Mapping Quantitative Trait Loci under the Multivariate-T Model

Jie Peng[1], David Siegmund[2]

[1]University of California, Davis, CA 95616

[2]Stanford University, Stanford, CA 94305

July 23, 2006

Abstract

The basic theory associated with standard variance-component models for QTL mapping in linkage analysis usually assumes multivariate normality. Robust score statistic based on the normality assumption has been proposed and showed to have the correct type-I error free of the true phenotypic distribution (Tang and Siegmund (2001)), however very low power has been observed when phenotypic distributions are far away from normal. Since the normality assumption is often violated in practice, we propose to use multivariate-t distribution as an alternative. This assumption leads to a mathematically tractable statistic. We show that it keeps most of the desirable properties of the normal-based statistic with respect to parameter estimation, ascertainment correction, tail approximation, etc. Moreover, the proposed statistic is more robust to outliers and more powerful when phenotypic distributions are heavily-tailed and/or skewed. This is partially because the t-distribution fits the data better by using one additional parameter–the degrees of freedom. When the phenotypic distribution is near normal, the two statistics are about equivalent, so there is no loss of power by using the new statistic. Numerical results under various phenotypic distributions w/out ascertainment sampling are presented and substantial gain in power is observed. Several data-transformation techniques including Copula, log-transformation are also discussed and compared to the proposed method.

1 Introduction

In this paper, we derive statistics under the variance-component model (Almsay and Blangero 1998) to map quantitative trait loci (QTL). Although the basic theory associated with the stan-

dard variance-component model for QTL mapping in linkage analysis usually assumes multivariate normality, we find that the power of the statistics derived under the normality assumption could be very low when the phenotypic distribution is actually non-normal. Therefore we propose new statistics derived from a multivariate-t model (Lange *et al.* (1989)). By simulations, we find the power can be increased as much as from 11% for the normal-score to 91% for the t-score (Table 3). As a generalization of the normal-score, the t-statistics are still asymptotically locally optimal when the phenotypic distribution is multivariate normal. We also compare the t-statistics to several popular data transformation methods such as the copula (Wang and Huang (2002)) and the log-transformation. We find that under some cases the t-statistics are more powerful than the other two, while under other cases they could be less powerful. Nevertheless, all three of them are much more powerful than the normal-score when the phenotypic distribution is non-normal.

2 Variance-component model

We first introduce the variance component model for mapping QTL and derive the efficient score under the normality assumption. For more details, see Peng and Siegmund (2006). We assume Hardy-Weinberg equilibrium and linkage equilibrium throughout. Our model goes back to the classic paper of Fisher (1918) for the case of diallelic genes; the general case is discussed by Kempthorne (1957). The basic model for the phenotype Y having a mean value μ is

$$Y = \mu + \alpha_m + \alpha_f + \delta_{m,f} + e, \tag{2.1}$$

where $\alpha_x = \alpha_x(\tau)$ denotes the additive genetic effect of allele x at locus τ, and $\delta_{x,y}$ denotes the dominance deviation of alleles x, y. The subscript $m(f)$ denotes the allele contributed by the mother (father). By standard analysis of variance arguments, we may assume that $\mathrm{E}(\alpha_m) = \mathrm{E}(\alpha_f) = \mathrm{E}(e) = \mathrm{E}[\delta_{m,f}|m] = \mathrm{E}[\delta_{m,f}|f] = 0$. By the assumption of Hardy-Weinberg equilibrium m and f are independent (unless the parents are inbred), so the different genetic effects in the model (2.1) are uncorrelated. We also assume the residual term e, which may include the genetic effects from other QTL, is uncorrelated with the explicitly modelled genetic effects from τ. It

follows that $\sigma_Y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2$, where $\sigma_A^2 = 2\mathrm{E}(\alpha_m^2),\ \sigma_D^2 = \mathrm{E}(\delta_{m,f}^2),\ \sigma_e^2 = \mathrm{E}(e^2)$ are the additive variance, dominance variance and residual variance, respectively.

Although the analysis below can be applied to more general pedigrees, for simplicity, we focus on sibships of the same size, s in this paper. Under model (2.1), it is easy to calculate variance components, as well as covariance matrices.

Consider a pair of siblings satisfying model (2.1). Denote by $\nu = \nu(\tau)$ the number of alleles identical by descent at τ. Letting Y_i denote the phenotypic value of the ith sibling ($i = 1, 2$), we have

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2|\nu) &= \sigma_A^2\nu/2 + \sigma_D^2 1_{\{\nu=2\}} + \sigma_e^2 r \\ &= \mathrm{Cov}(Y_1, Y_2) + \alpha_0(\nu - 1) + \delta_0(1/2 - 1\{\nu = 1\}) \end{aligned}$$

where $r = \mathrm{corr}(e_1, e_2)$ accounts for the correlation between sibs that arises from other QTL and from a shared environment, while

$$\alpha_0 = \frac{\sigma_A^2 + \sigma_D^2}{2}, \quad \delta_0 = \frac{\sigma_D^2}{2}.$$

Note that $0 \le \delta_0 \le \alpha_0$.

The null hypothesis we want to test is $\alpha_0 = 0$ (which implies that $\delta_0 = 0$ as well). The *working assumption* for the variance-component model is that the conditional distribution of the phenotypes in a pedigree given the pairwise IBD sharing at a QTL is multivariate normal. The usefulness of this assumption is due to the mathematical tractability of the multivariate normal distribution. However, it cannot be expected to be exactly true. For a given sibship, let $\mathbf{Y}^T = (Y_1, \cdots, Y_s)$ denote the vector of phenotypes and M denote the matrix of all marker genotypes. The observed data for a sampled sibship is $(\mathbf{Y}, M)$. Also let A_ν denote the $s \times s$ matrix with entries $\nu_{ij} - 1$ for $i \neq j$ and zeroes along the diagonal. Let $L = \mathrm{P}(\mathbf{Y}, M)$ denote the likelihood of a randomly sampled sibship. Let $\ell'(\alpha_0)$ denote $\partial \log(L)/\partial\alpha_0$. Under the normality assumption, when the markers are fully informative, i.e, the IBD sharing matrix A_ν is observable,

the efficient score at a putative trait locus t, evaluated with $\alpha_0 = 0$, is (Tang & Siegmund (2001))

$$\ell'(0) = -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}A_{\nu(t)}) + \frac{1}{2}(\mathbf{Y}-\mu\mathbf{1})^T\Sigma^{-1}A_{\nu(t)}\Sigma^{-1}(\mathbf{Y}-\mu\mathbf{1}), \tag{2.2}$$

where $\Sigma = \mathrm{E}[(\mathbf{Y}-\mu\mathbf{1})(\mathbf{Y}-\mu\mathbf{1})^T]$ is the phenotypic covariance matrix. When markers are not fully informative, the likelihood can be written as

$$L = \mathrm{P}(\mathbf{Y}, M) = \sum \mathrm{P}(\mathbf{Y}|A_\nu)\mathrm{P}(A_\nu|M)\mathrm{P}(M),$$

where the summation is taken over all possible configurations of the IBD sharing matrix A_ν and we use the fact that $\mathrm{P}(\mathbf{Y}|M, A_\nu) = \mathrm{P}(\mathbf{Y}|A_\nu)$. Since (2.2) is linear in A_ν, it is easy to see that the efficient score for α_0 is obtained by replacing A_ν with its conditional expectation $A_{\hat{\nu}} = \mathrm{E}(A_\nu|M)$ in (2.2), i.e.,

$$\ell'(0) = -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}A_{\hat{\nu}}) + \frac{1}{2}(\mathbf{Y}-\mu\mathbf{1})^T\Sigma^{-1}A_{\hat{\nu}}\Sigma^{-1}(\mathbf{Y}-\mu\mathbf{1}). \tag{2.3}$$

Now let $\mathcal{A}$ denote the event that a particular sibship is ascertained. We assume the following *measurable ascertainment assumption*: each pedigree is ascertained through its phenotypes and possibly some additional randomization, but not its genotypes. The conditional likelihood of the data given $\mathcal{A}$ is

$$L_{\mathcal{A}} = \mathrm{P}(\mathbf{Y}, M|\mathcal{A}) = [\mathrm{P}(\mathbf{Y}, M)/\mathrm{P}(\mathcal{A})]I(\mathcal{A}) = [L/P(\mathcal{A})]I(\mathcal{A}),$$

where $I(.)$ is the indicator function. Peng and Siegmund (2006) shows that, the efficient score of the conditional likelihood $L_{\mathcal{A}}$ is the same as the efficient score of the unconditional likelihood L, given by (2.3). They also show that under the normality assumption and the measurable ascertainment assumption, for the conditional likelihood $L_{\mathcal{A}}$, the nuisance parameters are orthogonal to the linkage parameters under the null hypothesis of no linkage.

At a putative trait locus t, the efficient score for one ascertained sibship is given by (2.3). To use this score as a test statistic, we will standardize it to have unit variance under the null hypothesis and substitute estimates for the unknown segregation parameters μ, Σ. The variance

of the efficient score $\ell'(0)$ under the null hypothesis is given by

$$I_{\alpha\alpha} = \mathrm{E}_0(\ell'^2(0)|\mathcal{A}). \tag{2.4}$$

Therefore at a putative locus t, for a sample of ascertained sibships, the *standardized score* is

$$Z_t = \sum_n \ell'_n(0) \Big/ \big[\sum_n I_{\alpha\alpha,n}\big]^{1/2}, \tag{2.5}$$

where the summation is taken over all ascertained sibships. A second statistic involving $\ell'_\delta(0)$ can be defined similarly. Since $\alpha_0 = (\sigma_A^2 + \sigma_D^2)/2,\ \delta_0 = \sigma_D^2/2$, a two dimensional statistic rarely has more power than (2.5) unless the dominance effect σ_D^2 is quite large. Therefore for the following discussion we only consider (2.5) . Since we do not know the location of the QTL τ, we can scan the genome by the *test statistic*

$$Z_{\max} = \max_t Z_t,$$

where the maximum is taken over all marker loci t throughout the genome.

When the normality assumption is violated, the standardization (2.4) of the statistics is in general incorrect and leads to the possibility of an inflated false positive error probability. This problem can be mitigated by using the conditional variance of $\ell'(0)$, given the phenotypic values, $\mathbf{Y}_1, \cdots, \mathbf{Y}_N$ for all the ascertained sibships. Then the normalized score at marker t becomes

$$Z_t = \sum_n \ell'_n(0) \Big/ \big[\sum_n I^R_{\alpha\alpha,n}\big]^{1/2}, \tag{2.6}$$

where $I^R_{\alpha\alpha,n} = \mathrm{E}_0[\ell'^2_n(0)|\mathbf{Y}_n]$. If the estimator $\hat{\theta}$ is a function (only) of the phenotypes, the statistic (2.6) is asymptotically normally distributed with mean zero, variance one. In the case of fully informative markers, the testing process $\{Z(t)\}$ is a discretely observed Ornstein-Uhlenbeck process. An approximation to the genome-wide false positive rate, $\mathrm{P}_0(Z_{\max} \geq z)$, is given by Feingold *et al.* (1993),

$$\mathrm{P}_0\{\max_i Z_{i\Delta} \geq z\} \approx 1 - \exp\{-c[1 - \Phi(z)] - L\beta z\varphi(z)h[z(2\beta\Delta)^{1/2}]\}. \tag{2.7}$$

When markers are only partially informative, the process is still approximately a Gaussian process, but its p-value is slightly smaller. One can either use the fully informative approximation, which is slightly conservative, or use a Monte Carlo method to approximate the p-value. In the following we refer this statistic as the *(robust) normal-score statistic*. If the normality assumption in fact is true, (2.6) has the same asymptotic noncentrality as (2.5). Nevertheless, (2.6) is robust in validity (Miller 1978) in the sense that a correct (or slightly conservative) type-I error rate can be obtained by (2.7) regardless of the true phenotypic distribution. However, as we will see later, it is not robust in efficiency in the sense that it has very low power when the normality assumption is violated.

3 Multivariate-t Model and T-transformations

In the previous section, the primary role of the normality assumption is to suggest the form of the statistic given in (2.2), which can be regarded as a covariance between a function of phenotypes and identity-by-descent counts. An alternative to the multivariate normal distribution that is equally tractable is the multivariate t-distribution (Lange *et al.* (1989)). This leads to a similar statistic, but because of the heavier tails of the t-distribution it is more robust to outliers in the data. Moreover, although it cannot avoid problems that arise from modelling the complexities of multivariate dependence by multivariate distributions that measure dependence only by pair wise correlations, the t-distribution results in a more powerful statistic when the phenotypic distribution is heavy-tailed or skewed. This is because it has one more parameter– the degrees of freedom k to fit the data.

In a multivariate-t model, we assume the conditional distribution of the phenotypes in a pedigree given the pairwise IBD sharing at a QTL is multivariate-t up to a location and scale transformation

$$\mathbf{Y}|A_\nu \sim \sqrt{\frac{k-2}{k}} T_{(k)}(\Sigma_\nu) + \mu\mathbf{1}, \tag{3.8}$$

where $T_{(k)}(\Sigma_\nu) = X/\sqrt{Y/k}$, with $X \sim N_s(0, \mathbf{\Sigma}_\nu)$, $Y \sim \chi^2_{(k)}$ and X is independent of Y. For sibships, $\Sigma_\nu = \Sigma + \alpha_0 A_\nu$. The efficient score evaluated with $H_0 : \alpha_0 = 0$ at a putative trait locus

t for one sibship is

$$\ell'(0) = -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}A_{\widehat{\nu}(t)}) + \frac{1}{2}\frac{k+s}{k}\frac{\mathrm{tr}(\Sigma^{-1}A_{\widehat{\nu}(t)}\Sigma^{-1}\mathbf{Z}\mathbf{Z}')}{1+\mathbf{Z}'\Sigma^{-1}\mathbf{Z}/k}, \tag{3.9}$$

where $\mathbf{Z} = \sqrt{\frac{k}{k-2}}\frac{\mathbf{Y}-\mu\mathbf{1}}{\sigma_Y}$. If in the normal score (2.2) we replace $\mathbf{Y}-\mu\mathbf{1}$ by $\widehat{\mathbf{Y}}$, then we will get the t-score (3.9), where

$$\widehat{\mathbf{Y}} = \sqrt{\frac{k+s}{k}}\frac{\mathbf{Z}}{\sqrt{1+\frac{1}{k}\mathbf{Z}'\Sigma^{-1}\mathbf{Z}}}, \tag{3.10}$$

which is referred as the *t-transformation* hereafter. Note that, the above transformation is data dependent since k needs to be estimated from the data. In the following we list some properties of the t-score (3.9) and t-transformation (3.10) and many of them are carried over from the normal score.

1. Under the multivarite-t model, for both random sampling and ascertainment, the efficient score is (3.9) and the population parameters are orthogonal to the linkage parameters under the null.

2. The robust Fisher information $\mathrm{E}_0[(\ell'(0))^2|\mathbf{Y}]$ of the t-score is given by the corresponding robust Fisher information $I^R_{\alpha\alpha}$ of the normal score with $\mathbf{Y}$ replaced by $\widehat{\mathbf{Y}}$.

3. The t-transformation (3.10) makes the data less skewed and less heavy-tailed: $\widehat{\mathbf{Y}}$ has a smaller skewness and kurtosis than $\mathbf{Y}$ (Table 1).

4. When the degrees of freedom k is large, the t-transformation is just a standardization of the data. Therefore the resulting t-score is close to the normal score. Since multivariate normal distribution is a special case of the multivariate-t distribution with the degrees of freedom k being ∞, when the phenotypic distribution is indeed multivariate normal, using the t-score with k fitted from the data is very close to using the normal score. This means that, there is no loss of efficiency by using the t-score under the normal case.

5. The asymptotic distribution of the testing process derived from the t-score (3.9) is an O-U process, thus the formula (2.7) for the tail probabilities is still applicable. (2.7) is actually more accurate, since the Gaussian approximation via the central limit theorem converges faster under the transformed data $\widehat{\mathbf{Y}}$ due to the smaller skewness and kurtosis.

For a data set, we can either use the normalized t-score based on (3.9), or we can apply the t-transformation (3.10) and then fit the robust normal score (2.6) to $\widehat{\mathbf{Y}}$. Since usually $\widehat{\mathbf{Y}}$ is not standardized to have mean zero and variance one, and the correlation between $\widehat{Y}_1$ and $\widehat{Y}_2$ is different from that between Y_1 and Y_2, the above two statistics are different (Table 1).

In order to apply the t-score or the t-transformation, we need estimate $\mu,\ \sigma_Y,\ \rho,\ k$ from the data. If pedigrees are randomly sampled, one can use the sample estimators of the population parameters and pick up the k which maximizes the log likelihood under the multivariate-t model (3.8). The set of k we examined in the simulation study is $\Omega_k = \{4,\ 5,\ \cdots,\ 102,\ 103\}$ with $k = 4$ corresponding to a very heavy-tailed distribution and $k = 103$ corresponding to the multivariate normal distribution. Since this estimate is very close to the maximum likelihood estimate, therefore we refer it as the MLE. If pedigrees are ascertained, for each given k, the population parameters need to be estimated. For example, we can use the conditional MLE under the multivariate-t model (3.8), where the conditional likelihood is obtained from conditioning on the probands' phenotypes (Peng and Siegmund 2006). Since this is very computationally intensive, in the simulation study, we estimate the population parameters by the conditional MLE under the normality assumption, then find the k in Ω_k which maximizes the log likelihood given the estimators of the population parameters. We expect this estimate to work reasonably well, since Peng and Siegmund (2006) shows that even when the phenotypic distribution is non-normal, the conditional MLE based on the normality assumption gives reasonable estimates of the population parameters.

The t-transformation (3.10) can be successively applied several times to the phenotypic data $\mathbf{Y}$ before a normal score or a t-score is used. By simulation, we observe that, repeating the t-transformation several times until the estimated k is very large usually gives the highest power. We think this is partially because the transformed data has a smaller skewness and kurtosis. In Table 1, we sample data from a multivariate gamma model which results in a skewed phenotypic distribution. We then successively apply the t-transformation three times on this data set. As can be seen from Table 1, the skewness and kurtosis become smaller as more t-transformations being applied. After two t-transformations, the fitted degrees of freedom becomes 103. This means that by applying the third t-transformation, we simply standardize the data to have mean zero and variance one, and applying more t-transformations will not have any effects any more.

Table 1: 1000 sibships of size 4 are randomly sampled from a multivariate gamma phenotype with $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25.\ \hat{\mu},\ \hat{\sigma}_Y,\ \hat{\rho},\ \hat{k}$ are MLE under model (3.8).

Data	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\rho}$	Skewness	Kurtosis	$\hat{k}$
Original data	-0.010	0.954	0.243	5.403	42.800	4
After one t-transformation	-0.243	0.789	0.344	2.055	3.775	4
After two t-transformations	-0.256	1.022	0.441	1.358	1.095	103
After three t-transformations	-0.010	1.002	0.444	1.336	1.036	103

Therefore, we say that for this data set, the *t-transformation converges at the third step*. It is also observed that, unless $\hat{k}$ is already large, the t-transformation usually changes the phenotypic mean, variance and correlation.

4 Copula Transformation

To deal with non-normality, Wang and Huang (2002) "recommend transforming the trait values first, on the basis of the empirical normal quantile-distribution transformation" (p. 415). The transformation is as following: suppose we have phenotypic data for N sibships, each of size s: $\left\{Y_i^{(k)},\ k = 1, \cdots, N,\ i = 1, \cdots, s\right\}$. Let $r_i^{(k)}$ be the rank of $Y_i^{(k)}$ among all the phenotypic data. Then transform $Y_i^{(k)}$ to

$$C_i^{(k)} = \Phi^{-1}\left(\frac{r_i^{(k)}}{1 + Ns}\right), \tag{4.11}$$

where Φ is the standard normal cumulative distribution function. The idea of transformation (4.11) is to make the multivariate data marginally normal. Since the above procedure is often referred as the "multivariate (empirical) normal copula" model, in this paper, we call the transformation (4.11) the *copula transformation*. We then assume, given the IBD sharing matrix, the joint distribution of $\mathbf{C}^{(k)} = \left(C_1^{(k)}, \cdots, C_s^{(k)}\right)'$ is multivariate normal. The normal score is then applied to the transformed data. If there is ascertainment, the conditional (normal) likelihood conditioning on the probands' (transformed) phenotypes will be used for the purpose of parameter estimation.

If $\left\{\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(N)}\right\}$ are randomly sampled multivariate normal data, then by Kolmogorov-Smirnov and symmetry, the transformed data $\{\mathbf{C}^{(k)}\}$ is approximately the same as $\{\mathbf{Y}^{(k)}\}$.

We then study what happens when applying the copula transformation to the ascertained multivariate normal data. Use Y^* to denote a random sample of size 1 from $\{Y_i^{(k)},\ k=1,\cdots,N,\ i=1,\cdots,s\}$; and for each i, use Y_i^* to denote a random sample of size 1 from $\{Y_i^{(k)},\ k=1,\cdots,N\}$. When the sample size N is large, $Y_j^{(k)}$ is nearly independent of both Y^* and $\{Y_i^*, i=1,\cdots,s\}$. Therefore given $Y_j^{(k)}=y$, by symmetry

$$\mathrm{P}(Y^\star<y)=\frac{s-1}{s}\mathrm{P}(Y_2^\star<y)+\frac{1}{s}\mathrm{P}(Y_1^\star<y)\approx\frac{r_j^{(k)}}{1+Ns}.$$

Now assume that $\mathbf{Y}^{(1)},\cdots,\mathbf{Y}^{(N)}$ are i.i.d. from $\mathbf{Y}\sim N_s(0,\Sigma)$ conditioned on $Y_1>b$, where $\Sigma=(1-\rho)\mathrm{I}_s+\rho\mathbf{1}_s\mathbf{1}_s'$. When b is not too large and s is at least moderate (say, $s\geq 4$), we can approximate the RHS of the above expression by $\mathrm{P}(Y_2^\star<y)$, which in turn is approximately $\mathrm{P}(Z_2<y|Z_1>b)$, where $(Z_1,Z_2)'$ is bivariate normal with mean 0, variance 1 and correlation ρ. When b is not small,

$$\mathrm{P}(y\leq Z_2\leq y+dy|Z_1>b)=\phi(y)\frac{1-\Phi\Big(\frac{b-\rho y}{\sqrt{1-\rho^2}}\Big)}{1-\Phi(b)}dy\approx\frac{1}{\sqrt{1-\rho^2}}\phi\Big(\frac{y-\rho b}{\sqrt{1-\rho^2}}\Big)dy.$$

This means that when b is not small, we can approximate the distribution of $Z_2|Z_1>b$ by $N(\rho b,1-\rho^2)$, then

$$\Phi^{-1}(\mathrm{P}(Y_2^*<y))\approx\Phi^{-1}(\mathrm{P}(Z_2<y|Z_1>b))\approx\frac{y-\rho b}{\sqrt{1-\rho^2}}.$$

Therefore when s is (at least) moderate and b is moderate, the copula transformation becomes

$$C_j^{(k)}=\Phi^{-1}\left(\frac{r_j^{(k)}}{1+Ns}\right)\approx\frac{Y_j^{(k)}-\rho b}{\sqrt{1-\rho^2}}. \tag{4.12}$$

Since a linear transformation does not affect the normal score (2.2), when phenotypic data is multivariate normal and sibships are ascertained through one extreme sib, the normal score of the copula transformed data $\{\mathbf{C}^{(k)}\}$ is roughly the same as that of the original data $\{\mathbf{Y}^{(k)}\}$ provided that the sibship size is not too small and the ascertainment criterion is not too stringent.

In Figure 1 and Figure 2, sibships are ascertained through $\{Y_1>b\}$ with the phenotypic

distribution being multivariate normal with $\mu = 0,\ \sigma_Y = 1,\ \Sigma = (1-\rho)\mathrm{I}_s + \rho \mathbf{1}_s \mathbf{1}_s',\ \rho = 0.25$. In each figure, the z-axis is the copula transformed data and the x-axis is the linearly transformed data $X = (Y - \rho b)/\sqrt{1-\rho^2}$. The red points stand for the (transformed) phenotypes of the second sibling: $\left(C_2^{(k)}, X_2^{(k)}\right)$; while the green points stand for the (transformed) phenotypes of the first sibling: $\left(C_1^{(k)}, X_1^{(k)}\right)$. In each figure, the blue line is the diagonal line $x = z$, while the purple line is the corrected line $x = z - d$ with d being the average difference between the two transformations. As can be seen from Figure 1, when the sibship size is $s = 4$, and the threshold b is the 75%-percentile of the phenotypic distribution, the red and green points closely reside on the line $x = z + 0.44$. This means that the copula transformed data is nearly a simple linear transformation of $\mathbf{Y}$ under this case, although the transformation is not exactly as expected by (4.12). As can be seen from Figure 2, when sibship size is $s = 2$ which is small and the threshold b is the 99%-percentile which is large, not only the average difference between the two transformations becomes larger: $d = -1.10$, but also the points do not reside on a straight line any more. This means that the copula transformed data is no longer a linear transformation of $\mathbf{Y}$, and in turn it results in a different normal score. We also observe that, for the case in Figure 1, the sample covariances of $\{\mathbf{C}^{(k)}\}$ and $\{\mathbf{X}^{(k)}\}$ are very close which again implies that they are different only up to a location transformation. In contrast, for the case in Figure 2, the two sample covariances are very different. The effects of the copula transformation on the power is examined by simulation in section 6, which are consistent with the above findings.

Remark 1. Note that, if $F_i^{(k)} = f\left(Y_i^{(k)}\right),\ i = 1, \cdots, s,\ k = 1, \cdots, N$ with f a monotonic function, e.g., $f(x) = \exp(x)$, then the copula transformation of $\{F^{(k)}\}$ is the same as that of $\{\mathbf{Y}^{(k)}\}$.

Remark 2. In Diao and Lin (2005), the authors propose to use a nonparametric log likelihood (equation (3)) and then estimate both finite and infinite parameters based on this likelihood. We believe that it is equivalent as the "copula" transformation discussed above at least for randomly sampled phenotypes. This is because (i) the transformation function H is step wise and only changes values at the observed points Y_{ij}; The MLE of the nuisance parameters depend only on the ranks of the data points; (ii) If we first apply the copula transformation on $\{Y_{ij}\}$ and get $\{X_{ij}\}$, then $\mathbf{X}$ and $\mathbf{Y}$ has the same rank in the sense that if the rank of Y_{ij} among all Y is k, then rank of X_{ij} is also k. Therefore by applying this method on the copula transformed data $\mathbf{X}$, we

Figure 1: Comparison between the linear transformation and the copula transformation ($s = 4,\ b = 0.674$). The z-axis stands for the copula transformed data and the x-axis stands for the linearly transformed data. The phenotype $\mathbf{Y}$ for each sibship is normally distributed with $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25$. 400 sib-quads are ascertained through $\{Y_1 > 0.674\}$.

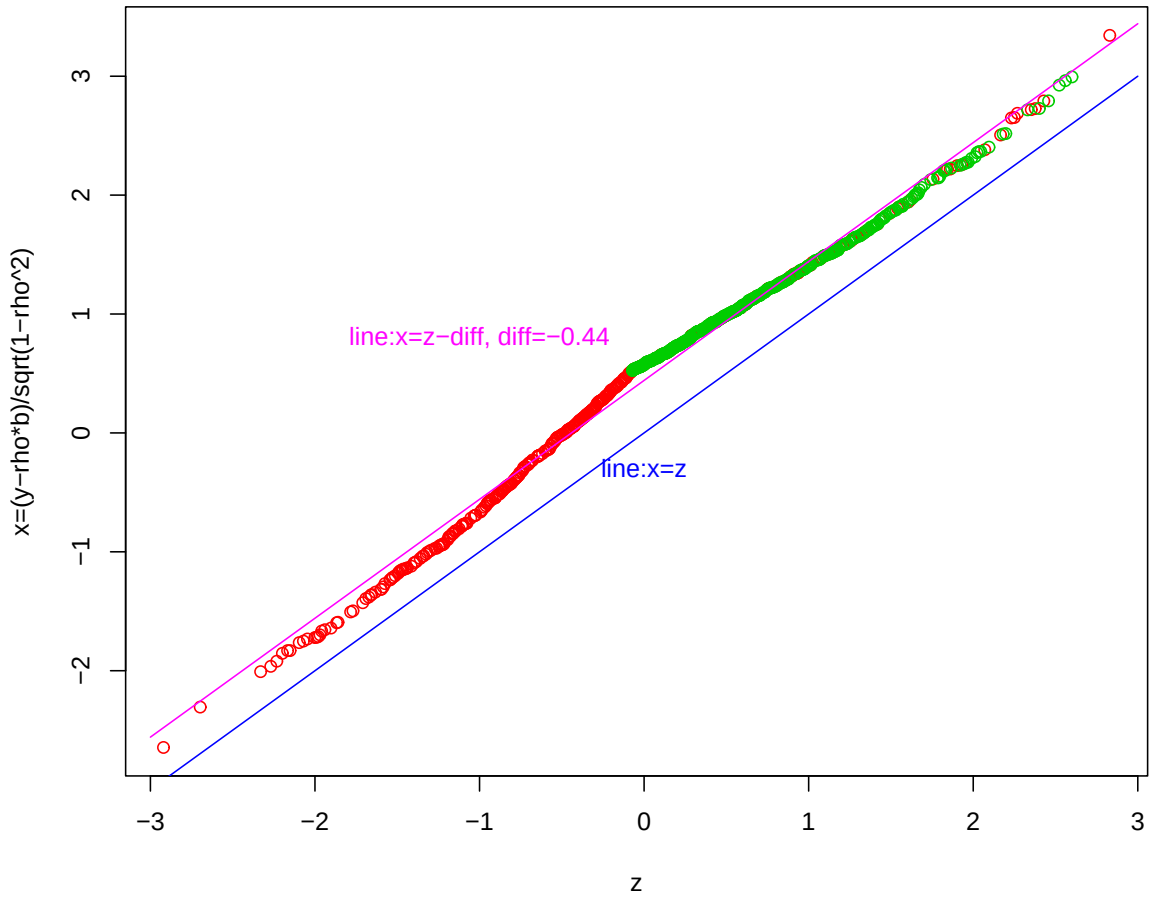
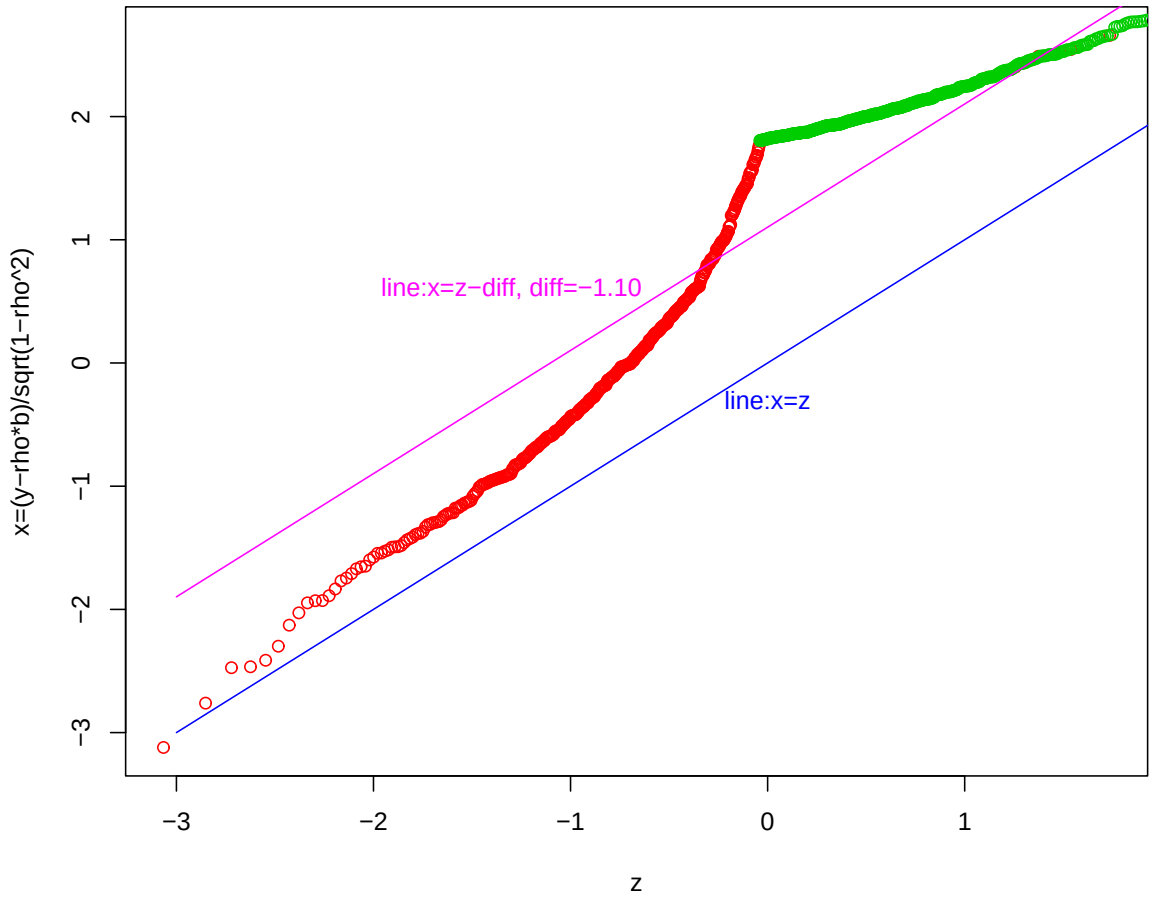


Figure 2: Comparison between the linear transformation and the copula transformation ($s = 2,\ b = 2.33$). The z-axis stands for the copula transformed data and the x-axis stands for the linearly transformed data. The phenotype $\mathbf{Y}$ for each sibpair is normally distributed with $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25$. 500 sibpairs are ascertained through $\{Y_1 > 2.33\}$.



get the same MLE of the finite-dimensional nuisance parameters (i.e., the population parameters $\mu,\ \sigma_Y^2$, etc.). As for the transformation function H_X, since H_X satisfies that $H_X(\mathbf{X})$ marginally normal, together with the fact that $\mathbf{X}$ is marginally standard normal (at least approximately), so H_X should be (approximately) the identity function. Therefore obtaining the test statistics based on the copula transformed $\mathbf{X}$ or obtaining that based on $H(\mathbf{Y})$ should be about equivalent.

5 Log-normal Model and Log-transformation

Some traits are usually modeled as having a log-normal distribution, for example, the age of onset of a disease. Therefore it is worthwhile studying what happens if we assume a log-normal distribution for the phenotype. Firstly, we briefly review the definition of a log-normal distribution.

$$(Y_1, Y_2)' \sim \log N\left((\tilde{\mu}_1, \tilde{\mu}_2)', \begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} \\ \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} & \tilde{\sigma}_2^2 \end{pmatrix}\right)$$

is said to have a bivariate log-normal distribution, if

$$(\log Y_1, \log Y_2)' \sim N\left((\tilde{\mu}_1, \tilde{\mu}_2)', \begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} \\ \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} & \tilde{\sigma}_2^2 \end{pmatrix}\right).$$

It is easy to see that for $i = 1, 2$

$$\mu_i = \mathrm{E}(X_i) = e^{\tilde{\mu}_i + \frac{\tilde{\sigma}_i^2}{2}}, \quad \sigma_i^2 = \mathrm{Var}(X_i) = \mu_i^2\big(e^{\tilde{\sigma}_i^2} - 1\big), \quad \rho = \mathrm{corr}(X_1, X_2) = \frac{e^{\tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho}} - 1}{\sqrt{\big(e^{\tilde{\sigma}_1^2} - 1\big)\big(e^{\tilde{\sigma}_2^2} - 1\big)}}. \tag{5.13}$$

Same as the multivariate normal and multivariate-t distributions, by definition, the log-normal distribution is uniquely determined by its mean and covariance. The marginal distribution of each component is skewed as well as heavy-tailed.

Under the log-normal model, we assume that, given the IBD sharing matrix A_ν at the trait locus τ, the phenotypic distribution of a sibship is log-normal ($\nu = \nu(\tau)$)

$$\mathbf{Y}|A_\nu \sim \log N\big(\tilde{\mu}, \tilde{\Sigma}_\nu\big). \tag{5.14}$$

Denote the phenotypic mean and covariance to be $\mathrm{E}(\mathbf{Y}) = \mu\mathbf{1},\ \mathrm{Cov}(\mathbf{Y}, \mathbf{Y}|A_\nu) = \Sigma_\nu = (\sigma^2\rho_\nu(i,j))$, then by (5.13)

$$\tilde{\mu} = \log\Big(\frac{\mu}{\sqrt{\sigma^2/\mu^2+1}}\Big),\ \tilde{\sigma}^2 = \log\Big(\sigma^2/\mu^2{+}1\Big),\ \tilde{\rho}_\nu(i,j) = \frac{\log(\rho_\nu(i,j)\sigma^2/\mu^2+1)}{\log(\sigma^2/\mu^2+1)},\ \tilde{\Sigma}_\nu(i,j) = \tilde{\sigma}^2\tilde{\rho}_\nu(i,j).$$

Note that for sibships, $\rho_\nu(i,j) = \rho + \frac{\alpha_0}{\sigma^2}(\nu_{ij}-1)$. Considering the Taylor expansion of $\tilde{\rho}_\nu(i,j)$ with respect to α_0, we get $\tilde{\rho}_\nu(i,j) = \tilde{\rho} + \frac{1}{\big(\rho\sigma^2+\mu^2\big)\log\big(\frac{\sigma^2}{\mu^2}+1\big)}(\nu_{ij}-1)\alpha_0 + o(\alpha_0)$, where $\tilde{\rho} = \frac{\log\big(\rho\frac{\sigma^2}{\mu^2}+1\big)}{\log\big(\frac{\sigma^2}{\mu^2}+1\big)}$. Let $f(\mu,\sigma,\rho) = \frac{1}{\big(\rho\sigma^2+\mu^2\big)\log\big(\frac{\sigma^2}{\mu^2}+1\big)}$ and $\tilde{\alpha}_0 = f(\mu,\sigma,\rho)\alpha_0$, then

$$\frac{\partial\tilde{\Sigma}_\nu}{\partial\alpha_0}|_{\alpha_0=0} = \tilde{\sigma}^2 f(\mu,\sigma,\rho)A_\nu.$$

Let $\tilde{\Sigma} = \tilde{\Sigma}_\nu|_{\alpha_0=0}$ and $X = \log(Y)$. Then it is easy to see that the efficient score of $\mathbf{Y}$ evaluated with $\alpha_0 = 0$ at a putative trait locus t is

$$\begin{aligned}\ell'(0) &= \tilde{\sigma}^2 f(\mu,\sigma,\rho)\Big\{-\frac{1}{2}\mathrm{tr}\big(\tilde{\mathbf{\Sigma}}^{-1}A_{\nu(t)}\big) + \frac{1}{2}(X-\tilde{\mu})'\tilde{\mathbf{\Sigma}}^{-1}A_{\nu(t)}\tilde{\mathbf{\Sigma}}^{-1}(X-\tilde{\mu})\Big\}\\ &= \tilde{\sigma}^2 f(\mu,\sigma,\rho)\tilde{\ell}'(0),\end{aligned} \tag{5.15}$$

where $\tilde{\ell}'(0)$ is the corresponding efficient score of the log-transformed data $\mathbf{X}$. Therefore the normalized efficient score of $\mathbf{Y}$ is the same as that of $\mathbf{X}$. (However, the likelihood ratio statistic of $\mathbf{Y}$ is not the same as that of $\mathbf{X}$.) Also the MLE of $\mathbf{X}$ implies the MLE of $\mathbf{Y}$ via (5.13). Therefore, for the log-normal data with parameters (μ,Σ,α_0), the power of the statistic based on (5.15) is the same as that of the statistic based on the normal score (2.2) for the multivariate normal data with corresponding parameters $\big(\tilde{\mu},\tilde{\Sigma},\tilde{\alpha}_0\big)$.

If a phenotype Y takes value on $(0,\infty)$, we can apply the *log-transformation* on Y directly by $X = \log(Y)$ and then fit the normal score on the transformed data X. If the phenotype Y can take negative values, we first take a location transformation to make all the phenotypic values positive, then apply the log-transformation. Let $y_0 = \min(\min(\mathbf{Y}), 0)$, then define the log-transformation as

$$X = \log\Big(Y - y_0 + \epsilon\Big), \tag{5.16}$$

where $\epsilon > 0$ is small. For the simulation study in Section 6, we take $\epsilon = 10^{-6}$. The point to

use a small positive ϵ is that we want to map the data onto the whole real line $\mathbb{R}$. Since this transformation is affected a lot by outliers, a preliminary detection of outliers may be necessary. After the log-transformation has been applied, we fit the robust normal score statistic (2.6) on the transformed data. If pedigrees are ascertained, we should apply ascertainment corrections on the transformed data for the purpose of parameter estimation via conditional MLE as discussed before.

When the phenotypic distribution is indeed log-normal, the log-transformation results in the actual normalized efficient score which is asymptotically locally optimal. By the discussion in the previous section, when the phenotypic distribution is log-normal and the pedigrees are randomly sampled, the copula transformation is equivalent to the log-transformation.

6 Numerical Study

In this section we study the power of the t-transformation, copula transformation and log-transformation for various phenotypic distributions w/o ascertainment. For genotypic data, 23 idealized human chromosomes of 150cM each are used. For each chromosome, there are 31 fully informative, equally spaced markers. Under the alternative, one purely additive QTL is located on the 16th marker of chromosome 1.

In this section, N_i and T_i $(i = 1, 2, \cdots)$ stand for the normal score and t-score applied to the data after $i-1$ times t-transformations, respectively, with data after zero times t-transformation meaning the original data; "copula" stands for the normal score applied to the copula-transformed data; "log" stands for the normal score applied to the log-transformed data. For all transformations, ascertainment corrections are applied when pedigrees are ascertained.

Phenotypic data is generated from four different distributions w/o ascertainment (Figure 3). For ascertainment, sibships are ascertained through the first sibling's phenotype being above a threshold.

1. *Multivariate normal traits*: phenotypic data is generated from the multivariate normal model, thus the phenotypic skewness and kurtosis are both zero. The population parameters are set to be $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25$ and the linkage parameter under the alternative is $\alpha_0 = 0.1$ which means that the QTL effect accounts for 20% of the phenotypic variance.

Consider three cases:

case 1.1: 1000 sibpairs are ascertained through $\{Y_1 > 1.65\}$, where 1.65 is the 95%−percentile of the phenotypic distribution.

case 1.2: 500 sibpairs are ascertained through $\{Y_1 > 2.33\}$, where 2.33 is the 99%−percentile of the phenotypic distribution.

case 1.3: 400 sib-trios are ascertained through $\{Y_1 > 0.674\}$, where 0.674 is the 75%−percentile of the phenotypic distribution.

2. *Multivariate-t traits*: phenotypic data is generated from the multivariate-t model with degrees of freedom $k = 4$, thus the phenotypic skewness and kurtosis are zero and ∞, respectively. The population parameters are set to be $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25$ and the linkage parameter under the alternative is $\alpha_0 = 0.1$. Consider two cases:

   case 2.1: 1000 sib-quads are randomly sampled.

   case 2.2: 400 sib-quads are ascertained through $\{Y_1 > 0.741\}$, where 0.741 is the 75%−percentile of the phenotypic distribution.

3. *Multivariate gamma traits*: phenotypic data is generated from a multivariate gamma model. The model is set such that the phenotypic skewness and kurtosis are around 6 and 50, respectively. The population parameters are set to be $\mu = 0,\ \sigma_Y = 1,\ \rho = 0.25$ and the linkage parameter under the alternative is $\alpha_0 = 0.1$. For a detailed definition of the multivariate gamma model, see Peng and Siegmund (2006). Consider two cases:

   case 3.1: 1000 sib-quads are randomly sampled.

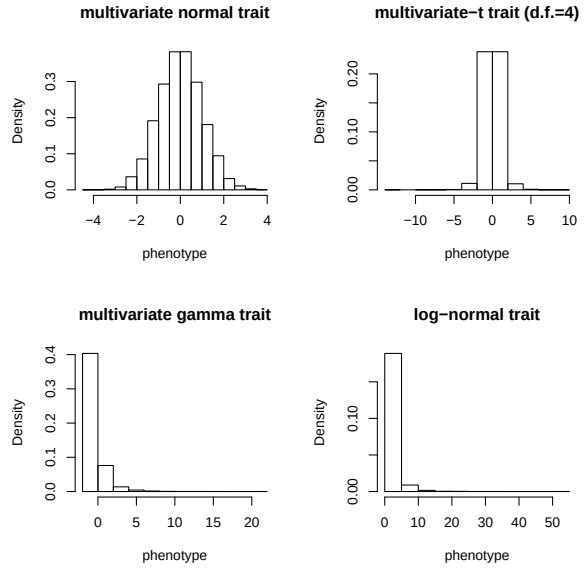
   case 3.2: 400 sib-quads are ascertained through $\{Y_1 > -0.168\}$, where -0.168 is the 75%−percentile of the phenotypic distribution.

4. *Log-normal traits*: phenotypic data is generated from the log-normal model. The population parameters are set to be $\mu = 1.649,\ \sigma_Y = 2.161,\ \rho = 0.165$ and the linkage parameter under the alternative is $\alpha_0 = 0.0747$. The phenotypic skewness and kurtosis are 5.7 and 72, respectively. (The log-transformed phenotype has $\tilde{\mu} = 0,\ \tilde{\sigma}_Y = 1,\ \tilde{\rho} = 0.25$ and $\tilde{\alpha} = 0.1$ (under the alternative).) Consider three cases:

   case 4.1: 1000 sib-quads are randomly sampled.

   case 4.2: 400 sib-trios are ascertained through $\{Y_1 > 1.963\}$, where 1.963 is the 75%−percentile

Figure 3: Marginal distributions of four different phenotypes.



of the phenotypic distribution.

case 4.3: 500 sibpairs are ascertained through $\{\mathbf{Y}_1 > 10.24\}$, where 10.24 is the 99%−percentile of the phenotypic distribution.

In Table 2, 60000 replications under the null are used to determine the thresholds b for the genome-wide 0.05 significance level (over 23 chromosomes, each with 31 fully informative, equally spaced markers with a inter marker space $\Delta = 5$cM.): $P(\max_t Z(t) > b) \approx 0.05$. As can be seen from Table 2, the resulting thresholds are quite close to 3.809 which is suggested by formula (2.7).

In Table 3, 500 replications under the alternative are used to get the power for the genome-wide 0.05 significance level. As can be seen from Table 3, the copula transformation has similar performance as the normal score on the original data (N_1) for the multivariate normal traits when the sibship size s is at least moderate and the ascertainment rule is moderate (case 1.3). However when the sibship size s becomes smaller and the ascertainment rule becomes more stringent, there is more and more loss of power by using the copula transformation compared to N_1 (case 1.1 and case 1.2). These observations are consistent with the discussion in Section 4. For the multivariate-t phenotypes (case 2.1 and case 2.2), the t-score (T_1) is more powerful than "copula"

Table 2: Thresholds for the genome-wide 0.05 significance level; by 60000 replications under the null

Case	N_1	T_1	N_2	T_2	N_3	T_3	copula	log
case 1.1	3.81	3.81	na	na	na	na	3.81	na
case 1.2	3.81	3.81	na	na	na	na	3.81	na
case 1.3	3.81	3.81	na	na	na	na	3.81	na
case 2.1	3.81	3.81	na	na	na	na	3.81	na
case 2.2	3.81	3.81	na	na	na	na	3.81	na
case 3.1	4.10	3.85	3.85	3.80	3.80	3.85	3.90	3.85
case 3.2	4.10	3.85	3.85	3.80	3.80	3.80	3.90	3.85
case 4.1	3.70	3.75	3.75	3.75	3.75	3.75	3.80	3.80
case 4.2	3.70	3.80	3.75	3.75	3.75	3.75	3.80	3.80
case 4.3	na	na	na	na	na	na	3.80	3.80

which in turn is more powerful than N_1. This is under expectation, since now T_1 is based on the true efficient score thus is optimal. For both cases, the log transformation has no power at all. For the multivariate gamma phenotypes (case 3.1 and case 3.2), the normal score N_1 has very low power since the phenotypic distribution is badly skewed. "copula" and "log" are much less powerful than the final t-score T_3. The fitted degrees of freedom $\hat{k}$ becomes larger when more t-transformations are applied and reaches a very large number at the third step. The gain in power becomes smaller as more t-transformations have been applied. These suggest that the t-transformation is converging. For case 4.1, the copula transformation has almost the power as the log-transformation since they are roughly equivalent for the randomly sampled log-normal data as discussed in Section 5. Both transformations are much more powerful than the t-transformation which converges at T_2. Note that, under this case the copula transformation and log-transformation are equivalent to the statistic based on the true efficient score. All three transformations are much more powerful than the normal score N_1 which has no power at all in this case. For case 4.2, the copula transformation is a little bit less powerful than the optimal log-transformation, while both are much more powerful than the t-score T_3. Again the normal score N_1 has no power at all. Similarly as in the multivariate normal case, if small sibships are ascertained through a stringent ascertainment rule, there is a larger drop of power by the copula transformation (case 4.3).

Table 3: Power under the genome-wide 0.05 significance level; by 500 replications under the alternative

Case	N_1	T_1	N_2	T_2	N_3	T_3	copula	log
case 1.1	0.876	0.872	na	na	na	na	0.83	na
case 1.2	0.77	0.77	na	na	na	na	0.706	na
case 1.3	0.408	0.412	na	na	na	na	0.442	na
case 2.1	0.454	0.952	na	na	na	na	0.754	0
case 2.2	0.192	0.650	na	na	na	na	0.48	0
case 3.1	0.108	0.71	0.778	0.864	0.892	0.908	0.762	0.794
case 3.2	0.134	0.814	0.846	0.924	0.94	0.942	0.694	0.568
case 4.1	0.076	0.386	0.424	0.436	0.442	0.448	0.756	0.758
case 4.2	0.048	0.168	0.220	0.220	0.22	0.22	0.376	0.414
case 4.3	na	na	na	na	na	na	0.70	0.77

7 Conclusions

The results in the above section show that, applying the three transformations: t-transformation, copula transformation and log-transformation leads to statistics which are much more powerful than the normal score applied to the original data (N_1) when the phenotypic distribution is non-normal. The t-transformation outperforms the copula transformation and the log-transformation for the multivariate normal, multivariate-t and multivariate gamma phenotypes (cases 1–3). However, the copula transformation and log-transformation are more powerful than the t-transformation for the log-normal traits (case 4), in which case the former two are actually equivalent to the optimal statistic based on the efficient score. We also find out that, we should apply the t-transformation successively several times until it converges as this yields the largest power.

References

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees, *Am. J. Hum. Genet.* **62**, 1198-1211.

Diao, G. and Lin, D.Y. (2005). A powerful and robust method for mapping quantitative trait loci in general pedigrees,*Am. J. Hum. Genet.* **77**, 97-111.

Fisher, R.A. (1918). The correlation of relatives on the assumption of Mendelian inheritance,

Proc. Roy. Soc. Edinburgh.

Kempthorne, O. (1957). *Genetic Statistics*, John Wiley and Sons, New York.

Lange, K., Little, R.J.A. and Taylor, J.M.G. (1989). Robust statistical inference using the t distribution, *J. Am. Statist. Assoc.* **84**, 881-896.

Miller, R.G. (1986). *Beyond ANOVA: Basics of Applied Statistics*, Chapman & Hall/CRC.

Peng, J. and Siegmund, D. (2006). Mapping quantitative traits under ascertainment, *Ann. Hum. Genet.*, doi:10.1111/j.1469-1809.2006.00286.x.

Tang, H.-K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models, *Biostatistics* **2**, 147-162.

Wang, K. and Huang J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size, *Am. J. Hum. Genet.* **70**, 412-424.