta method cannot control the coverage probability (CP) well, and the LR and Fieller's methods have similar performance in the interval estimation, while the Fieller's method is computationally efficient. Thus, the Fieller's method is recommended in practice. However, both the LR and Fieller's methods may yield unbounded CIs when the denominators in the ratios used to estimate γ are close to 0. Fortunately, the penalized Fieller's (PF) method, which was proposed by Wang et al. [41], can be used to conduct the ratio estimation and always get the bounded CIs by choosing an appropriate penalty parameter. However, it has not been applied to the estimation of the degree of the skewness of the XCI yet. On the other hand, the above-mentioned methods do not consider the constraint condition of $\gamma \in [0, 2]$, and simply cut off the point estimates and the CIs within $[0, 2]$, which may result in extreme point estimates (0 or 2) and empty sets or noninformative CIs (i.e., $[0, 2]$). In contrast, the Bayesian methods can effectively utilize the prior information of each unknown parameter in the analysis, and have been widely used in statistical genetics [42].

Therefore, in this paper, we borrow the idea of the burden test, aggregate all the variants in a gene under study into a burden variable by selecting the appropriate weights, and then estimate the mean degree of the skewness of the XCI over all the SNPs in the gene based on the burden variable. We first propose the point estimate and the penalized point estimate of γ for the gene, and then derive its CIs based on the Fieller's and PF methods, respectively. Then, by considering the constraint condition of $\gamma \in [0, 2]$, we propose the Bayesian methods to obtain the point estimates and the credible intervals of γ . Specifically, after getting enough samples drawn from the posterior distribution of γ , we calculate the mode of the samples as the point estimate of γ and the highest posterior density interval (HPDI) as the credible interval of γ [43]. We conduct extensive simulation studies to compare the performances of the proposed point estimation methods and the interval estimation methods for γ . Finally, we demonstrate the practical utility of the proposed methods by applying them to the Minnesota Center for Twin and Family Research (MCTFR) data.

2. Materials and Methods

2.1. Notations

Suppose that we only collect n female subjects, because male subjects provide no information on the XCI skewing. Consider an X-linked trait (quantitative or qualitative) and let y_i represent the trait value of the i th female ($i = 1, 2, \dots, n$), then $Y = (y_1, y_2, \dots, y_n)^T$ is the vector of the trait values for all the females. Assume that a gene which contains J SNPs is associated with this trait, and let d_j and D_j denote the major allele and the minor

allele at the j th SNP ($j = 1, 2, \dots, J$), respectively. Let G_{ij} be the genotype at the j th SNP of the i th female (i.e., $G_{ij} = d_j d_j, D_j d_j$ or $D_j D_j$). If we use $\gamma \in [0, 2]$ to measure the mean degree of the skewness of the XCI for all the SNPs in the gene, then $g_{ij} = 0, \gamma$ and 2 can be used to denote the genotypic values for genotypes $d_j d_j, D_j d_j$ and $D_j D_j$, respectively. As such, $\mathbf{G}_i = (g_{i1}, g_{i2}, \dots, g_{iJ})^T$ is the vector of the genotypic values at the J SNPs of the i th female. Therefore, we consider the association between the gene and the trait based on the following generalized linear model

$$h(\mu_i) = \beta_0 + \boldsymbol{\beta}^T \mathbf{G}_i + \mathbf{b}^T \mathbf{Z}_i, \tag{1}$$

where $h(\cdot)$ is a link function; $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{im})^T$ is the vector of m covariates of the i th female, which are needed to be adjusted, and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n)^T$ is an $n \times m$ covariate matrix; $\mu_i = E(y_i | \mathbf{G}_i, \mathbf{Z}_i)$ is the conditional mean of the i th female's trait value given \mathbf{G}_i and \mathbf{Z}_i ; β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_J)^T$ is the vector of the regression coefficients of \mathbf{G}_i , and $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$ is an $m \times 1$ vector of the regression coefficients of \mathbf{Z}_i .

Based on the idea of the burden test [27], we aggregate all the SNPs in the gene into a burden variable and let $X_i = \sum_{j=1}^J \omega_j g_{ij}$, where ω_j is a weight for the j th SNP. Here we assume that ω_j is a function with respect to the MAF at the j th SNP (denoted as MAF_j), i.e., $\omega_j = \text{Beta}(\text{MAF}_j, 0.5, 0.5)$ [35]. So, model (1) can be rewritten as

$$h(\mu_i) = \beta_0 + \beta_c X_i + \mathbf{b}^T \mathbf{Z}_i, \tag{2}$$

where β_c is the regression coefficient of X_i . Next, we consider two variables $g_{ij}^{(1)} = I_{\{G_{ij}=D_j d_j \text{ or } D_j D_j\}}$ and $g_{ij}^{(2)} = I_{\{G_{ij}=D_j D_j\}}$, where $I_{\{\cdot\}}$ is the indicator function. Thus, $g_{ij}^{(1)} = 1$ means that the i th female contains at least one minor allele at the j th SNP, and $g_{ij}^{(2)} = 1$ denotes that the female is a homozygote $D_j D_j$ at the j th SNP. Through simple transformations, we can get $g_{ij} = \gamma g_{ij}^{(1)} + (2 - \gamma) g_{ij}^{(2)}$, and X_i can be expressed as $X_i = \sum_{j=1}^J \omega_j [\gamma g_{ij}^{(1)} + (2 - \gamma) g_{ij}^{(2)}] = \gamma X_i^{(1)} + (2 - \gamma) X_i^{(2)}$, where $X_i^{(1)} = \sum_{j=1}^J \omega_j g_{ij}^{(1)}$ and $X_i^{(2)} = \sum_{j=1}^J \omega_j g_{ij}^{(2)}$. Further, let $\mathbf{X}^{(1)} = (X_1^{(1)}, X_2^{(1)}, \dots, X_n^{(1)})^T$ and $\mathbf{X}^{(2)} = (X_1^{(2)}, X_2^{(2)}, \dots, X_n^{(2)})^T$. To estimate the mean degree of the XCI skewing for the gene (i.e., γ), we substitute $X_i = \gamma X_i^{(1)} + (2 - \gamma) X_i^{(2)}$ into model (2) and get

$$h(\mu_i) = \beta_0 + \beta_c [\gamma X_i^{(1)} + (2 - \gamma) X_i^{(2)}] + \mathbf{b}^T \mathbf{Z}_i. \tag{3}$$

For quantitative traits, $h(\cdot)$ is the identity function, and model (3) can be written as $y_i = \beta_0 + \beta_c [\gamma X_i^{(1)} + (2 - \gamma) X_i^{(2)}] + \mathbf{b}^T \mathbf{Z}_i + \varepsilon_i$, where ε_i is the random error and follows $N(0, \sigma^2)$. In this case, the unknown parameters are $\boldsymbol{\theta}_1 = (\beta_0, \beta_c, \gamma, \mathbf{b}^T, \sigma)^T$, and the corresponding likelihood function of the sample is

$$L_1(\boldsymbol{\theta}_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{\sum_{i=1}^n [y_i - \beta_0 - \gamma\beta_c X_i^{(1)} - (2 - \gamma)\beta_c X_i^{(2)} - \mathbf{b}^T \mathbf{Z}_i]^2}{2\sigma^2} \right\}.$$

As for qualitative traits, $h(\cdot)$ is the logit function, and model (3) is written as $\text{Logit}\left(\Pr\left(y_i = 1 \mid X_i^{(1)}, X_i^{(2)}, \mathbf{Z}_i\right)\right) = \beta_0 + \beta_c \left[\gamma X_i^{(1)} + (2 - \gamma) X_i^{(2)}\right] + \mathbf{b}^T \mathbf{Z}_i$. The unknown parameters are $\boldsymbol{\theta}_2 = \left(\beta_0, \beta_c, \gamma, \mathbf{b}^T\right)^T$ and the likelihood function is

$$L_2(\boldsymbol{\theta}_2) = \prod_{i=1}^n \pi_i^{I_{\{y_i=1\}}} (1 - \pi_i)^{I_{\{y_i=0\}}},$$

where $y_i = 1$ and 0 respectively indicate that the i th female is a case and a control, and $\pi_i = 1 / \left\{1 + \exp\left[-\beta_0 - \gamma \beta_c X_i^{(1)} - (2 - \gamma) \beta_c X_i^{(2)} - \mathbf{b}^T \mathbf{Z}_i\right]\right\}$. Let $\beta_c^{(1)} = \gamma \beta_c$ and $\beta_c^{(2)} = (2 - \gamma) \beta_c$, and we have

$$h(\mu_i) = \beta_0 + \beta_c^{(1)} X_i^{(1)} + \beta_c^{(2)} X_i^{(2)} + \mathbf{b}^T \mathbf{Z}_i. \tag{4}$$

As such, we obtain $\beta_c = \left(\beta_c^{(1)} + \beta_c^{(2)}\right) / 2$ and γ can be expressed as

$$\gamma = \frac{\beta_c^{(1)}}{\beta_c} = \frac{2\beta_c^{(1)}}{\beta_c^{(1)} + \beta_c^{(2)}}. \tag{5}$$

By assuming that the degree of the skewness of the XCI at the j th SNP is γ_j , γ satisfies, under a certain condition (the proof is given in Appendix A),

$$\gamma = \frac{\sum_{j=1}^J \omega_j \left(g_j^{(1)} - g_j^{(2)}\right) \gamma_j}{\sum_{j=1}^J \omega_j \left(g_j^{(1)} - g_j^{(2)}\right)},$$

where $g_j^{(1)} = \sum_{i=1}^n g_{ij}^{(1)}$ is the number of the females who contain at least one minor allele at the j th SNP, and $g_j^{(2)} = \sum_{i=1}^n g_{ij}^{(2)}$ is the number of the females whose genotypes at the j th SNP are $D_j D_j$. So, γ is the weighted mean of the γ_j 's for all the SNPs in the gene with the weights being $\omega_j \left(g_j^{(1)} - g_j^{(2)}\right) / \sum_{j=1}^J \omega_j \left(g_j^{(1)} - g_j^{(2)}\right)$. When there are rare variants at some SNPs or when the variation of the γ_j 's in the gene is large, γ is still well defined for the whole gene. On the other hand, from Equation (5), γ can be well defined if there is an association between the gene and the trait (i.e., $\beta_c = \left(\beta_c^{(1)} + \beta_c^{(2)}\right) / 2 \neq 0$). Further, $\gamma = 0$ if and only if $\beta_c^{(1)} = 0$ and $\beta_c^{(2)} \neq 0$, which means that all the γ_j 's are 0 and the XCI-S is completely skewed towards the minor allele for each SNP, and $\gamma = 2$ only when $\beta_c^{(1)} \neq 0$ and $\beta_c^{(2)} = 0$, indicating that all the γ_j 's are 2 and the XCI-S is completely skewed towards the major allele for each SNP. However, $\gamma = 1$ means that on the average, the gene undergoes the XCI-R or the XCI-E. After obtaining the estimates of $\beta_c^{(1)}$ and $\beta_c^{(2)}$, respectively denoted by $\hat{\beta}_c^{(1)}$ and $\hat{\beta}_c^{(2)}$ which can be derived by the maximum likelihood method, the point estimate of γ can be expressed as $\hat{\gamma} = 2\hat{\beta}_c^{(1)} / \left(\hat{\beta}_c^{(1)} + \hat{\beta}_c^{(2)}\right)$.

2.2. Point Estimate and CI of γ by Fieller's Method

Note that γ should take the possible values from the interval $[0, 2]$. So, the original estimate $\hat{\gamma} = 2\hat{\beta}_c^{(1)} / \left(\hat{\beta}_c^{(1)} + \hat{\beta}_c^{(2)}\right)$ needs to be cut off in $[0, 2]$ and the resulting estimate is denoted by $\hat{\gamma}_{GF}$. Further, we utilize the Fieller's method to get the CI of γ . Specifically, borrowing the idea of Wang et al. [39], we have $\hat{\beta}_c = \left(\hat{\beta}_c^{(1)} + \hat{\beta}_c^{(2)}\right) / 2$, $\hat{\text{Var}}\left(\hat{\beta}_c\right) = \frac{1}{4} \left[\hat{\text{Var}}\left(\hat{\beta}_c^{(1)}\right) + \hat{\text{Var}}\left(\hat{\beta}_c^{(2)}\right) + 2\hat{\text{Cov}}\left(\hat{\beta}_c^{(1)}, \hat{\beta}_c^{(2)}\right)\right]$ and $\hat{\text{Cov}}\left(\hat{\beta}_c^{(1)}, \hat{\beta}_c\right) = \frac{1}{2} \hat{\text{Var}}\left(\hat{\beta}_c^{(1)}\right) + \frac{1}{2} \hat{\text{Cov}}\left(\hat{\beta}_c^{(1)}, \hat{\beta}_c^{(2)}\right)$. To construct the CI of γ , we first establish a Wald test under the null

hypothesis $H_0 : \gamma = \gamma_0$, where γ_0 is a pre-specified value (e.g., 1, which means that on the average, the gene undergoes the XCI-R or the XCI-E). As such, we have $\beta_c^{(1)} - \gamma_0\beta_c = 0$, and the Wald test statistic is

$$\frac{\hat{\beta}_c^{(1)} - \gamma_0\hat{\beta}_c}{\sqrt{\hat{\text{Var}}(\hat{\beta}_c^{(1)}) + \gamma_0^2\hat{\text{Var}}(\hat{\beta}_c) - 2\gamma_0\hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c)}} \sim N(0, 1).$$

Therefore, the $100(1 - \alpha)\%$ CI of γ can be derived by solving the following equation

$$\left[\frac{\hat{\beta}_c^{(1)} - \gamma_0\hat{\beta}_c}{\sqrt{\hat{\text{Var}}(\hat{\beta}_c^{(1)}) + \gamma_0^2\hat{\text{Var}}(\hat{\beta}_c) - 2\gamma_0\hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c)}} \right]^2 = Z_{1-\alpha/2}^2,$$

where $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ upper quantile of the standard normal distribution. Rearrange the above equation with respect to γ_0 into a quadratic equation

$$A\gamma_0^2 + B\gamma_0 + C = 0, \tag{6}$$

where $A = \hat{\beta}_c^2 - Z_{1-\alpha/2}^2\hat{\text{Var}}(\hat{\beta}_c)$, $B = 2[Z_{1-\alpha/2}^2\hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c) - \hat{\beta}_c^{(1)}\hat{\beta}_c]$ and $C = (\hat{\beta}_c^{(1)})^2 - Z_{1-\alpha/2}^2\hat{\text{Var}}(\hat{\beta}_c^{(1)})$. When $\Delta = \sqrt{B^2 - 4AC} = 0$ or $A = 0$, the CI of γ will degenerate to be a point. The CI of γ for other cases is as follows

$$\left\{ \begin{array}{l} \left(\frac{-B-\sqrt{\Delta}}{2A}, \frac{-B+\sqrt{\Delta}}{2A} \right) \cap [0, 2], \text{ if } \Delta > 0 \text{ and } A > 0 \\ \left((-\infty, \frac{-B+\sqrt{\Delta}}{2A}) \cup \left(\frac{-B-\sqrt{\Delta}}{2A}, +\infty \right) \right) \cap [0, 2], \text{ if } \Delta > 0 \text{ and } A < 0 \\ [0, 2], \text{ if } \Delta < 0 \text{ and } A < 0 \\ \emptyset, \text{ if } \Delta < 0 \text{ and } A > 0 \end{array} \right.$$

It should be noted that even in the case of $\Delta > 0$, the CI of γ obtained by the Fieller’s method may still be an empty set. And in the case of $\Delta > 0$ and $A < 0$, the CI may be composed of two parts, which is the discontinuous interval.

2.3. Penalized Point Estimate and CI of γ by PF Method

As mentioned above, we construct $\hat{\gamma} = \hat{\beta}_c^{(1)} / \hat{\beta}_c$ as the point estimate of γ , where $\hat{\beta}_c = (\hat{\beta}_c^{(1)} + \hat{\beta}_c^{(2)}) / 2$. However, if the denominator $\hat{\beta}_c$ is very close to 0, $\hat{\gamma}$ will tend to the infinity. The CI of γ based on the Fieller’s method before the truncation is usually unbounded. To deal with this issue in the ratio estimate and borrow the idea of Wang et al. [41], we propose the following PF method to obtain the penalized point estimate of γ and the corresponding CI. Consider the penalized log-likelihood function of β_c as follows: $pl = -(\hat{\beta}_c - \beta_c)^2 / (2\hat{\text{Var}}(\hat{\beta}_c)) + \lambda \log|\beta_c|$, where $\lambda > 0$ is a penalty parameter and is taken to be $Z_{1-\alpha/2}^2/4$ as suggested by Wang et al. [41] because the CI obtained by the PF method is always bounded with $\lambda = Z_{1-\alpha/2}^2/4$. By maximizing the function pl , we have the penalized denominator $\tilde{\beta}_c = \hat{\beta}_c/2 + \text{sign}(\hat{\beta}_c)\sqrt{\hat{\beta}_c^2/4 + \lambda\hat{\text{Var}}(\hat{\beta}_c)}$, where $\text{sign}(\cdot)$ is the signum function. Further, we can get $\hat{\text{Var}}(\tilde{\beta}_c) = \zeta^2\hat{\text{Var}}(\hat{\beta}_c) + O(n^{-3})$, where $\zeta = \tilde{\beta}_c / (2\tilde{\beta}_c - \hat{\beta}_c)$. If we replace $\hat{\beta}_c$ by $\tilde{\beta}_c$ to obtain the point estimate $\tilde{\gamma} = \hat{\beta}_c^{(1)} / \tilde{\beta}_c$, then $\tilde{\gamma}$ is a biased estimate of γ . To reduce this bias, we need to correct the numerator $\hat{\beta}_c^{(1)}$ by $\tilde{\beta}_c^{(1)} = \hat{\beta}_c^{(1)} + \tilde{\gamma}(\tilde{\beta}_c - \hat{\beta}_c)$. Correspondingly, we can get $\hat{\text{Var}}(\tilde{\beta}_c^{(1)}) = \zeta^{-2}\hat{\text{Var}}(\hat{\beta}_c^{(1)}) - 4(\zeta^{-1} - 1)\tilde{\gamma}\hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c) + 4(1 - \zeta)^2\tilde{\gamma}^2\hat{\text{Var}}(\hat{\beta}_c)$ and $\hat{\text{Cov}}(\tilde{\beta}_c^{(1)}, \tilde{\beta}_c) = \hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c) - 2\zeta(1 - \zeta)\tilde{\gamma}\hat{\text{Var}}(\hat{\beta}_c)$. After ob-

taining the corrected denominator $\tilde{\beta}_c$ and the corrected numerator $\tilde{\beta}_c^{(1)}$, $\hat{\gamma}^* = \tilde{\beta}_c^{(1)} / \tilde{\beta}_c$ truncated by $[0, 2]$ is the penalized point estimate of γ , which is denoted by $\hat{\gamma}_{GPF}$. The construction process of the corresponding CI of $\hat{\gamma}_{GPF}$ is the same as the Fieller's method, except for respectively replacing $\hat{\beta}_c, \hat{\beta}_c^{(1)}, \hat{\text{Var}}(\hat{\beta}_c), \hat{\text{Var}}(\hat{\beta}_c^{(1)})$ and $\hat{\text{Cov}}(\hat{\beta}_c^{(1)}, \hat{\beta}_c)$ by $\tilde{\beta}_c, \tilde{\beta}_c^{(1)}, \hat{\text{Var}}(\tilde{\beta}_c), \hat{\text{Var}}(\tilde{\beta}_c^{(1)})$ and $\hat{\text{Cov}}(\tilde{\beta}_c^{(1)}, \tilde{\beta}_c)$ in Equation (6). However, it should be noted that although the CI of γ based on the PF method is always bounded when $\lambda = Z_{1-\alpha/2}^2/4$, it may be out of $[0, 2]$ and we need to truncate it by $[0, 2]$.

2.4. Point Estimate and Credible Interval of γ by Bayesian Method

Note that the point estimates ($\hat{\gamma}_{GF}$ and $\hat{\gamma}_{GPF}$), and the corresponding CIs mentioned above, are cut off in the interval $[0, 2]$ and cannot directly incorporate the information on $\gamma \in [0, 2]$. Therefore, in this subsection, we introduce the Bayesian method to give the point estimate and the credible interval of γ by considering the prior information of $\gamma \in [0, 2]$. Specifically, we have the posterior distribution of the unknown parameter θ as follows

$$f(\theta | \mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Z}) = \frac{f(\theta) L(\theta)}{\int f(\theta) L(\theta) d\theta},$$

where $f(\theta)$ is the joint prior distribution of θ ; when the traits are quantitative, $\theta = \theta_1$ and $L(\theta) = L_1(\theta_1)$; when the traits are qualitative, $\theta = \theta_2$ and $L(\theta) = L_2(\theta_2)$. However, in general, we cannot get the analytical solutions of $f(\theta | \mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Z})$. Therefore, it is not feasible to directly sample from the posterior distribution. Fortunately, there are several algorithms for sampling from an approximate distribution of the posterior distribution, such as the Hamiltonian Monte Carlo (HMC) algorithm which can be implemented by the "rstan" package in R [43]. On the other hand, according to Annis et al. [43], the correlation between the parameters has little influence on the HMC algorithm. To simplify the operations, and improve the sampling efficiency, we assume that the unknown parameters in θ are independent of each other, and use the HMC algorithm to sample from the approximate posterior distribution of θ . In other words, we choose the prior distribution for each unknown parameter separately.

The prior distributions of the parameters in θ are selected as follows. To reduce the influence of the selection of the prior distributions on the results, for nuisance parameters β_0, β_c and \mathbf{b} (there is an additional nuisance parameter σ when the trait is quantitative), we choose the weak prior distributions [44]. Specifically, we assume that the prior distributions of β_0 and β_c are both $N(0, 10^2)$, and that of \mathbf{b} is $MVN(\mathbf{0}, \text{diag}(10^2, 10^2, \dots, 10^2))$. For quantitative traits, we also specify the prior distribution of σ to be an exponential distribution, i.e., $\sigma \sim \exp(1)$. As for the parameter γ of interest, which is used to measure the mean degree of the skewness of the XCI over all the SNPs in the gene and satisfies the constraint condition of $\gamma \in [0, 2]$, we consider two possible prior distributions. The first one is the truncated normal distribution, with both parameters being 1 and the values ranging from 0 to 2, and the probability density function of the prior distribution is

$$f(\gamma) = \begin{cases} \frac{\phi(\gamma-1)}{\frac{1}{\sqrt{2\pi}} \int_0^2 \exp[-\frac{1}{2}(x-1)^2] dx}, & 0 \leq \gamma \leq 2 \\ 0, & \text{otherwise} \end{cases},$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution. In this way, γ not only satisfies the constraint condition of $\gamma \in [0, 2]$, but also the probability of γ being close to 1 is the highest, which is consistent with the literature [2], i.e., most of the SNPs on the X chromosome undergo the XCI-R. Meanwhile, the selected truncated normal distribution of γ also avoids that the probability of γ taking the extreme value (0 or 2) is too low, which may be more suitable for practical applications. The second prior distribution of γ is a uniform distribution, i.e., $\gamma \sim U(0, 2)$.

After specifying the prior distributions of all the unknown parameters, we can get enough samples of γ through the HMC algorithm, and then calculate the mode of the samples as the point estimate of γ , and the highest posterior density interval (HPDI) as the credible interval of γ . Here, we denote the Bayesian methods with the truncated normal prior and the uniform prior as GBN and GBU, and the point estimates obtained by these two methods are denoted as $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$, respectively.

3. Results

3.1. Simulation Settings

We conducted extensive simulation studies to evaluate the performances of the proposed point estimation and interval estimation methods. The number of female subjects (i.e., the sample size n) is set to be 500 and 2000. Consider a gene associated with the trait under study and the number of the SNPs in the gene (i.e., J) is fixed at 100, i.e., we assume that all the 100 SNPs are associated with the trait. Meanwhile, we define η as the proportion of rare variants among the 100 SNPs. To explore the effect of η on the proposed methods, we set $\eta = 0, 0.4$ and 1 , which correspond to the cases of all the 100 SNPs only including common variants, the 100 SNPs simultaneously containing common and rare variants, and all the 100 SNPs only consisting of rare variants, respectively. Among them, the MAFs for common variants are sampled from $U(0.01, 0.5)$, while the MAFs for rare variants are randomly simulated from $U(0.005, 0.01)$ [45–47]. We generate the genotypes of n female subjects by referring to the ideas of Wang et al. [45], Basu et al. [46], and Turkmen et al. [47]. We first generate a latent vector $\mathbf{V} = (V_1, V_2, \dots, V_{100})^T$ from the multivariate normal distribution with the mean vector being $\mathbf{0}$ and the elements of the variance-covariance matrix satisfying $\text{Var}(V_j) = 1$ and $\text{Corr}(V_j, V_k) = \rho^{|j-k|}$ ($j, k = 1, 2, \dots, 100$) [45,47], where the linkage disequilibrium among the SNPs is taken into consideration. For simplicity, we set $\rho = 0.5$ in our simulation studies. Once \mathbf{V} is generated, it is then transformed to 0 (major allele) or 1 (minor allele) determined by the corresponding MAFs. This process is repeated twice, and two simulated vectors of length 100 are put together to form the genotypes at the 100 SNPs for a female subject. After simulating the genotypes of n female subjects, we have an $n \times 100$ genotypic value matrix $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_n)^T$ with the elements being 0, 1 or 2, and then we replace the elements of \mathbf{G} equal to 1 with γ to simulate the information on the XCI-S. Note that to simplify the simulation and better evaluate the performances of our proposed methods (e.g., the calculation of the mean squared errors (MSEs) of the point estimates requires a single true value of γ for each replicate; the details are given later), we set the degrees of the XCI skewing γ_j 's at all the 100 SNPs to be the same in the simulation study (i.e., $\gamma_j = \gamma, j = 1, 2, \dots, 100$).

We only consider a covariate Q , which is generated from the standard normal distribution. For the quantitative trait, we simulate the trait value y_i of the i th female according to the following model

$$y_i = \beta_0 + \beta_1 g_{i1} + \beta_2 g_{i2} + \dots + \beta_{100} g_{i100} + \delta Q_i + \varepsilon_i,$$

where ε_i is the random error, which is generated from the standard normal distribution; β_0 is the intercept and δ is the regression coefficient of the covariate Q , and both the parameters are set to be 0.5 [36]; $|\beta_j| = e^{|\log_{10} \text{MAF}_j|/2}$ is the regression coefficient of the genotypic value g_{ij} at the j th SNP [33,34,36], where e is the tuning parameter and is used to avoid the effect of a SNP being too large or too small [36]. To highlight the effects of rare variants on the trait, we set $e = 1.5$ when the j th SNP has a rare variant, otherwise $e = 1.1$. Further, notice that the directions of the effects of different SNPs on the trait may be different. Therefore, we consider the proportion of the SNPs with positive effects among the 100 SNPs (denoted by τ) and set τ to be 0.6 and 1, indicating that the effect directions of some SNPs are positive and some are negative, and all the SNP effects are positive, respectively. We do not simulate the case of $\tau = 0$ (i.e., all the SNP effects are negative) because the

results with $\tau = 0$ are similar to those with $\tau = 1$. As for the qualitative trait, the model for generating the affection status y_i of the i th female is as follows

$$\text{Logit}(Pr(y_i = 1 | G_i, Q_i)) = \beta_0 + \beta_1 g_{i1} + \beta_2 g_{i2} + \dots + \beta_{100} g_{i100} + \delta Q_i.$$

All of the parameters are the same as when simulating the quantitative trait, except that we need to set the case-control ratio to be 1:1.

After simulating the genotypes and the trait values, we use model (4) to obtain the estimates of $\beta_c^{(1)}$ and $\beta_c^{(2)}$, where $X_i^{(1)} = \sum_{j=1}^{100} \omega_j g_{ij}^{(1)}$, $X_i^{(2)} = \sum_{j=1}^{100} \omega_j g_{ij}^{(2)}$, $\omega_j = \text{Beta}(\text{MAF}_j, 0.5, 0.5)$, and $\hat{\text{MAF}}_j$ is the estimate of the MAF at the j th SNP. Then, we get the point estimate $\hat{\gamma}_{GF}$, the penalized point estimate $\hat{\gamma}_{GPF}$, and the CIs of γ derived by the Fieller's and the PF methods. As for the Bayesian methods, the HMC algorithm is implemented through the "sampling" function in the R package "rstan". We set 8 chains for the parallel sampling in the simulation. For each chain, we extract 10,000 samples, and the first 5000 are used for warm-up. So, we finally get 40,000 samples. To ensure the convergence, the target acceptance rate is set to be 0.99.

The above simulation steps are all implemented in the R software (version 4.1.1, <http://r-project.org>, accessed on 5 January 2022). For each simulation setting, the number of the replicates is fixed to be 500, and for each replicate, the true value of γ is sampled from the uniform distribution $U(0, 2)$. To evaluate the accuracy and the robustness of $\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$, we calculate the MSEs of these point estimates. Here, $\text{MSE} = \sum_{s=1}^{500} (\hat{\gamma}_s - \gamma_s)^2 / 500$, where γ_s represents the true value of γ and $\hat{\gamma}_s$ is the point estimate in the s th replicate ($s = 1, 2, \dots, 500$). Note that $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ are always between 0 and 2, so we only compute the proportions of $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ taking the extreme values (0 or 2), respectively. Meanwhile, scatter plots are used to show the relationship between the four point estimates and the true values of γ . To compare the performances of the GBN, GBU, PF and Fieller's methods in the interval estimation, we calculate the CP as well as the mean, the median, the standard deviation and the interquartile range of the widths of the 95% HPDIs or CIs (denoted by W_{mean} , W_{median} , W_{sd} and W_{iqr}), respectively. For the PF and Fieller's methods, we also compute the proportions of the empty sets (EP), the noninformative intervals (NP), and the discontinuous intervals (DP) to further compare the effectiveness of these two methods, where the noninformative interval means the CI being $[0, 2]$. However, it should be noted that the GBN and GBU methods avoid the cases of empty sets, noninformative intervals, and discontinuous intervals occurring. In addition, we draw the scatter plots between the interval widths of the four proposed methods and the true values of γ .

3.2. Simulation Results

The proportions of the extreme values (0 or 2) for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ are shown in Table 1. It can be seen from the table that the proportions of the point estimates equal to 0 are the same for both $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$, while the proportion of the point estimates equal to 2 for $\hat{\gamma}_{GPF}$ is reduced. This is because before the truncation, both $\hat{\gamma}^* = \tilde{\beta}_c^{(1)} / \tilde{\beta}_c$ and $\hat{\gamma} = \hat{\beta}_c^{(1)} / \hat{\beta}_c$ always have the same sign, and $\hat{\gamma}^*$ is bounded. Specifically, when $\hat{\gamma}^*$ and $\hat{\gamma}$ are negative, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ are both 0. On the other hand, when $\hat{\gamma}^*$ and $\hat{\gamma}$ are positive, compared with $\hat{\gamma}$, the proportion of $\hat{\gamma}^*$ being greater than 2 decreases. Further, from Table 1, with the increase of the sample size or the trait changing from qualitative to quantitative, the proportions of the extreme values for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ both become less. Next, let us take a look at the effects of the proportion of the rare variants (η) and the proportion of the SNPs with the positive effects (τ) among all the SNPs on the proportions of the extreme values for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$. Under the situation that the trait is quantitative and $\tau = 0.6$ (i.e., the effect directions of some SNPs are positive and some are negative), the proportions of the extreme values (0 and 2) for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ with $\eta = 0$ (all the SNPs only include common variants) are less than those with $\eta = 1$ (all the SNPs only consist of rare variants), irrespective of the sample

size (n). As for the qualitative trait, when $n = 2000$ and $\tau = 0.6$, the proportion of the extreme values equal to 0 for $\hat{\gamma}_{GPF}$ and the proportions of the extreme values (0 and 2) for $\hat{\gamma}_{GF}$ with $\eta = 0$ are smaller than those with $\eta = 1$, while the proportion of the extreme values equal to 2 for $\hat{\gamma}_{GPF}$ with $\eta = 0$ (12.8%) is larger than that with $\eta = 1$ (10.4%). When the trait is qualitative, $n = 500$ and $\tau = 0.6$, the results are similar to those with $n = 2000$, except that the proportion of the extreme values equal to 2 for $\hat{\gamma}_{GF}$ with $\eta = 0$ (20.0%) and that with $\eta = 1$ (19.2%) are very close to each other. In addition, the proportions of the extreme values (0 or 2) for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ have no obvious trends for other cases of different values of η and τ .

Table 1. Proportions (%) of extreme values of $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ among 500 replications.

Trait	n	η^a	τ^b	$\hat{\gamma}_{GPF}$			$\hat{\gamma}_{GF}$		
				0	2	Total	0	2	Total
Quantitative	500	0	0.6	8.6	10.6	19.2	8.6	11.8	20.4
	500	0	1	7.6	19.2	26.8	7.6	21.4	29.0
	500	0.4	0.6	9.6	8.2	17.8	9.6	10.6	20.2
	500	0.4	1	11.2	16.0	27.2	11.2	21.2	32.4
	500	1	0.6	13.4	11.8	25.2	13.4	15.0	28.4
	500	1	1	9.0	9.0	18.0	9.0	15.8	24.8
	2000	0	0.6	5.2	6.0	11.2	5.2	6.2	11.4
	2000	0	1	5.0	9.4	14.4	5.0	9.6	14.6
	2000	0.4	0.6	5.6	4.6	10.2	5.6	5.0	10.6
	2000	0.4	1	6.4	10.8	17.2	6.4	11.2	17.6
	2000	1	0.6	9.8	7.0	16.8	9.8	7.2	17.0
	2000	1	1	1.4	12.2	13.6	1.4	13.8	15.2
Qualitative	500	0	0.6	19.6	12.8	32.4	19.6	20.0	39.6
	500	0	1	23.8	17.0	40.8	23.8	20.4	44.2
	500	0.4	0.6	18.8	12.8	31.6	18.8	22.0	40.8
	500	0.4	1	29.2	10.0	39.2	29.2	19.2	48.4
	500	1	0.6	22.0	9.0	31.0	22.0	19.2	41.2
	500	1	1	27.8	0.6	28.4	27.8	7.8	35.6
	2000	0	0.6	9.4	12.8	22.2	9.4	14.6	24.0
	2000	0	1	8.0	19.4	27.4	8.0	21.4	29.4
	2000	0.4	0.6	14.6	10.8	25.4	14.6	13.2	27.8
	2000	0.4	1	13.4	16.4	29.8	13.4	20.0	33.4
	2000	1	0.6	11.8	10.4	22.2	11.8	15.4	27.2
	2000	1	1	16.2	5.0	21.2	16.2	13.0	29.2

^a Proportion of rare variants among all the SNPs; ^b proportion of the SNPs with positive effects among all the SNPs.

The MSEs of the four point estimates ($\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$) are listed in Table 2. From Table 2, we can see that the MSEs of $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ are smaller than those of $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$, and the MSE of $\hat{\gamma}_{GBN}$ is the smallest. When the sample size increases or the trait turns from qualitative to quantitative, the MSEs of these four point estimates decrease significantly. In general, the MSEs of the four point estimates gradually become larger when η changes from 0, 0.4 to 1 (i.e., higher proportion of rare variants) and other parameters are kept unchanged, except for the case when the trait is quantitative, $n = 500$ and $\tau = 1$. On the other hand, the MSEs of the four point estimates with $\tau = 0.6$ (i.e., the effect directions of some SNPs are positive and some are negative) are smaller than those with $\tau = 1$ (i.e., all the SNP effects are positive), when other parameters are fixed.

Table 2. Mean squared errors of $\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ among 500 replications.

Trait	n	η^a	τ^b	$\hat{\gamma}_{GBN}$	$\hat{\gamma}_{GBU}$	$\hat{\gamma}_{GPF}$	$\hat{\gamma}_{GF}$
Quantitative	500	0	0.6	0.0976	0.1022	0.1236	0.1287
	500	0	1	0.1409	0.1601	0.2344	0.2549
	500	0.4	0.6	0.1335	0.1395	0.1579	0.1633
	500	0.4	1	0.1953	0.2248	0.3008	0.3601
	500	1	0.6	0.1414	0.1592	0.2079	0.2363
	500	1	1	0.1623	0.1703	0.2690	0.3475
	2000	0	0.6	0.0359	0.0379	0.0403	0.0405
	2000	0	1	0.0541	0.0642	0.0793	0.0805
	2000	0.4	0.6	0.0480	0.0512	0.0555	0.0558
	2000	0.4	1	0.0755	0.0773	0.0922	0.0959
	2000	1	0.6	0.0481	0.0509	0.0578	0.0591
	2000	1	1	0.0687	0.0727	0.0962	0.1160
Qualitative	500	0	0.6	0.2765	0.3382	0.4849	0.5503
	500	0	1	0.3100	0.4038	0.5286	0.5788
	500	0.4	0.6	0.3320	0.4087	0.5785	0.6344
	500	0.4	1	0.3826	0.4700	0.6416	0.7254
	500	1	0.6	0.3405	0.4329	0.5915	0.6369
	500	1	1	0.7519	0.7673	1.0190	1.0193
	2000	0	0.6	0.1207	0.1367	0.1595	0.1668
	2000	0	1	0.1362	0.1503	0.2133	0.2306
	2000	0.4	0.6	0.1320	0.1492	0.1937	0.2090
	2000	0.4	1	0.2168	0.2460	0.3347	0.3647
	2000	1	0.6	0.1431	0.1615	0.2144	0.2364
	2000	1	1	0.3163	0.3263	0.4684	0.5145

^a Proportion of rare variants among all the SNPs; ^b proportion of the SNPs with positive effects among all the SNPs.

Figures 1, 2 and S1–S6 are the scatter plots of the four point estimates against the true values of γ under different simulation settings. These figures can more intuitively compare the performances of the four point estimates. For example, Figures 1 and 2 are the scatter plots of the four point estimates against the true values of γ for the quantitative trait with $n = 500$, and $\tau = 0.6$ and 1, respectively. In each figure, subplots (a)–(d) (four subplots in the first row) are respectively the scatter plots of $\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ with $\eta = 0$; subplots (e)–(h) (four subplots in the second row) and subplots (i)–(l) (four subplots in the third row) are the corresponding scatter plots with $\eta = 0.4$ and 1, respectively. By comparing the four subplots in the same row of each figure, we find that the two point estimates ($\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$) obtained by the Bayesian methods are closer to the true values of γ , and both perform better than $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$. On the other hand, note that the distribution of the true value of γ is $U(0, 2)$, and it can be seen from the figures that the distributions of $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ are more uniform, while the distributions of $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ are skewed towards the extreme values (0 and 2). Meanwhile, by respectively comparing subplots (a), (e) and (i) for $\hat{\gamma}_{GBN}$ with subplots (b), (f) and (j) for $\hat{\gamma}_{GBU}$, there is a little greater dispersion for $\hat{\gamma}_{GBU}$ than $\hat{\gamma}_{GBN}$. In addition, from subplots (c), (g) and (k) for $\hat{\gamma}_{GPF}$ and subplots (d), (h) and (l) for $\hat{\gamma}_{GF}$, we observe that there exist many extreme point estimates for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ (represented by the blue points). Moreover, the scatter plots for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ provide the additional information that most of the extreme point estimates generally occur when the true values of γ are less than 0.5 or greater than 1.5. Further, by comparing the subplots in different rows of each figure when $\tau = 0.6$ (Figures 1, S1, S3 and S5), i.e., η changing from 0, 0.4 to 1, the dispersions of the four point estimates generally increase, indicating that, in general, the MSEs of the four point estimates become larger, which are consistent with the results in Table 2. The numbers of the blue points in subplots (c) and (d) with $\eta = 0$ are much less than those in subplots (k) and (l) with $\eta = 1$, respectively. However, for those figures with $\tau = 1$ (Figures 2, S2, S4 and S6), there is no obvious trend for the number of the blue points. Compared to Figure 1 ($\tau = 0.6$), the agreements between the four point

estimates and the true values of γ in Figure 2 ($\tau = 1$) are worse, which can also be seen in other figures (Figures S1, S3 and S5 vs. Figures S2, S4 and S6, respectively). Observing Figure 2, we find that the four point estimation methods may underestimate γ when $\tau = 1$. Finally, these four point estimation methods perform better for the quantitative trait than for the qualitative trait (Figures 1, 2, S1 and S2 vs. Figures S3–S6, respectively), and when the sample size increases (Figures S1, S2, S5 and S6 vs. Figures 1, 2, S3 and S4, respectively).

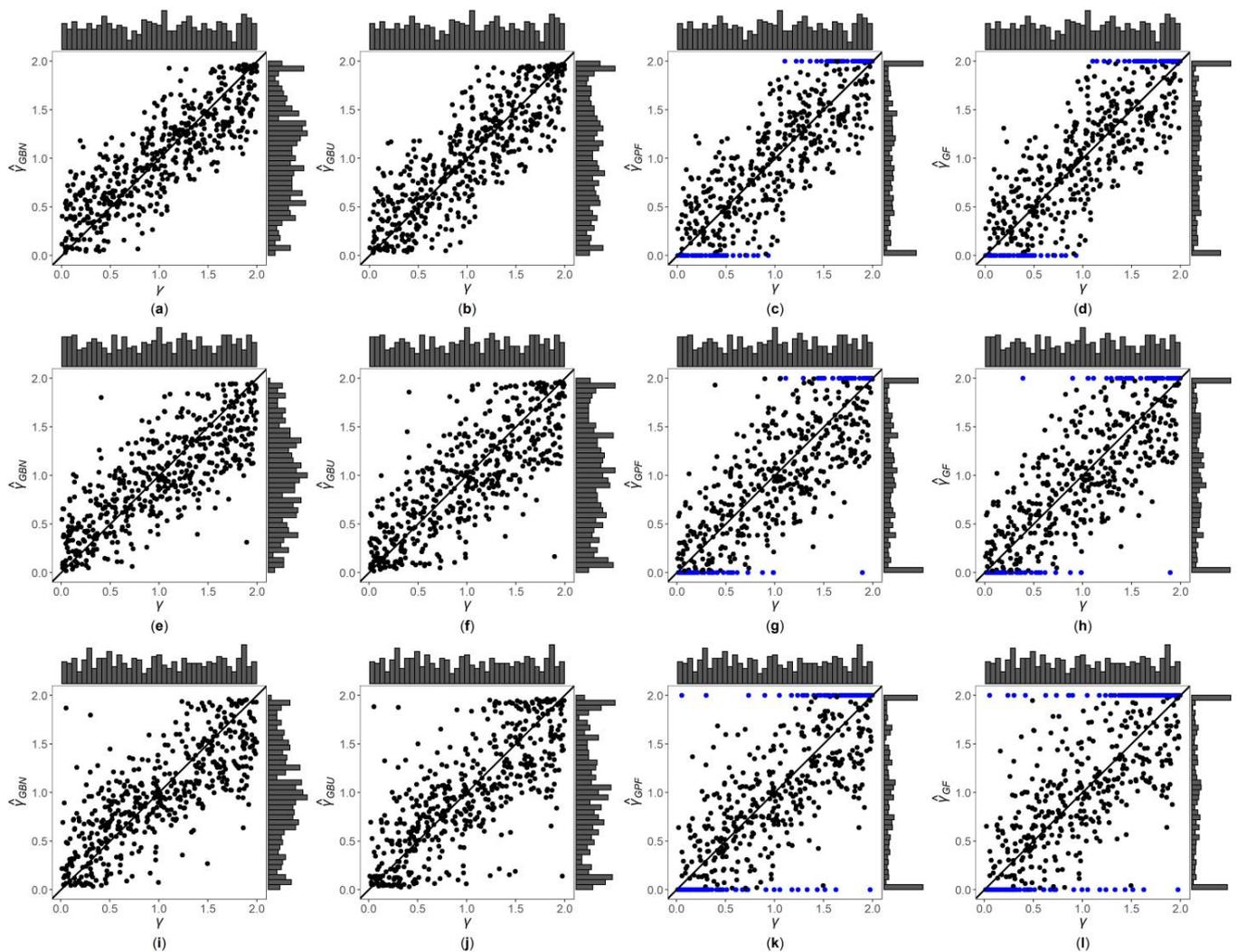


Figure 1. Scatter plots of point estimates of γ against true values of γ for quantitative trait with $n = 500$ and $\tau = 0.6$. The blue points represent the extreme values (0 or 2). (a) $\hat{\gamma}_{GBN}$ with $\eta = 0$; (b) $\hat{\gamma}_{GBU}$ with $\eta = 0$; (c) $\hat{\gamma}_{GPF}$ with $\eta = 0$; (d) $\hat{\gamma}_{GF}$ with $\eta = 0$; (e) $\hat{\gamma}_{GBN}$ with $\eta = 0.4$; (f) $\hat{\gamma}_{GBU}$ with $\eta = 0.4$; (g) $\hat{\gamma}_{GPF}$ with $\eta = 0.4$; (h) $\hat{\gamma}_{GF}$ with $\eta = 0.4$; (i) $\hat{\gamma}_{GBN}$ with $\eta = 1$; (j) $\hat{\gamma}_{GBU}$ with $\eta = 1$; (k) $\hat{\gamma}_{GPF}$ with $\eta = 1$; (l) $\hat{\gamma}_{GF}$ with $\eta = 1$.

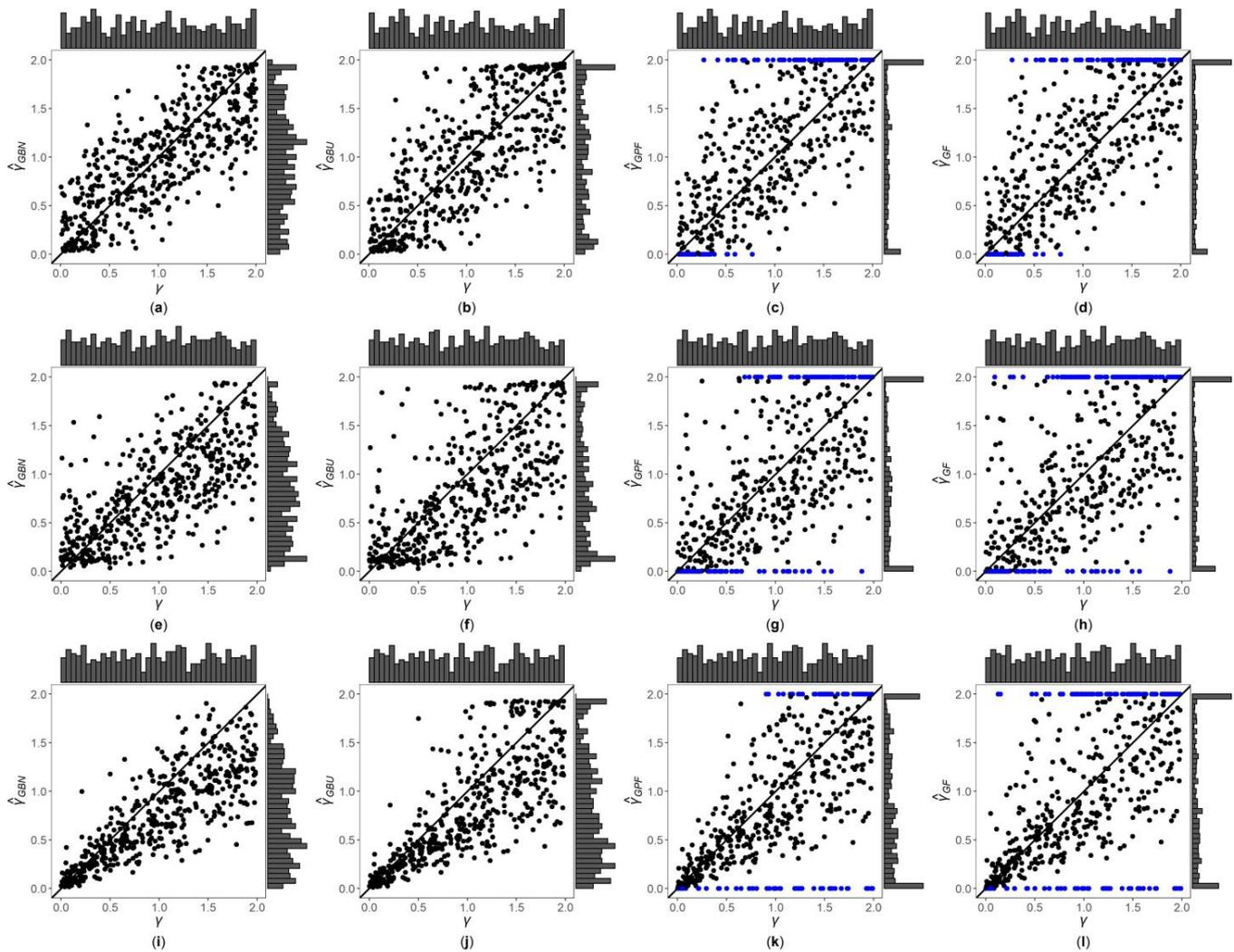


Figure 2. Scatter plots of point estimates of γ against true values of γ for quantitative trait with $n = 500$ and $\tau = 1$. The blue points represent the extreme values (0 or 2). (a) $\hat{\gamma}_{GBN}$ with $\eta = 0$; (b) $\hat{\gamma}_{GBU}$ with $\eta = 0$; (c) $\hat{\gamma}_{GPF}$ with $\eta = 0$; (d) $\hat{\gamma}_{GF}$ with $\eta = 0$; (e) $\hat{\gamma}_{GBN}$ with $\eta = 0.4$; (f) $\hat{\gamma}_{GBU}$ with $\eta = 0.4$; (g) $\hat{\gamma}_{GPF}$ with $\eta = 0.4$; (h) $\hat{\gamma}_{GF}$ with $\eta = 0.4$; (i) $\hat{\gamma}_{GBN}$ with $\eta = 1$; (j) $\hat{\gamma}_{GBU}$ with $\eta = 1$; (k) $\hat{\gamma}_{GPF}$ with $\eta = 1$; (l) $\hat{\gamma}_{GF}$ with $\eta = 1$.

Table 3 displays the EPs, NPs and DPs of the PF and Fieller's methods. From Table 3, we observe that the EPs of the PF method are generally smaller than, or equal to, those of the Fieller's method, except for the quantitative trait with $n = 500$, $\eta = 0.4$ or 1, and $\tau = 1$, and the qualitative trait with $n = 500$ or 2000, $\eta = 0.4$ or 1, and $\tau = 1$. However, the NPs of the PF method are always smaller than, or equal to, those of the Fieller's method. Note that when we use the PF and Fieller's methods to calculate the CIs of γ , we need to truncate the CIs by the interval $[0, 2]$. As such, compared to the Fieller's method, the PF method can get shorter CIs, which means that the PF method reduces the possibility of the truncated CIs being the noninformative intervals. On the other hand, if the CIs before the truncation are disjoint from the interval $[0, 2]$, the PF method will increase the possibility that the truncated CIs are empty sets, which is the reason why the PF method may have bigger EPs than the Fieller's method in some scenarios. In addition, all the DPs of the PF method are equal to 0. This is because we consider the penalty parameter $\lambda = Z_{1-\alpha/2}^2/4$, and the CIs derived by the PF method are always continuous. With increase of the sample size, the NPs of the PF and Fieller's methods and the DPs of the Fieller's method become smaller. Moreover, under the same simulation settings, the NPs of both methods, and the DPs of

the Fieller’s method, for the quantitative trait are less than those for the qualitative trait. Under the situation that $\tau = 0.6$, when η changes from 0, 0.4 to 1 and other parameters are kept unchanged, the EPs of both methods have no obvious trends, while the NPs of both methods and the DPs of the Fieller’s method generally become larger. As for $\tau = 1$, when η changing from 0, 0.4 to 1 and other parameters being fixed, the EPs of the PF method appear larger except for the quantitative trait and $n = 2000$, while the DPs of the Fieller’s method are relatively stable, and the NPs of the PF and Fieller’s methods show a trend of first increasing and then decreasing on most occasions. On the other hand, when other parameters are fixed, the EPs and NPs of the PF and Fieller’s methods with $\tau = 0.6$ are smaller than those with $\tau = 1$ in most cases, and the DPs of the Fieller’s method with $\tau = 0.6$ are larger than or equal to those with $\tau = 1$.

Table 3. Proportions (%) of empty sets (EPs), noninformative intervals (NPs), and discontinuous intervals (DPs) of PF and Fieller’s methods among 500 replications.

Trait	n	η^a	τ^b	PF			Fieller		
				EP	NP	DP	EP	NP	DP
Quantitative	500	0	0.6	0.0	7.2	0.0	0.8	16.6	1.0
	500	0	1	0.0	19.0	0.0	0.2	21.8	0.0
	500	0.4	0.6	0.2	10.2	0.0	0.2	22.2	0.4
	500	0.4	1	1.4	27.2	0.0	0.4	33.8	0.0
	500	1	0.6	0.0	14.8	0.0	0.8	31.2	2.8
	500	1	1	6.8	3.6	0.0	1.0	3.6	0.0
	2000	0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
	2000	0	1	0.6	0.0	0.0	0.6	0.2	0.0
	2000	0.4	0.6	0.0	0.0	0.0	0.2	0.0	0.0
	2000	0.4	1	0.0	2.4	0.0	0.4	4.2	0.0
	2000	1	0.6	0.0	0.2	0.0	0.0	2.2	0.0
2000	1	1	0.2	0.2	0.0	0.2	0.6	0.0	
Qualitative	500	0	0.6	0.0	43.4	0.0	0.6	65.0	2.8
	500	0	1	1.4	58.2	0.0	1.4	64.4	0.0
	500	0.4	0.6	0.0	45.4	0.0	0.0	68.2	4.0
	500	0.4	1	1.8	55.2	0.0	1.2	64.0	1.0
	500	1	0.6	0.0	44.0	0.0	0.4	75.0	3.6
	500	1	1	10.4	53.4	0.0	0.0	54.2	0.0
	2000	0	0.6	0.0	10.8	0.0	0.4	19.8	0.6
	2000	0	1	0.4	20.8	0.0	0.6	25.2	0.0
	2000	0.4	0.6	0.0	14.4	0.0	0.2	27.0	1.4
	2000	0.4	1	1.2	26.2	0.0	0.6	31.0	0.2
	2000	1	0.6	0.0	19.0	0.0	0.2	36.6	2.2
2000	1	1	12.4	4.8	0.0	0.2	16.0	0.0	

^a Proportion of rare variants among all the SNPs; ^b proportion of the SNPs with positive effects among all the SNPs.

The CPs, W_{mean} and W_{median} of the GBN, GBU, PF and Fieller’s methods are displayed in Table 4, and the corresponding W_{sd} and W_{iqr} are given in Table 5. Table 4 demonstrates that, for the quantitative trait, the CPs of the GBN, GBU and Fieller’s methods are controlled around 95%. However, when $n = 500$, $\eta = 1$ and $\tau = 1$, the CP of the PF method is underestimated (87.8%). As the sample size increases to 2000 and other parameters remain unchanged, the CP of the PF method is 96.6%. For the qualitative trait, when $n = 500$, the CPs of the GBN, GBU and PF methods are underestimated in most situations. With the increase of the sample size to 2000, the CPs of these three methods generally increase to be around 95%, but the CPs when $\eta = 1$ and $\tau = 1$ are still underestimated. Thus, for this simulation setting, we conduct an additional simulation study with larger sample sizes (3000 and 4000), and the corresponding results are presented in Table S1. It is shown in Table S1 that the CPs of these three methods are closer to 95% when the sample size continues to increase. This is explainable by the fact that qualitative traits generally require

larger samples to achieve the same CPs than quantitative traits. In addition, we can see from Table 4 that the Fieller’s method has higher CPs under various simulation settings for the qualitative trait. However, according to Table 3, when the trait is qualitative, the NPs of the Fieller’s method are relatively high, which means that many CIs obtained by the Fieller’s method are the noninformative intervals (i.e., [0, 2]). This may explain why the CPs of the Fieller’s method are on the high side. Further, from Tables 4 and 5, the W_{mean} , W_{median} , W_{sd} and W_{iqr} of the GBN and GBU methods are smaller than those of the PF and Fieller’s methods in most situations. The GBN method has the smallest W_{mean} , W_{median} and W_{iqr} in most cases, and it also has the smallest W_{sd} under all the simulated settings. As can be seen from Table 4, when the trait is qualitative and $n = 500$, the W_{median} ’s of the Fieller’s method are all 2, which indicates that in this case, more than half of the CIs based on the Fieller’s method are the noninformative intervals. This is consistent with the results of the NPs in Table 3. When the sample size increases, or the trait turns from qualitative into quantitative, the W_{mean} ’s and W_{median} ’s of the four interval estimation methods greatly decrease. However, for the W_{sd} and W_{iqr} , there are different trends in some situations. For example, when the trait is qualitative, the W_{sd} ’s and W_{iqr} ’s of the four methods become larger in most cases as the sample size increases. Note that the widths of the intervals obtained by the four methods are closer to 2 and the corresponding variation will be smaller when $n = 500$. With the sample size increasing, the widths of the intervals gradually decrease and the corresponding variation appears larger, which may cause the bigger W_{sd} and W_{iqr} .

Table 4. Coverage probability (CP, in %), W_{mean} and W_{median} of GBN, GBU, PF and Fieller’s methods among 500 replications.

Trait	n	η^a	τ^b	CP				W_{mean}				W_{median}				
				GBN	GBU	PF	Fieller	GBN	GBU	PF	Fieller	GBN	GBU	PF	Fieller	
Quantitative	500	0	0.6	96.2	95.8	95.8	95.2	1.2357	1.2524	1.2338	1.2674	1.2439	1.2571	1.2072	1.2328	
	500	0	1	96.2	97.0	97.8	95.8	1.3536	1.3695	1.4593	1.4375	1.3959	1.4152	1.4749	1.5010	
	500	0.4	0.6	95.0	95.6	95.6	96.2	1.2663	1.2862	1.2815	1.3305	1.2662	1.2973	1.2449	1.2682	
	500	0.4	1	95.6	96.6	94.2	95.6	1.4718	1.4977	1.5555	1.5887	1.5158	1.5571	1.6734	1.6888	
	500	1	0.6	96.2	96.6	95.4	94.2	1.3457	1.3689	1.3363	1.3767	1.4001	1.4490	1.2991	1.3461	
	500	1	1	94.6	95.4	87.8	93.8	1.2841	1.2983	1.2918	1.3827	1.3135	1.3316	1.4814	1.4465	
	2000	0	0.6	94.6	94.2	94.8	94.6	0.7216	0.7258	0.7377	0.7413	0.7149	0.7230	0.7406	0.7425	
	2000	0	1	95.8	96.0	95.8	94.2	0.8934	0.8946	0.9184	0.9249	0.9068	0.9035	0.9396	0.9469	
	2000	0.4	0.6	94.0	95.4	94.4	94.6	0.7895	0.7958	0.8067	0.8152	0.7770	0.7850	0.8087	0.8124	
	2000	0.4	1	95.6	96.2	97.4	96.2	1.0439	1.0505	1.0800	1.0950	1.0415	1.0420	1.0857	1.0828	
	2000	1	0.6	95.8	96.6	96.2	96.2	0.8284	0.8325	0.8406	0.8539	0.7933	0.7974	0.8211	0.8190	
	2000	1	1	95.4	95.6	96.6	95.0	0.9483	0.9560	0.9750	1.0066	0.9988	0.9982	1.0294	1.0527	
	Qualitative	500	0	0.6	92.6	94.2	95.4	95.0	1.6289	1.6667	1.6720	1.7236	1.7202	1.7749	1.8354	2.0000
		500	0	1	94.0	96.0	90.0	94.8	1.6575	1.6934	1.7053	1.7578	1.7387	1.7848	2.0000	2.0000
500		0.4	0.6	93.0	94.6	93.6	96.0	1.6782	1.7193	1.6986	1.7668	1.7516	1.8033	1.8721	2.0000	
500		0.4	1	93.0	94.8	84.6	94.0	1.6775	1.7154	1.6108	1.7788	1.7360	1.7830	2.0000	2.0000	
500		1	0.6	92.6	94.8	93.0	96.0	1.7318	1.7742	1.6981	1.7965	1.7837	1.8283	1.8659	2.0000	
500		1	1	77.0	74.4	74.2	99.4	1.3896	1.3523	1.4088	1.8704	1.4854	1.4788	2.0000	2.0000	
2000		0	0.6	94.6	95.8	96.6	95.0	1.2519	1.2686	1.2531	1.2774	1.2388	1.2710	1.1933	1.2177	
2000		0	1	97.0	96.8	97.2	95.6	1.3832	1.4010	1.4869	1.4734	1.4162	1.4502	1.5404	1.5295	
2000		0.4	0.6	96.2	96.6	96.8	95.2	1.3468	1.3682	1.3443	1.3908	1.4163	1.4514	1.3443	1.3965	
2000		0.4	1	95.0	95.8	93.6	95.4	1.4765	1.5029	1.5565	1.5781	1.5153	1.5623	1.6985	1.6909	
2000		1	0.6	96.4	96.8	94.2	95.0	1.4216	1.4488	1.3842	1.4516	1.5241	1.5772	1.3174	1.4640	
2000		1	1	89.8	89.6	84.6	98.6	1.3833	1.3967	1.3764	1.6143	1.4576	1.4936	1.7096	1.6751	

^a Proportion of rare variants among all the SNPs; ^b proportion of the SNPs with positive effects among all the SNPs.

Table 5. W_{sd} and W_{iqr} of GBN, GBU, PF and Fieller’s methods among 500 replications.

Trait	n	η^a	τ^b	W_{sd}				W_{iqr}			
				GBN	GBU	PF	Fieller	GBN	GBU	PF	Fieller
Quantitative	500	0	0.6	0.3309	0.3619	0.4066	0.4851	0.5036	0.5697	0.5403	0.6862
	500	0	1	0.3020	0.3364	0.4429	0.4948	0.4613	0.5274	0.6959	0.7625
	500	0.4	0.6	0.3312	0.3624	0.4198	0.4868	0.5334	0.5910	0.5862	0.8516
	500	0.4	1	0.2631	0.2917	0.4881	0.4498	0.3741	0.4244	0.6279	0.6390
	500	1	0.6	0.3585	0.3890	0.4492	0.5487	0.5765	0.6382	0.7346	1.0386
	500	1	1	0.2563	0.2891	0.6080	0.4568	0.3086	0.3487	0.7633	0.5616
	2000	0	0.6	0.1961	0.2118	0.2251	0.2350	0.2369	0.2684	0.2520	0.2564
	2000	0	1	0.2623	0.2874	0.3381	0.3514	0.3609	0.4000	0.4281	0.4336
	2000	0.4	0.6	0.2214	0.2419	0.2500	0.2723	0.2874	0.3203	0.2952	0.3094
	2000	0.4	1	0.3084	0.3386	0.4154	0.4447	0.3816	0.4537	0.5550	0.5927
	2000	1	0.6	0.2720	0.2941	0.3049	0.3455	0.3455	0.3840	0.3589	0.3830
	2000	1	1	0.3184	0.3442	0.4515	0.4661	0.3969	0.4519	0.6674	0.6647
Qualitative	500	0	0.6	0.2535	0.2727	0.3893	0.4565	0.2800	0.2841	0.5975	0.4816
	500	0	1	0.2005	0.2194	0.5012	0.4440	0.2140	0.2336	0.4291	0.3656
	500	0.4	0.6	0.1998	0.2129	0.3599	0.4105	0.2059	0.1966	0.5632	0.3726
	500	0.4	1	0.1611	0.1782	0.6086	0.4317	0.1748	0.1658	0.6470	0.2657
	500	1	0.6	0.1553	0.1632	0.3705	0.4144	0.1162	0.1055	0.5430	0.0354
	500	1	1	0.2933	0.3707	0.8749	0.2417	0.3847	0.5508	1.9212	0.1898
	2000	0	0.6	0.3501	0.3824	0.4415	0.5142	0.5624	0.6511	0.6639	0.8792
	2000	0	1	0.2936	0.3261	0.4417	0.4911	0.4447	0.5120	0.7372	0.8589
	2000	0.4	0.6	0.3518	0.3824	0.4411	0.5098	0.5682	0.6366	0.6747	1.0159
	2000	0.4	1	0.2487	0.2780	0.4936	0.4545	0.3529	0.3963	0.6457	0.6883
	2000	1	0.6	0.3456	0.3758	0.4350	0.5209	0.5482	0.6068	0.7529	0.9691
	2000	1	1	0.2762	0.3174	0.7095	0.3578	0.2032	0.2535	0.7992	0.3615

^a Proportion of rare variants among all the SNPs; ^b proportion of the SNPs with positive effects among all the SNPs.

In the case of $\tau = 0.6$, the four methods have larger W_{mean} 's and W_{median} 's in most cases when η changes from 0, 0.4 to 1, while for the scenario of $\tau = 1$, the four methods show a trend of first increasing and then decreasing on most occasions, except that the W_{mean} 's and W_{median} 's of the Fieller's method are gradually larger for the qualitative trait. When the trait is quantitative and $\tau = 0.6$, the W_{sd} 's and W_{iqr} 's of the four methods become larger with η increasing from 0, 0.4 to 1, irrespective of the sample size. When the trait is qualitative, $n = 500$ and $\tau = 0.6$, as η is bigger, the W_{sd} 's and W_{iqr} 's of the four methods generally are smaller, while when $n = 2000$, the W_{sd} 's of the four methods and the W_{iqr} 's of the GBN and GBU methods are relatively stable, and the W_{iqr} 's of the PF and Fieller's methods generally become larger. For the quantitative trait with $n = 500$ and $\tau = 1$, with the increase of η , the W_{sd} 's and W_{iqr} 's of the GBN, GBU and Fieller's methods appear smaller and those of the PF method are larger in most situations, while in the case of $n = 2000$, the four methods usually have larger W_{sd} 's and W_{iqr} 's. When the trait is qualitative and $\tau = 1$, with η increasing, the W_{sd} 's and W_{iqr} 's of the GBN and GBU methods present a tendency of first decreasing and then increasing on most occasions, while those of the PF method are larger in most cases, and those of the Fieller's method become smaller, irrespective of the sample size. On the other hand, when other parameters are fixed, the W_{mean} 's and W_{median} 's of the four methods with $\tau = 0.6$ are smaller than those with $\tau = 1$, except for the W_{mean} 's of the GBN, GBU and PF methods and the W_{median} 's of the GBN and GBU methods for the quantitative trait with $n = 500$ and $\eta = 1$, and those for the qualitative trait with $n = 500$ or 2000, and $\eta = 1$. Under the scenarios where η is kept unchanged, the W_{sd} 's and W_{iqr} 's of the GBN, GBU and Fieller's methods with $\tau = 0.6$ are generally larger than those with $\tau = 1$ for the quantitative trait with $n = 500$, and the qualitative trait with $n = 500$ or 2000, while there are different trends for the quantitative trait with $n = 2000$. In addition, the W_{sd} 's and W_{iqr} 's of the PF method with $\tau = 0.6$ generally are smaller than those with $\tau = 1$, when other parameters are fixed.

Figures 3, 4 and S7–S12 are the scatter plots of the widths of the 95% HPDIs or CIs obtained by the four interval estimation methods (GBN, GBU, PF and Fieller) against the true values of γ under different simulation settings. We can clearly observe the distributions of the widths of the HPDIs or CIs through these figures. For example, Figures 3 and 4 are the scatter plots of the widths of the HPDIs or CIs against the true values of γ for the quantitative trait with $n = 500$, and $\tau = 0.6$ and 1, respectively. In each figure, subplots (a)–(d) (four subplots in the first row) are respectively the scatter plots for the GBN, GBU, PF and Fieller's methods with $\eta = 0$; subplots (e)–(h) (four subplots in the second row) and subplots (i)–(l) (four subplots in the third row) are the corresponding scatter plots with $\eta = 0.4$ and 1, respectively. It can be seen from the four subplots in the same row of each figure that the distributions of the widths of the HPDIs for the GBN and GBU methods are similar, and both have smaller dispersions than those of the CIs for the PF and Fieller's methods. Furthermore, these figures display that the distributions of the interval widths for the PF and Fieller's methods are greatly more skewed towards 2 than the GBN and GBU methods. We respectively compare subplots (a), (e) and (i) for the GBN method with subplots (b), (f) and (j) for the GBU method and find that the dispersions of the widths of the HPDIs for the GBN method are slightly smaller than the GBU method. Additionally, subplots (c), (g) and (k) for the PF method, and subplots (d), (h) and (l) for the Fieller's method, show that the PF and Fieller's methods may yield empty sets or noninformative intervals (displayed by the blue points), and the Fieller's method may also get discontinuous intervals (shown by the orange points). By comparing the subplots in different rows of each figure (Figures 3 and S7) when the trait is quantitative and $\tau = 0.6$, the dispersions of the widths of the HPDIs or CIs become slightly larger as η changing from 0, 0.4 to 1, and it can also be seen from Figure 3 that the distributions of the interval widths are a little more skewed towards 2. On the other hand, when the trait is qualitative with $\tau = 0.6$ (Figures S9 and S11), there are no obvious trends in the dispersions of the interval widths, except that their distributions are more skewed towards 2. However, under the situation that $\tau = 1$ (Figures 4, S8, S10 and S12), the points in these figures become less discrete in most cases when η increases, and the overall widths of the four interval estimation methods also somewhat decrease, except for the scenarios where the trait is quantitative and $n = 2000$, and the trait is qualitative and $n = 500$. Further, by comparing the figures for different values of τ (Figures 3, S7, S9 and S11 vs. Figures 4, S8, S10 and S12, respectively), it can be found that the overall widths of the HPDIs or the CIs obtained by the four interval estimation methods with $\tau = 0.6$ are generally smaller than those with $\tau = 1$, except for those with $\eta = 1$. Lastly, as the trait turns from qualitative into quantitative (Figures S9–S12 vs. Figures 3, 4, S7 and S8, respectively) or the sample size increases (Figures 3, 4, S9 and S10 vs. Figures S7, S8, S11 and S12, respectively), the performances of the four interval estimation methods are greatly improved.

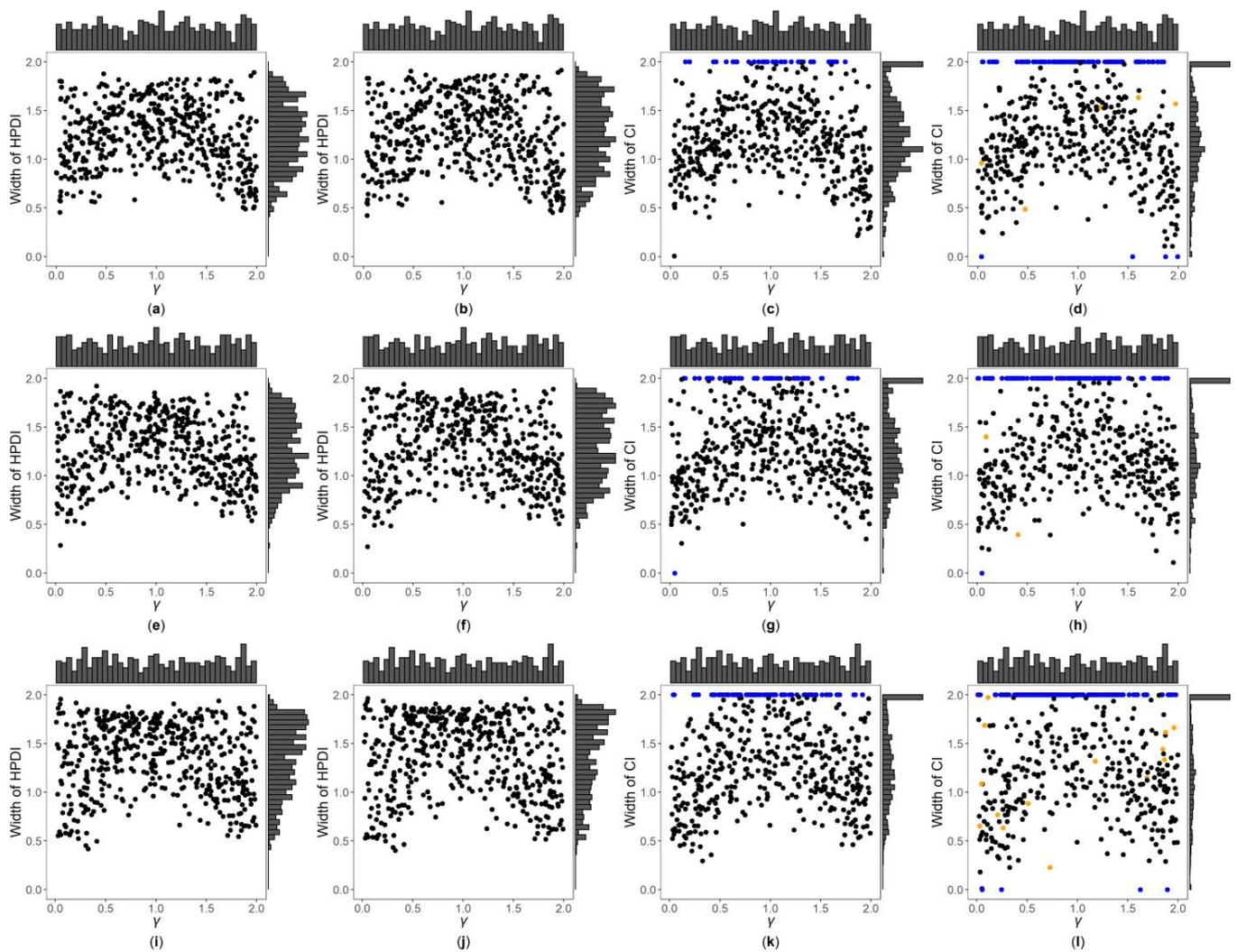


Figure 3. Widths of highest posterior density intervals (HPDIs) or confidence intervals (CIs) of GBN, GBU, PF and Fieller's methods against true values of γ for quantitative trait with $n = 500$ and $\tau = 0.6$. The blue points represent the widths of the empty sets or the noninformative intervals, and the orange points represent the widths of the discontinuous intervals. (a) GBN with $\eta = 0$; (b) GBU with $\eta = 0$; (c) PF with $\eta = 0$; (d) Fieller with $\eta = 0$; (e) GBN with $\eta = 0.4$; (f) GBU with $\eta = 0.4$; (g) PF with $\eta = 0.4$; (h) Fieller with $\eta = 0.4$; (i) GBN with $\eta = 1$; (j) GBU with $\eta = 1$; (k) PF with $\eta = 1$; (l) Fieller with $\eta = 1$.

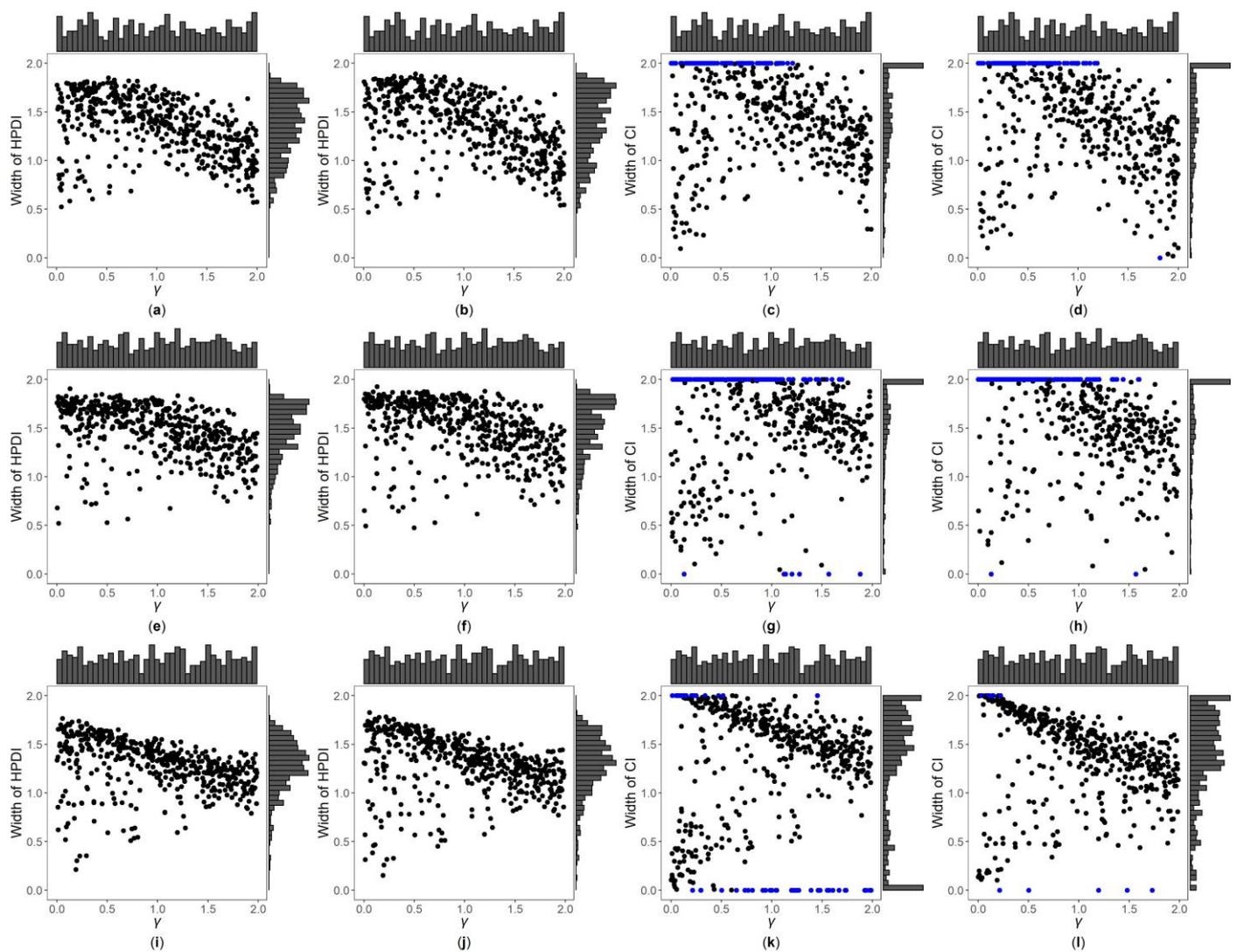


Figure 4. Widths of HPDIs or CIs of GBN, GBU, PF and Fieller’s methods against true values of γ for quantitative trait with $n = 500$ and $\tau = 1$. The blue points represent the widths of the empty sets or the noninformative intervals. (a) GBN with $\eta = 0$; (b) GBU with $\eta = 0$; (c) PF with $\eta = 0$; (d) Fieller with $\eta = 0$; (e) GBN with $\eta = 0.4$; (f) GBU with $\eta = 0.4$; (g) PF with $\eta = 0.4$; (h) Fieller with $\eta = 0.4$; (i) GBN with $\eta = 1$; (j) GBU with $\eta = 1$; (k) PF with $\eta = 1$; (l) Fieller with $\eta = 1$.

3.3. Application to MCTFR Data

The MCTFR Genome-Wide Association Study of Behavioral Disinhibition is a family-based epidemiological study of substance abuse and related psychopathology. The dataset can be made available from the database of Genotypes and Phenotypes with accession numbers 86747-6 and 95621-5 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000620.v1.p1, accessed on 5 January 2022). The dataset includes 2183 families and 7377 participants (3831 female subjects and 3546 male subjects). Among them, only 5960 subjects have both the phenotypic and genotypic data, while others do not have phenotypic data or do not have genotypic data. There are five quantitative traits included in this dataset: the nicotine composite score, the alcohol consumption composite score, the alcohol dependence composite score, the illicit drug composite score and the non-substance use related behavioral disinhibition composite score. To avoid the influence of family structure on the results, we exclude offspring from the real data application. At the same time, we only need the information of female subjects, so we also exclude male subjects from the analysis. Meanwhile, 12,354 SNPs on the X chromosome are included in the dataset. We use

the following quality control criteria to filter the SNPs [48,49]: (1) genotype call rate being less than 99%, (2) MAF being smaller than 1×10^{-5} , (3) individual call rate being below 99%, and (4) the p value of the Hardy–Weinberg equilibrium test being less than 1×10^{-6} . Finally, 1994 female subjects and 12,342 SNPs on the X chromosome are utilized to conduct real data analysis. Since we estimate the degree of the skewness of the XCI based on genes, we first need to find the genes which each SNP belongs to. Based on the GRCH38 (Genome Reference Consortium Human Genome Build 38, <https://uswest.ensembl.org/>, accessed on 25 February 2022) reference, we use the “getBM” function in the R package “biomaRt” to match the SNPs to the genes on the X chromosome [45]. As such, we find 733 matched genes, while there are some genes containing only a single SNP in the dataset. As there have been several methods available to estimate the degree of the skewness of the XCI for a single SNP, we exclude genes consisting of only one SNP. Therefore, only 493 genes are included in the subsequent analysis.

Note that estimating γ requires the genes on the X chromosome to be associated with the traits. So, we need to test if the associations between the genes and the traits exist before using our proposed methods to estimate the degree of skewness of the XCI. Notice that the five traits in the MCTFR dataset do not follow normal distributions; therefore, we use the rank-based inverse normal transformation to transform the trait data [50]. Further, to adjust the effects of other variables, we incorporate two covariates, age and year of birth, into the application [48]. Due to the fact that we only use female subjects, we still apply the adaptive sum test proposed by Iuliana et al. [35] to test for the association between each gene and each trait. Unlike other multi-locus association analysis methods, when there are both rare and common variants in a gene, the adaptive sum test still maintains high test power. We set the significance level to be $\alpha = 0.05 / (5 \times 493) = 2.03 \times 10^{-5}$ based on the Bonferroni correction. After identifying the genes associated with the traits, we calculate the four point estimates of γ ($\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$), and then use the GBN, GBU, PF and Fieller’s methods to obtain the corresponding HPDIs or CIs.

We finally identify only one gene, *TMEM47*, statistically significantly associated with the alcohol dependence composite score (p value = 2.32×10^{-6}). There are two SNPs (*rs10522027* and *rs5928615*) included in the gene. The estimated MAFs of these two SNPs are 0.1407 and 0.0998, respectively, which means that both SNPs only contain common variants. It has been confirmed that *TMEM47* is located in the NC_000023.11 region and includes three exons. Studies have shown that the gene is expressed in the bladder, adipose and 23 other tissues and found that the overexpression of *TMEM47* may induce resistance in patients to certain chemotherapy drugs [51,52]. The four point estimates ($\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$) of γ for the gene are 0.4703, 0.4547, 0.4816 and 0.4847, and the 95% HPDIs or CIs derived by the GBN, GBU, PF and Fieller’s methods are (0.0023, 1.2380), (0.0337, 1.3083), (0.0562, 1.2410) and (0.0557, 1.3896), respectively. That is to say, the point estimates are all less than 0.5, while the 95% HPDIs or CIs all contain 1, which means that the XCI pattern for *TMEM47* on the alcohol dependence composite score may be the XCI-R or the XCI-E. By comparing the interval widths of these four interval estimation methods, we find that the width of the CI obtained by the PF method is the shortest, followed by the HPDI obtained by the GBN method, and the longest is the CI yielded by the Fieller’s method.

4. Discussion

In this paper, we propose four point estimates ($\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$) and four interval estimation methods (GBN, GBU, PF and Fieller) of the degree of the skewness of the XCI for a gene (i.e., γ). Among the point estimates, $\hat{\gamma}_{GF}$ is constructed by truncating the ratio of the two regression coefficients by the interval $[0, 2]$. And, $\hat{\gamma}_{GPF}$ is obtained by choosing the penalty parameter $\lambda = Z_{1-\alpha/2}^2 / 4$, and respectively correcting the denominator and the numerator, which is also truncated by $[0, 2]$. Both the $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ are developed, based on the Bayesian theory, by considering the prior information of $\gamma \in [0, 2]$, and the corresponding prior distributions of γ are respectively a truncated normal distribution and a uniform distribution. Use of $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ can avoid the extreme point

estimates of γ (0 or 2) occurring. Among the interval estimation methods, the Fieller's method has been widely used to construct the CIs of a ratio estimate. The PF method can always get the bounded CIs by choosing an appropriate penalty parameter. The GBN and GBU methods calculate the HPDIs of the samples randomly chosen from the approximate posterior distributions of γ as the credible intervals, which can avoid empty sets, noninformative intervals (i.e., $[0, 2]$) and discontinuous intervals to occur. We conducted extensive simulation studies to compare their performances, by simulating different types of traits (quantitative and qualitative), different sample sizes ($n = 500$ and 2000), different proportions of rare variants among all the SNPs considered ($\eta = 0, 0.4$ and 1), and different proportions of the SNPs with positive effects among all the SNPs considered ($\tau = 0.6$ and 1). The simulation results showed that there may exist some extreme point estimates for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$, especially when the sample size is small or the proportion of rare variants is high. The least MSE, in most situations, is derived from $\hat{\gamma}_{GBN}$, and the MSEs of $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ are smaller than those of $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$. As for the interval estimation, the CIs derived by the Fieller's method may be empty sets, noninformative intervals and discontinuous intervals. Although the PF method can avoid discontinuous intervals, the resulting CIs can be empty sets and noninformative intervals. In addition, most of the CPs of the GBN and GBU methods can be controlled around 95%, and a larger sample size is required only when the trait is qualitative and all the SNPs are rare variants. For qualitative traits, the CPs of the PF method appear a little low when the sample size is relatively small. However, the CPs of the Fieller's method seem to be well controlled, which is due to the large proportion of noninformative intervals. The GBN method has the smallest W_{mean} , W_{median} and W_{iqr} in most situations, and the least W_{sd} under all the simulation settings. Therefore, we recommend using $\hat{\gamma}_{GBN}$ and the GBN method to estimate the degree of the XCI skewing in practical applications.

On the other hand, concerning the simulation settings and the simulation results, we further discuss the following issues. Firstly, we consider the influence of the proportion of rare variants (η) and the proportion of the SNPs with positive effects (τ) among all the SNPs in the gene under study on the estimation results. When $\tau = 0.6$ and other parameters are fixed, the proportions of the extreme values (0 and 2) for $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ with $\eta = 0$ are generally less than those with $\eta = 1$, while they have no obvious trends for other cases of different values of η and τ . In general, the MSEs of the four point estimates generally become larger as η changes from 0, 0.4 to 1 and other parameters are kept unchanged. The four point estimates with $\tau = 0.6$ always have smaller MSEs than $\tau = 1$. The changing trends of the EPs, NPs and DP of the PF and Fieller's methods with the increase of η are related to τ . Furthermore, the EPs and NPs of the PF and Fieller's methods with $\tau = 0.6$ generally are smaller than $\tau = 1$, while the DP of the Fieller's method with $\tau = 0.6$ are larger than or equal to those with $\tau = 1$. On the other hand, in the case of $\tau = 0.6$, the four interval estimation methods have larger W_{mean} 's and W_{median} 's in most cases with η changing from 0, 0.4 to 1, while for the scenario of $\tau = 1$, those of the four methods show a trend of first increasing and then decreasing on most occasions. The changing tendencies of the W_{sd} 's and W_{iqr} 's of the four methods, with η increasing, are affected by the trait type, n and τ . When other parameters are kept unchanged, the W_{mean} 's and W_{median} 's of the four methods with $\tau = 0.6$ are smaller than those with $\tau = 1$ in most cases. Besides this, the findings, by comparing the W_{sd} 's and W_{iqr} 's of the GBN, GBU and Fieller's methods for $\tau = 0.6$ with those for $\tau = 1$, are related to the trait type and n , while the W_{sd} 's and W_{iqr} 's of the PF method with $\tau = 0.6$ are generally smaller than those with $\tau = 1$. Secondly, to better evaluate the performances of the proposed methods, we set the degrees of the XCI skewing at all the SNPs in the gene to be the same in our simulation studies. For example, when we calculate the MSEs of the point estimates and the CPs of the HPDIs or the CIs, a single true value of γ for each replicate is required. However, note that there may be different degrees of the XCI skewing at different SNPs, and, actually, we can also consider this issue in our simulation studies, although we have no appropriate evaluation indexes to assess the performances of the proposed methods for this situation. Finally, when we

simulate quantitative traits, the random error ε_i is generated from the standard normal distribution, where the standard deviation (σ) is equal to 1. To further illustrate the effect of different values of σ on the estimation results, we conducted additional simulation studies with $n = 2000$ and assume that ε_i follows $N(0, 4)$, where $\sigma = 2$. The corresponding results are presented in Tables S2–S4 and Figures S13–S16. As can be seen from these tables and figures, the Bayesian methods still have obvious advantages in both the point estimation and the interval estimation. Further, the four point estimation methods, and the four interval estimation methods with $\sigma = 2$, perform worse than $\sigma = 1$.

We applied the proposed methods to the MCTFR data and identified a gene, *TMEM47*, which is statistically significantly associated with the alcohol dependence composite score. However, although the four point estimates of γ for the gene *TMEM47* on the alcohol dependence composite score are all smaller than 0.5, the corresponding 95% HPDIs or CIs all contain 1, which means that the XCI pattern for this gene may not be the XCI-S. Further, we observed that the width of the CI obtained by the PF method is the shortest, followed by the HPDI obtained by the GBN method, and the longest was the CI yielded by the Fieller's method. However, it should be noted that the PF method may not control the CP well (e.g., Table S3).

Last, but not least, there are still some issues in our proposed methods which need to be discussed. Firstly, we would like to further discuss the effect of the truncation by the interval $[0, 2]$ on the point estimation and the interval estimation of γ . When we use the $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ to estimate γ , both of them are truncated by $[0, 2]$. If the point estimates before the truncation ($\hat{\gamma}^*$ and $\hat{\gamma}$) lie outside $[0, 2]$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ become the extreme values (0 or 2). Correspondingly, when using the PF and Fieller's methods to construct the CIs of γ , it is easy to obtain empty sets or noninformative intervals. On the contrary, the Bayesian methods can avoid extreme point estimates, empty sets and noninformative intervals by specifying the appropriate prior distributions of γ and making full use of the constraint condition of $\gamma \in [0, 2]$. In addition, the extreme point estimate of 0 (2) means that the XCI is completely skewed towards the minor alleles (major alleles) at all the SNPs in a gene. However, these phenomena are not common in practice [2]. Meanwhile, it should be noted that empty sets and noninformative intervals are not informative, and the discontinuous CIs are also not useful, because the discontinuous CIs cannot be clearly explained in practice. Secondly, since the XCI patterns at different SNPs may be different, our estimated $\hat{\gamma}$ is just the mean degree of the skewness of the XCI over all the SNPs in the gene under study, and we cannot obtain the degree of the skewness of the XCI for each SNP in this gene. Meanwhile, in the process of estimating γ , the target allele is the minor one at each SNP, and it is not possible to distinguish the disease allele from the normal allele at each SNP. Therefore, we can only identify whether or not the XCI of the gene is skewed towards the minor alleles, but it is not possible to know whether the XCI is skewed towards the disease alleles or the normal alleles. Thirdly, the proposed Bayesian methods need to specify the prior distributions of all the unknown parameters in advance, and the selection of the prior distributions may have a certain impact on the results. For simplicity, we only considered two possible prior distributions for γ , and one prior distribution for each of the other unknown parameters. However, the prior distributions of these parameters are usually unknown, and we cannot guarantee that the weak prior distributions we used are the most appropriate. We provide an R package named GEXCIS, which is publicly available at <https://github.com/Meng-KaiLi/GEXCIS> (accessed on 30 April 2022), and can be used to estimate the degree of the skewness of the XCI for genes through the proposed methods in this paper. This R package also allows researchers to specify the prior distribution of each unknown parameter from their own research backgrounds. Fourthly, the Bayesian methods use the HMC algorithm for the sampling, which is not affected by the correlation between unknown parameters. Therefore, to improve computational efficiency, we assumed that all the unknown parameters are independent. However, the Bayesian methods, taking the correlation between the parameters into account, should have better performance, which is our future work. Fifthly, if the HPDIs or CIs we get contain 1, which means that the XCI

pattern for the gene is the XCI-R or the XCI-E, our proposed methods cannot distinguish them. Therefore, in our future study, we will consider including males' information to distinguish the XCI-R from the XCI-E. Finally, the proposed methods are only applicable to independent female subjects, and we will extend them in future so that they could accommodate the family data.

5. Conclusions

We propose four point estimates and four interval estimation methods to estimate γ of genes. Among the four point estimates, $\hat{\gamma}_{GF}$ may have the extreme point estimates, and $\hat{\gamma}_{GPF}$ can only reduce the occurrence of the extreme point estimates equal to 2, while $\hat{\gamma}_{GBN}$ and $\hat{\gamma}_{GBU}$ can avoid the extreme point estimates occurring. As for the four interval estimation methods, the Fieller's method may derive empty sets, discontinuous intervals and noninformative intervals, and the PF method can avoid the occurrence of discontinuous intervals and get less noninformative intervals, while the GBN and GBU methods do not yield these three types of the intervals. However, it should be noted that through these proposed methods, we cannot obtain the degree of the skewness of the XCI for each SNP in the gene, and cannot know whether the XCI is skewed towards the disease alleles or the normal alleles. In summary, the point estimates obtained by the GBN method always have the least MSE, and the HPDIs of the GBN method generally have the shortest width and the lowest variation, so we recommend using the GBN method in practical applications.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13050827/s1>, Table S1: Results of point estimations and interval estimations for γ among 500 replications with $n = 3000$ and 4000 , $\eta = 1$, $\tau = 1$ and $\sigma = 1$ for qualitative trait; Table S2: MSEs of $\hat{\gamma}_{GBN}$, $\hat{\gamma}_{GBU}$, $\hat{\gamma}_{GPF}$ and $\hat{\gamma}_{GF}$ among 500 replications with $n = 2000$ and $\sigma = 2$ for quantitative trait; Table S3: CPs (%), W_{mean} and W_{median} of GBN, GBU, PF and Fieller's methods among 500 replications with $n = 2000$ and $\sigma = 2$ for quantitative trait; Table S4: W_{sd} and W_{iqr} of GBN, GBU, PF and Fieller's methods among 500 replications with $n = 2000$ and $\sigma = 2$ for quantitative trait; Figures S1–S6: Scatter plots of point estimates of γ against true values of γ for quantitative ($\sigma = 1$) or qualitative trait with $n = 500$ and 2000 , and $\tau = 0.6$ and 1 ; Figures S7–S12: Widths of HPDIs or CIs of GBN, GBU, PF and Fieller's methods against true values of γ for quantitative ($\sigma = 1$) or qualitative trait with $n = 500$ and 2000 , and $\tau = 0.6$ and 1 ; Figures S13 and S14: Scatter plots of point estimates of γ against true values of γ for quantitative trait with $n = 2000$, $\tau = 0.6$ and 1 , and $\sigma = 2$; Figures S15 and S16: Widths of HPDIs or CIs of GBN, GBU, PF and Fieller's methods against true values of γ for quantitative trait with $n = 2000$, $\tau = 0.6$ and 1 , and $\sigma = 2$.

Author Contributions: Conceptualization, J.-Y.Z.; methodology, M.-K.L. and Y.-X.Y.; software, M.-K.L., Y.-X.Y. and J.-Y.Z.; validation, M.-K.L., Y.-X.Y., B.Z. and K.-W.W.; writing—original draft preparation, M.-K.L. and Y.-X.Y.; writing—review and editing, B.Z., K.-W.W., W.K.F. and J.-Y.Z.; supervision, W.K.F. and J.-Y.Z.; project administration, J.-Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 82173619 and 81773544, the Science and Technology Planning Project of Guangdong Province, grant number 2020B1212030008, and the Hong Kong Research Grants Council, grant number 17302919.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs00062.0.v1.p1 (accessed on 5 January 2022).

Acknowledgments: A Minnesota Center for Twin and Family Research (MCTFR) was supported by the National Institute on Drug Abuse, grant number U01 DA024417. The sample ascertainment and data collection in MCTFR data were supported by the National Institute on Drug Abuse, grant

numbers R37 DA05147 and R01 DA13240, the National Institute on Alcohol Abuse and Alcoholism, grant numbers R01 AA09367 and R01 AA11886, and the National Institute of Mental Health, grant number R01 MH66140.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

We assume that γ is the mean degree of the skewness of the XCI for the gene under study. For the i th female, we have $X_i = \sum_{j=1}^J \omega_j [\gamma g_{ij}^{(1)} + (2 - \gamma)g_{ij}^{(2)}]$. On the other hand, when supposing that the degree of the skewness of the XCI at the j th SNP is γ_j , the genotypic values of genotypes $d_j d_j$, $D_j d_j$ and $D_j D_j$ at the j th SNP of the i th female are 0, γ_j and 2, respectively. Similar to the construction process of X_i , we can get $X_i^* = \sum_{j=1}^J \omega_j [\gamma_j g_{ij}^{(1)} + (2 - \gamma_j)g_{ij}^{(2)}]$. Under the assumption of $\sum_{i=1}^n X_i = \sum_{i=1}^n X_i^*$, we have

$$\sum_{i=1}^n \sum_{j=1}^J \omega_j [\gamma g_{ij}^{(1)} + (2 - \gamma)g_{ij}^{(2)}] = \sum_{i=1}^n \sum_{j=1}^J \omega_j [\gamma_j g_{ij}^{(1)} + (2 - \gamma_j)g_{ij}^{(2)}],$$

and

$$\gamma \sum_{i=1}^n \sum_{j=1}^J \omega_j (g_{ij}^{(1)} - g_{ij}^{(2)}) = \sum_{i=1}^n \sum_{j=1}^J \omega_j (\gamma_j (g_{ij}^{(1)} - g_{ij}^{(2)})) \gamma_j.$$

Then,

$$\gamma \sum_{j=1}^J \omega_j (g_{.j}^{(1)} - g_{.j}^{(2)}) = \sum_{j=1}^J \omega_j (\gamma_j (g_{.j}^{(1)} - g_{.j}^{(2)})) \gamma_j,$$

where $g_{.j}^{(1)} = \sum_{i=1}^n g_{ij}^{(1)}$ and $g_{.j}^{(2)} = \sum_{i=1}^n g_{ij}^{(2)}$. Finally, we have

$$\gamma = \frac{\sum_{j=1}^J \omega_j (g_{.j}^{(1)} - g_{.j}^{(2)}) \gamma_j}{\sum_{j=1}^J \omega_j (g_{.j}^{(1)} - g_{.j}^{(2)})}.$$

References

1. Lyon, M.F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **1961**, *190*, 372–373. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Amos-Landgraf, J.M.; Cottle, A.; Plenge, R.M.; Friez, M.; Schwartz, C.E.; Longshore, J.; Willard, H.F. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am. J. Hum. Genet.* **2006**, *79*, 493–499. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Plenge, R.M.; Stevenson, R.A.; Lubs, H.A.; Schwartz, C.E.; Willard, H.F. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *Am. J. Hum. Genet.* **2002**, *71*, 168–173. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Shvetsova, E.; Sofronova, A.; Monajemi, R.; Galalova, K.; Draisma, H.; White, S.J.; Santen, G.; Chuva de Sousa Lopes, S.M.; Heijmans, B.T.; van Meurs, J.; et al. Skewed X-inactivation is common in the general female population. *Eur. J. Hum. Genet.* **2019**, *27*, 455–465. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Medema, R.H.; Burgering, B.M. The X factor: Skewing X inactivation towards cancer. *Cell* **2007**, *129*, 1253–1254. [\[CrossRef\]](#)
6. Deng, X.; Berletch, J.B.; Nguyen, D.K.; Disteche, C.M. X chromosome regulation: Diverse patterns in development, tissues and disease. *Nat. Rev. Genet.* **2014**, *15*, 367–378. [\[CrossRef\]](#)
7. Posynick, B.J.; Brown, C.J. Escape from X-chromosome inactivation: An evolutionary perspective. *Front. Cell Dev. Biol.* **2019**, *7*, 241. [\[CrossRef\]](#)
8. Peeters, S.B.; Cotton, A.M.; Brown, C.J. Variable escape from X-chromosome inactivation: Identifying factors that tip the scales towards expression. *Bioessays* **2014**, *36*, 746–756. [\[CrossRef\]](#)
9. Minks, J.; Robinson, W.P.; Brown, C.J. A skewed view of X chromosome inactivation. *J. Clin. Invest.* **2008**, *118*, 20–23. [\[CrossRef\]](#)
10. Chabchoub, G.; Uz, E.; Maalej, A.; Mustafa, C.A.; Rebai, A.; Mnif, M.; Bahloul, Z.; Farid, N.R.; Ozcelik, T.; Ayadi, H. Analysis of skewed X-chromosome inactivation in females with rheumatoid arthritis and autoimmune thyroid diseases. *Arthritis Res. Ther.* **2009**, *11*, R106. [\[CrossRef\]](#)
11. Sun, Z.; Fan, J.; Wang, Y. X-chromosome inactivation and related diseases. *Genet. Res.* **2022**, *2022*, 1391807. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Okumura, K.; Fujimori, Y.; Takagi, A.; Murate, T.; Ozeki, M.; Yamamoto, K.; Katsumi, A.; Matsushita, T.; Naoe, T.; Kojima, T. Skewed X chromosome inactivation in fraternal female twins results in moderately severe and mild haemophilia B. *Haemophilia* **2008**, *14*, 1088–1093. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Garagiola, I.; Mortarino, M.; Siboni, S.M.; Boscarino, M.; Mancuso, M.E.; Biganzoli, M.; Santagostino, E.; Peyvandi, F. X chromosome inactivation: A modifier of factor VIII and IX plasma levels and bleeding phenotype in Haemophilia carriers. *Eur. J. Hum. Genet.* **2021**, *29*, 241–249. [\[CrossRef\]](#)

14. Zuo, T.; Wang, L.; Morrison, C.; Chang, X.; Zhang, H.; Li, W.; Liu, Y.; Wang, Y.; Liu, X.; Chan, M.; et al. FOXP3 is an X-linked breast cancer suppressor gene and an important repressor of the HER-2/ErbB2 oncogene. *Cell* **2007**, *129*, 1275–1286. [[CrossRef](#)]
15. Li, G.; Jin, T.; Liang, H.; Tu, Y.; Zhang, W.; Gong, L.; Su, Q.; Gao, G. Skewed X-chromosome inactivation in patients with esophageal carcinoma. *Diagn. Pathol.* **2013**, *8*, 55. [[CrossRef](#)] [[PubMed](#)]
16. Simmonds, M.J.; Kavvoura, F.K.; Brand, O.J.; Newby, P.R.; Jackson, L.E.; Hargreaves, C.E.; Franklyn, J.A.; Gough, S.C. Skewed X chromosome inactivation and female preponderance in autoimmune thyroid disease: An association study and meta-analysis. *J. Clin. Endocrinol. Metab.* **2014**, *99*, E127–E131. [[CrossRef](#)] [[PubMed](#)]
17. Giliberto, F.; Radic, C.P.; Luce, L.; Ferreira, V.; de Brasi, C.; Szijan, I. Symptomatic female carriers of Duchenne muscular dystrophy (DMD): Genetic and clinical characterization. *J. Neurol. Sci.* **2014**, *336*, 36–41. [[CrossRef](#)]
18. Sangha, K.K.; Stephenson, M.D.; Brown, C.J.; Robinson, W.P. Extremely skewed X-chromosome inactivation is increased in women with recurrent spontaneous abortion. *Am. J. Hum. Genet.* **1999**, *65*, 913–917. [[CrossRef](#)] [[PubMed](#)]
19. Zhang, Y.; Xu, S.Q.; Liu, W.; Fung, W.K.; Zhou, J.Y. A robust test for X-chromosome genetic association accounting for X-chromosome inactivation and imprinting. *Genet. Res.* **2020**, *102*, e2. [[CrossRef](#)]
20. Zhang, L.; Martin, E.R.; Morris, R.W.; Li, Y.J. Association test for X-linked QTL in family-based designs. *Am. J. Hum. Genet.* **2009**, *84*, 431–444. [[CrossRef](#)]
21. Zheng, G.; Joo, J.; Zhang, C.; Geller, N.L. Testing association for markers on the X chromosome. *Genet. Epidemiol.* **2007**, *31*, 834–843. [[CrossRef](#)] [[PubMed](#)]
22. Clayton, D. Testing for association on the X chromosome. *Biostatistics* **2008**, *9*, 593–600. [[CrossRef](#)] [[PubMed](#)]
23. Wang, J.; Yu, R.; Shete, S. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genet. Epidemiol.* **2014**, *38*, 483–493. [[CrossRef](#)] [[PubMed](#)]
24. Liu, W.; Wang, B.Q.; Liu-Fu, G.; Fung, W.K.; Zhou, J.Y. X-chromosome genetic association test incorporating X-chromosome inactivation and imprinting effects. *J. Genet.* **2019**, *98*, 99. [[CrossRef](#)] [[PubMed](#)]
25. Ma, L.; Hoffman, G.; Keinan, A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics* **2015**, *16*, 241. [[CrossRef](#)]
26. Gao, F.; Chang, D.; Biddanda, A.; Ma, L.; Guo, Y.; Zhou, Z.; Keinan, A. XWAS: A software toolset for genetic data analysis and association studies of the X chromosome. *J. Hered.* **2015**, *106*, 666–671. [[CrossRef](#)]
27. Madsen, B.E.; Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **2009**, *5*, e1000384. [[CrossRef](#)]
28. Li, B.; Leal, S.M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **2008**, *83*, 311–321. [[CrossRef](#)]
29. Schork, N.J.; Murray, S.S.; Frazer, K.A.; Topol, E.J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **2009**, *19*, 212–219. [[CrossRef](#)]
30. Han, F.; Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **2010**, *70*, 42–54. [[CrossRef](#)]
31. Ionita-Laza, I.; Buxbaum, J.D.; Laird, N.M.; Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* **2011**, *7*, e1001289. [[CrossRef](#)] [[PubMed](#)]
32. Price, A.L.; Kryukov, G.V.; de Bakker, P.I.; Purcell, S.M.; Staples, J.; Wei, L.J.; Sunyaev, S.R. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **2010**, *86*, 832–838. [[CrossRef](#)] [[PubMed](#)]
33. Wu, M.C.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **2011**, *89*, 82–93. [[CrossRef](#)] [[PubMed](#)]
34. Lee, S.; Emond, M.J.; Bamshad, M.J.; Barnes, K.C.; Rieder, M.J.; Nickerson, D.A.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team; Christiani, D.C.; Wurfel, M.M.; Lin, X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **2012**, *91*, 224–237. [[CrossRef](#)] [[PubMed](#)]
35. Ionita-Laza, I.; Lee, S.; Makarov, V.; Buxbaum, J.D.; Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **2013**, *92*, 841–853. [[CrossRef](#)] [[PubMed](#)]
36. Ma, C.; Boehnke, M.; Lee, S.; GoT2D Investigators. Evaluating the calibration and power of three gene-based association tests of rare variants for the X chromosome. *Genet. Epidemiol.* **2015**, *39*, 499–508. [[CrossRef](#)] [[PubMed](#)]
37. Turkmen, A.S.; Lin, S. Detecting X-linked common and rare variant effects in family-based sequencing studies. *Genet. Epidemiol.* **2021**, *45*, 36–45. [[CrossRef](#)]
38. Xu, S.Q.; Zhang, Y.; Wang, P.; Liu, W.; Wu, X.B.; Zhou, J.Y. A statistical measure for the skewness of X chromosome inactivation based on family trios. *BMC Genet.* **2018**, *19*, 109. [[CrossRef](#)]
39. Wang, P.; Zhang, Y.; Wang, B.Q.; Li, J.L.; Wang, Y.X.; Pan, D.; Wu, X.B.; Fung, W.K.; Zhou, J.Y. A statistical measure for the skewness of X chromosome inactivation based on case-control design. *BMC Bioinformatics* **2019**, *20*, 11. [[CrossRef](#)]
40. Li, B.H.; Yu, W.Y.; Zhou, J.Y. A statistical measure for the skewness of X chromosome inactivation for quantitative traits and its application to the MCTFR data. *BMC Genom. Data* **2021**, *22*, 24. [[CrossRef](#)]
41. Wang, P.; Xu, S.; Wang, Y.X.; Wu, B.; Fung, W.K.; Gao, G.; Liang, Z.; Liu, N. Penalized Fieller’s confidence interval for the ratio of bivariate normal means. *Biometrics* **2021**, *77*, 1355–1368. [[CrossRef](#)] [[PubMed](#)]

42. Stephens, M.; Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **2009**, *10*, 681–690. [[CrossRef](#)] [[PubMed](#)]
43. Annis, J.; Miller, B.J.; Palmeri, T.J. Bayesian inference with Stan: A tutorial on adding custom distributions. *Behav. Res. Methods* **2017**, *49*, 863–886. [[CrossRef](#)] [[PubMed](#)]
44. Kruschke J., K. Bayesian data analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **2010**, *1*, 658–676. [[CrossRef](#)] [[PubMed](#)]
45. Wang, C.; Deng, S.; Sun, L.; Li, L.; Hu, Y.Q. A nonparametric test for association with multiple loci in the retrospective case-control study. *Stat. Methods Med. Res.* **2020**, *29*, 589–602. [[CrossRef](#)]
46. Basu, S.; Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **2011**, *35*, 606–619. [[CrossRef](#)] [[PubMed](#)]
47. Turkmen, A.S.; Yan, Z.; Hu, Y.Q.; Lin, S. Kullback-Leibler distance methods for detecting disease association with rare variants from sequencing data. *Ann. Hum. Genet.* **2015**, *79*, 199–208. [[CrossRef](#)] [[PubMed](#)]
48. McGue, M.; Zhang, Y.; Miller, M.B.; Basu, S.; Vrieze, S.; Hicks, B.; Malone, S.; Oetting, W.S.; Iacono, W.G. A genome-wide association study of behavioral disinhibition. *Behav. Genet.* **2013**, *43*, 363–373. [[CrossRef](#)]
49. Asadollahi, H.; Vaez Torshizi, R.; Ehsani, A.; Masoudi, A.A. An association of CEP78, MEF2C, VPS13A and ARRDC3 genes with survivability to heat stress in an F₂ chicken population. *J. Anim. Breed. Genet.* **2022**. [[CrossRef](#)]
50. McCaw, Z.R.; Lane, J.M.; Saxena, R.; Redline, S.; Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **2020**, *76*, 1262–1272. [[CrossRef](#)]
51. Ng, K.T.; Yeung, O.W.; Liu, J.; Li, C.X.; Liu, H.; Liu, X.B.; Qi, X.; Ma, Y.Y.; Lam, Y.F.; Lau, M.Y.; et al. Clinical significance and functional role of transmembrane protein 47 (TMEM47) in chemoresistance of hepatocellular carcinoma. *Int. J. Oncol.* **2020**, *57*, 956–966. [[CrossRef](#)]
52. Men, X.; Su, M.; Ma, J.; Mou, Y.; Dai, P.; Chen, C.; Cheng, X.A. Overexpression of TMEM47 induces tamoxifen resistance in human breast cancer cells. *Technol. Cancer Res. Treat.* **2021**, *20*, 15330338211004916. [[CrossRef](#)]