Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

A NOVEL THREE-CLASS ROC METHOD for eQTL ANALYSIS

WEICHAO XU[1], PEIKAI CHEN[1], Y. S. HUNG[1], and S. Y. KUNG[2]

[1]Department of Electrical & Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
[2]Department of Electrical Engineering, Princeton University, NJ 08544, US
E-MAIL: wcxu@eee.hku.hk, pkchen@eee.hku.hk, yshung@eee.hku.hk, kung@princeton.edu

Abstract

The problem of identifying genetic factors underlying complex and quantitative traits such as height, weight and disease susceptibility in natural populations has become a major theme of research in recent years. Aiming at revealing the inter-dependency and causal relationship between the underlying genotypes and observed phenotypes, researchers from different areas have developed a variety of methods for expression quantitative trait loci (eQTL) mapping. Most of these methods rely on resampling-based algorithms that are computationally very expensive. To overcome the disadvantages of the current techniques, we propose a novel nonparametric method based on the volume under surface (VUS) within the framework of three-class receiver operating characteristic (ROC) analysis. With the fast algorithms developed, we can reduce the computation time of the genomewide analysis from several months down to several days.

Keywords:

expression quantitative trait loci (eQTL); receiver operating characteristic (ROC); volume under surface (VUS); nonparametric; normal distribution.

1. Introduction

Genetic variability in natural populations affects the observed characteristics (traits) such as height, weight, crop yield and disease susceptibility [1]. The problem of identifying the genetic factors underlying complex and quantitative traits has long been the major theme of research among geneticists, biologists, clinicians, statisticians, and engineers, to name a few [2]. Aiming at revealing the inter-dependency and causal relationship between the underlying genotypes and observed phenotypes, researchers from different areas have developed a plenty of methods, which are generally termed as *expression quantitative trait loci* (eQTL) mapping [3]. At the heart of these mapping techniques is to quantify the intensity of association between genotype patterns and expression levels in a genetically diverse population [4] (See Fig. 1(a)). These methods, while of various rationales, fall roughly into three categories: 1) information-theory-based methods [5], 2) signal-processing-based methods [6], 3) statistics-based methods [2, 7–13].

There are many strengths and weaknesses of the eQTL mapping methods mentioned above. The information-theory-based methods employ the *mutual information* (MI) to quantify the degree of association between phenotypes and genotypes [5]. While possessing desired statistical properties (the logarithm of $2 \times \mathrm{MI}$ obeys χ^2 distribution under the null hypothesis of no association), this category of MI-based methods suffers the drawback of heavy (exponential) computational load [5]. On the other hand, the signal-processing-based methods, depending on independent component analysis (ICA) [6] and network component analysis (NCA) [14], have relatively efficient computational algorithms. Nevertheless these blind-source-separation-based methods perform the analysis between the phenotype and genotype in an indirect manner (involvements of estimation of the latent sources). This sometimes may result in artifacts that can obscure biological interpretations.

Compared to the methods just remarked, the leading role in eQTL mapping is played by the statistics-based methods, which, among others, consist of contingency-table-based techniques [7,8], Bayesian approaches [2,9,10] and regression-based methods [11–13]. Similar to MI, the contingency-table-based test statistic can also be transformed to a χ^2 distributed random variable under the null hypothesis of no association [5]. This approach, however, can only treat discrete phenotypes, and is therefore of limited service in practice. To overcome such limitation, other methods that based on Bayesian rule can be resorted to, with advantages of avoiding the difficult and often intractable mathematical treatments [15]. Nevertheless, the methods

978-1-4244-6527-9/10/$26.00 ©2010 IEEE

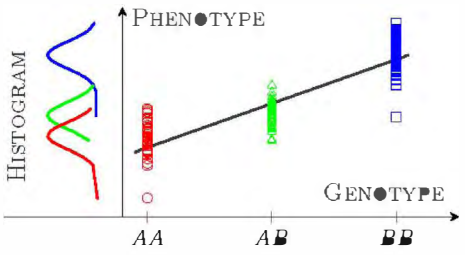
within Bayesian framework require the assignments of subjective *a priori* probabilistic terms, which might output misleading results. Due to the solid theoretical foundation and mathematical tractability, the linear regression models with assumption of Gaussian distribution of data are perhaps the most widely used methods in eQTL analysis. However, the linearity and normality assumed in the linear-regression-based methods are hardly justified by physical evidence. Besides, it is well known that the linear regression methods are very sensitive to *outliers*. Even a single outlier can distort severely the regression slope and thus result in problematic conclusion. In addition to the shortcomings just mentioned, nearly all the above methods rely on algorithms based on resampling techniques that are computationally demanding. Aiming at overcoming these limitations, in this project we propose to develop a new statistical framework for eQTL mapping based on a new 3-class ROC (receiver operating characteristic) approach. The rationale for using a 3-class ROC is that underlying each eQTL mapping, there is a 3-class classification problem— if the gene expression of a certain gene can be classified according to the class labels (AA, AB and BB) of a SNP locus, then this constitutes an eQTL [cf. Fig. 1(a)].

By developing a statistic capable of quantifying the "separability" of the expression levels with respect to three classes, in this paper we 1) formulate a framework for the eQTL mapping problem using a 3-class ROC approach, and derive the necessary statistical formulas for hypothesis testing; 2) develop fast computational algorithms that implement the 3-class ROC approach in a feasible manner for eQTL analysis on high-dimensional datasets; and 3) evaluate the statistical power of the new method by Monte Carlo simulations.

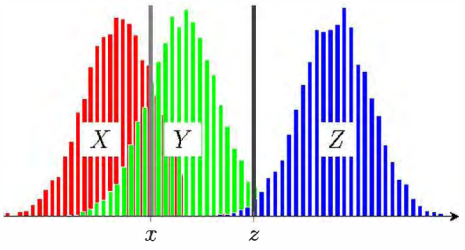
2. Methodology

2.1 Volume under ROC Surface

Let $\{x_i\}_{i=1}^m$, $\{y_i\}_{j=1}^n$ and $\{z_k\}_{k=1}^l$ be independent and identically distributed (IID) samples drawn from three classes with continuous cumulative distribution functions (cdf) $F_X(x)$, $F_Y(y)$ and $F_Z(z)$, respectively. These samples may be the observed gene expressions values corresponding to three groups of people with different genotypes, e.g. AA, AB, and BB, respectively (see the vertical axis in Fig. 1(a)). Assume we are to design a classifier that distinguishes between the three classes based on two thresholds th_1 $(= x)$ and th_2 $(= z)$, where $-\infty < x \leq z < +\infty$. As illustrated in Fig. 1(b), the two thresholds x and z divide the whole real axis into three parts. For a newly measured value u, a natural classification criterion is to decide u to class X, Y and Z if u falls in the region of $(-\infty, x)$, (x, z), and (z, ∞), respectively. Define



(a) Linear regression model



(b) Double-threshold based classifier

Figure 1: Illustration of the conventional linear regression model and the three-class classifier. The histogram in (b) is an enlarged and transposed version of the histogram in (a).

$$P_1(x, z) \triangleq \Pr(X < x) \tag{1}$$

$$P_2(x, z) \triangleq \Pr(x < Y < z) \tag{2}$$

$$P_3(x, z) \triangleq \Pr(Z > z). \tag{3}$$

Then it is clear that P_1, P_2 and P_3 are the probabilities that the classifier correctly classifies each sample to its true class. For each pair of the thresholds (x, z), there exists a corresponding triplet (P_1, P_2, P_3) in a three-dimensional space. With the variations of x and z, a surface, called ROC surface, can be described by the simultaneous equations (1)–(3). Fig. 2(a) shows schematically a three-dimensional ROC surface based on the procedure just mentioned. The volume of the solid that formed by the ROC surface and the three plenary walls plays a core role in three-way ROC analysis. This volume under the ROC surface (VUS) [16], is determined by

$$\theta = \int_0^1 \int_0^1 P_2(x, z) dP_1(x, z) dP_3(x, z) \tag{4}$$

which can be expressed, in terms of F_X, F_Y and F_Z, as

$$\theta = \int_{\substack{-\infty \\ x<z}}^{\infty} \int_{-\infty}^{\infty} [F_Y(z) - F_Y(x)] \, dF_X(x) dF_Z(z). \tag{5}$$

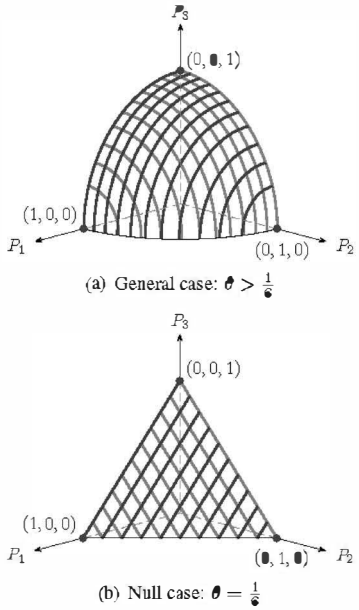


Figure 2: Geometric meaning of the three-class ROC analysis. (a) General appearance of the ROC surface. If the centroids of the three classes are far away from each other, the VUS is larger than 1/6. (b) Null case. If the three classes are totally overlapped, the surface degenerates to the plane $P_1+P_2+P_3=1$, and the volume of the tetrahedron equals 1/6.

In addition to its geometric meaning just established, it follows that (5) can also be interpreted as

$$\theta = \Pr(X < Y < Z) \tag{6}$$

which means that θ is the probability that the three random variables X, Y and Z are in ascending order. Then it follows that $\theta = 1$ if, from left to right, X, Y and Z are completely separable, and $\theta = 1/6$ if X, Y and Z are all overlapped together (the null case). The latter result can be obtained directly from Fig. 2(b).

2.2. Estimating VUS from Samples

Thus far we have established the population version of VUS, that is, we assume that the cumulative distribution functions of each class F_X, F_Y and F_Z are all known in our derivations. However, the forms of the distribution functions are rarely known in practice. We only have samples at hand and hence have to estimate the VUS based on the available samples. From (6), a nonparametric estimator of the VUS can be constructed, as

$$\hat{\theta} = \frac{1}{mnl}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} H(y_j - x_i)H(z_k - y_j) \tag{7}$$

where $H(t) = 1$ for $t > 0$ and $H(t) = 0$ for $t \leq 0$. It is easy to verify that $\hat{\theta}$ is an unbiased estimator of θ, namely $E(\hat{\theta}) = \theta$. Moreover, $\hat{\theta}$ is asymptotically normally distributed based on the theory of U-statistics [17]. Therefore, the distribution of $\hat{\theta}$ can be completely determined if we know the variance of $\hat{\theta}$. By a series of nontrivial derivations, it follows that the variance of $\hat{\theta}$ is

$$\begin{aligned}\mathrm{var}(\hat{\theta}) = \frac{1}{mnl}\big[&\theta(1-\theta)+(l-1)(q_{12}-\theta^2)+(n-1)(q_{13}-\theta^2)\\ &+(m-1)(q_{23}-\theta^2)+(n-1)(l-1)(q_1-\theta^2)\\ &+(m-1)(l-1)(q_2-\theta^2)+(m-1)(n-1)(q_3-\theta^2)\big]\end{aligned} \tag{8}$$

where

$$q_{12} = \int F_X(y)\left[1-F_Z(y)\right]^2 dF_Y(y) \tag{9}$$

$$q_{13} = \iint \left[F_Y(z)-F_Y(x)\right]^2 dF_X(x)dF_Z(z) \tag{10}$$

$$q_{23} = \int F_X^2(y)\left[1-F_Z(y)\right] dF_Y(y) \tag{11}$$

$$\begin{aligned}q_1 = \iint\limits_{x<\min(z,z')} &\left[F_Y(z)-F_Y(x)\right]\left[F_Y(z')-F_Y(x)\right]\\ &\times dF_X(x)dF_Z(z)dF_Z(z')\end{aligned} \tag{12}$$

$$q_2 = \int F_X^2(y)\left[1-F_Z(y)\right]^2 dF_Y(y) \tag{13}$$

and

$$\begin{aligned}q_3 = \iiint\limits_{z>\max(x,x')} &\left[F_Y(z)-F_Y(x)\right]\left[F_Y(z)-F_Y(x')\right]\\ &\times dF_X(x)dF_X(x')dF_Z(z).\end{aligned} \tag{14}$$

From the theoretical point of view, we have obtained the asymptotic behavior of $\hat{\theta}$, that is, when the sample sizes m, n, l are sufficiently large, $\hat{\theta}$ converges in distribution to a normal distribution with mean θ and variance $\mathrm{var}(\hat{\theta})$ determined by (8)–(14). Apparently, (8) has little practical value, since the expressions of the q-terms involves F_X, F_Y and F_Z, whose functional forms are often unknown in advance. Fortunately, it can be shown that, under the null case ($F_X = F_Y = F_Z$),

$$\sigma^2_{\mathrm{null}} = \frac{1}{180}\frac{1}{mnl}\left(4+5m+5l+2n+4mn+4nl+ml\right) \tag{15}$$

which depends only on the sample sizes m, n and l. In other words, $\hat{\theta}$ is a *distribution free* estimator under the null case. This desirable property constitutes the theoretical basis of

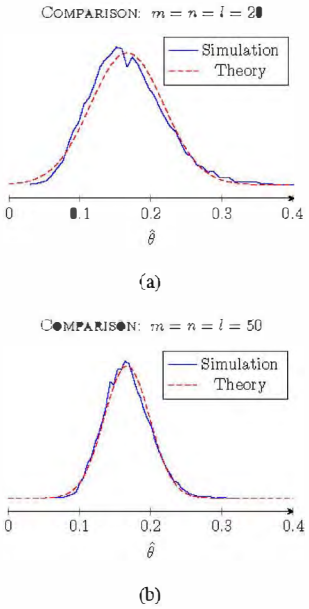


Figure 3: Illustration of the convergence of $\hat{\theta}$ to normal distributions. For convenience the sample sizes of m, n and l are set to be identical. The number of Monte Carlo trials is 10^4 for the simulation curves in (a) and (b).

the methodology in this work. The paradigm of the hypothesis test is thus shifted to

$$\begin{cases} H_0 : \theta = \frac{1}{6} \\ H_1 : \theta \neq \frac{1}{6}. \end{cases} \tag{16}$$

The false positive rate (Type I error) can be accurately controlled directly from the normal distribution with mean $1/6$ and variance σ^2_{null}, thus avoiding the time consuming resampling procedures commonly employed in the literature of biostatistics and system biology [18]. The associated criterion is: reject H_0 if $\hat{\theta} > \text{th}$ (for one-sided tests) or $|\hat{\theta} - 1/6| > \text{th}$ (for two-sided tests). The threshold th is dependent on a pre-specified false positive rate that often designated as α.

2.3 Fast Algorithm of Computing $\hat{\theta}$

Apparently the estimator $\hat{\theta}$ established in (5) is simple to implement. However, from the viewpoint of time complexity, $\hat{\theta}$ is of order $O(mnl)$, which is computationally very expensive and impractical even for medium-sized samples. Fortunately, an efficient algorithm can be constructed based on an identity with respective to the population version θ. With some tedious derivations, it follows that the right side of (5) simplifies to

$$\theta = \int_{-\infty}^{\infty} F_X(y)\left[1 - F_Z(y)\right] dF_Y(y) \tag{17}$$

suggesting the following sample version

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} \hat{F}_X(y_j)\left[1 - \hat{F}_Z(y_j)\right] \tag{18}$$

where $\hat{F}_X(y_j)$ and $\hat{F}_Z(y_j)$ are estimations of $F_X(y)$ and $F_Z(y)$, respectively. Now we proceed to developing the efficient algorithms for calculating $\hat{F}_X(y_j)$ and $\hat{F}_Z(y_j)$ based on samples $\{x_i\}_{i=1}^m$, $\{y_i\}_{j=1}^n$ and $\{z_k\}_{k=1}^l$ at hand. For convenience, write $W_1 = \{y_j\} \cup \{z_k\}$, $W_2 = \{x_i\} \cup \{z_k\}$ and $W_3 = \{x_i\} \cup \{y_j\}$. Let $r_x(i)$ $(= i)$ denote the *rank* of x_i, if x_i is the ith smallest in the x-array [19–22]. Similarly symbols $r_y(j)$ and $r_z(k)$ are employed to denote the respective ranks of y_j and z_k in the corresponding y-array and z-array. Denote $R_y(j)$, $R_z(k)$ as the respective ranks of y_j and z_k in the combined W_1-array; $S_x(i)$, $S_z(k)$ as the respective ranks of x_i and z_k in the combined W_2-array; $T_x(i)$, $T_y(j)$ as the respective ranks of x_i and y_j in the combined W_3-array. Then

$$\hat{F}_X(y_j) = \left[T_y(j) - r_y(j)\right]/m \tag{19}$$

$$\hat{F}_Z(y_j) = \left[R_y(j) - r_y(j)\right]/l \tag{20}$$

which, along with (18), yield

$$\hat{\theta} = \frac{1}{mnl}\sum_{j=1}^{n}\left[T_y(j) - r_y(j)\right]\left[l - R_y(j) + r_y(j)\right]. \tag{21}$$

It can be shown that (21) is numerically equivalent to (7), nevertheless, the computational load of the former is much lighter than that of the latter. Compared to $O(mnl)$, the cubic time complexity of (7), the most time-consuming operation in (21) is ranking y_j's in W_1-array and W_3-array, whose time complexity is $O(N \log N)$, where $N = \max(m, n, l)$.

3. Numerical Results

3.1 Verification of Normality

We have established the asymptotic normality of the sample estimate of the VUS, $\hat{\theta}$, base on the U-statistic theory [17]. The term "asymptotic" means that only when the sample size is large does the property of normality hold. Now an important question arise: how large is "large" of the sample size so as to ensure the validity of normal approximation when doing the hypothesis test in practice? To answer this question, we perform Monte Carlo simulations

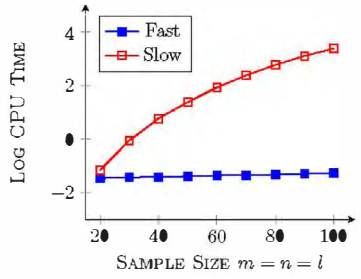


Figure 4: Comparative results of CPU time between the algorithms of (7) and (21). A log scale is used for better visual effect.

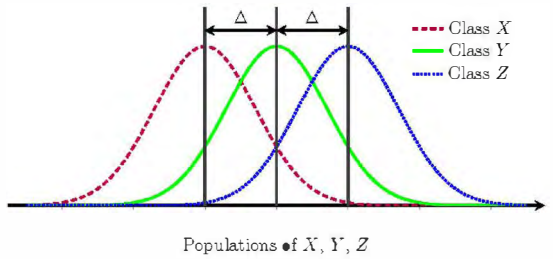


Figure 5: Setup for testing the discriminatory power of $\hat{\theta}$ in nonnull cases. For brevity, the differences between the centroids of the adjacent two classes are the same, i.e., $\mu_Y - \mu_X = \mu_Z - \mu_Y = \Delta\mu$.

for small-sized samples (below 50) drawn from normal populations. For convenience, we set $m = n = l$ in our simulations. As shown in Fig. 3, the normal curves agrees well with the corresponding simulation curves, even when the sample sizes are as small as 20. In other words, when the sample size of each class is larger than 20, a reasonable scenario in practice, it is safe to make use of the normal assumption when performing the hypothesis-test-based analyses.

3.2 Time Complexity Comparison

Fig. 4 compares the computational loads between the two sample versions of $\hat{\theta}$ in (7) and (21). For simplicity, the sample sizes of the three classes are set to be identical, that is, $m = n = l$. It is seen that when the sample size is small, the CPU time of (7) and (21) are comparable. However, with increase of the sample size, the computational time of (7) soars up rapidly, suggesting its inferiority in terms of computational load.

3.3 Statistical Power

Up to this stage, we have focused on the null distribution of $\hat{\theta}$, i.e., under the assumption that all the three classes obey

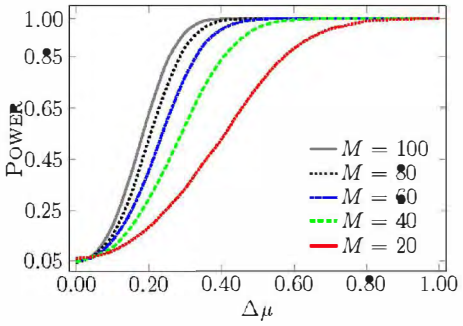


Figure 6: The relationship among the power of $\hat{\theta}$, the location shift parameter $\Delta\mu$ and the sample size M.

an identical distribution:

$$F(X) = F(Y) = F(Z).$$

However, for a statistic to be useful, the performance in the nonnull cases, often characterized by means of statistical *power* (or true positive rate), is also very important, if not more. In other words, it is desirable that the statistic is sensitive to small diviations of the null case. To investigate the discriminatory power of $\hat{\theta}$, we employ a location shift model in this study. As shown in Fig. 5, the shapes of the three populations are the same except for the three location parameters μ_X, μ_Y and μ_Z. Due to the lack of space, we only exam the normal cases under the equal-sample-size setup here. Specifically, $F(X) \sim N(\mu_X, 1)$, $F(Y) \sim N(\mu_Y, 1)$, $F(Z) \sim N(\mu_Z, 1)$ and $m = n = l \triangleq M$. For each $\Delta\mu$ and each M, the simulations are undertaken 10^4 times and the fraction of rejections is computed as an estimate of the power. It is seen in Fig. 6 that 1) the power of $\hat{\theta}$ increases monotonically with both the shift parameter $\Delta\mu$ and the sample size M; 2) all curves lie above 0.70 for $\Delta\mu > 0.5$, or SNR> -60 dB, indicating the high discriminatory power of $\hat{\theta}$; and 3) all curves start from around 0.05, the pre-assigned false positive probability.

4. Conclusion

In this paper we established a novel framework for the eQTL mapping problem using a 3-class ROC approach. Theoretical derivations and simulation results suggest that this framework is both theoretically solid and computationally feasible. The advantages of this new method enable it to be a useful alternative to the existing methods in the literature for genome-wide eQTL analysis, which often involves a massive amount of data to investigate. With the efficient algorithms of (21) and (15), we can reduce the computation time from an order of several months down to an order of several days. Moreover, some new biological findings may

be made from the developed technique due to its high discriminatory power revealed in this work.

References

[1] J. J. Michaelson, S. Loguercio, and A. Beyer, "Detection and interpretation of expression quantitative trait loci (eQTL)," *Methods*, vol. 48, no. 3, pp. 265–276, 2009.

[2] S. Sen and G. A. Churchill, "A statistical framework for quantitative trait mapping," *Genetics*, vol. 159, no. 1, pp. 371–387, 2001.

[3] Members of the Complex Trait Consortium, "The nature and identification of quantitative trait loci: a community's view," *Nat. Rev. Genet.*, vol. 4, no. 11, pp. 911–916, 2003.

[4] M. Sarkis, B. Goebel, Z. Dawy, J. Hagenauer, P. Hanus, and J. Mueller, "Gene mapping of complex diseases—a comparison of methods from statistics informnation theory, and signal processing," *IEEE Signal Proc. Mag.*, vol. 24, no. 1, pp. 83–90, Jan. 2007.

[5] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. Mueller, "Gene mapping and marker clustering using shannon's mutual information," *IEEE/ACM Trans. Comput. Biology Bioinformatics*, vol. 3, no. 1, pp. 47–56, 2006.

[6] W. Liebermeister, "Linear modes of gene expression determined by ICA," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

[7] D. J. Balding, M. Bishop, and C. Cannings, *Handbook of Statistical Genetics*. New York: Wiley, 2001.

[8] J. K. Percus, *Mathematics of Genome Analysis*, ser. Cambridge studies in mathematical biology. Cambridge: Cambridge University Press, 2002.

[9] R. Yang and S. Xu, "Bayesian shrinkage analysis of quantitative trait loci for dynamic traits," *Genetics*, vol. 176, no. 2, pp. 1169–1185, 2007.

[10] M. K. Kuhner, P. Beerli, J. Yamato, and J. Felsenstein, "Usefulness of Single Nucleotide Polymorphism Data for Estimating Population Parameters," *Genetics*, vol. 156, no. 1, pp. 439–447, 2000.

[11] S. Wang, N. Yehya, E. E. Schadt, H. Wang, T. A. Drake, and A. J. Lusis, "Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity," *PLoS Genet.*, vol. 2, no. 2, p. e15, 2006.

[12] H. J. Cordell and D. G. Clayton, "A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes." *Am. J. Hum. Genet.*, vol. 70, no. 1, pp. 124–141, January 2002.

[13] S. Kim, K.-A. Sohn, and E. P. Xing, "A multivariate regression approach to association analysis of a quantitative trait network," *Bioinformatics*, vol. 25, no. 12, pp. 204–212, 2009.

[14] C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics*, vol. 24, no. 11, pp. 1349–1358, 2008.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, ser. Prentice-Hall signal processing series. Englewood Cliffs, N.J.: PTR Prentice-Hall, 1993.

[16] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, no. 1, pp. 78–89, 1999.

[17] S. Dreiseitl, L. Ohno-Machado, and M. Binder, "Comparing three-class diagnostic tests by three-way ROC analysis," *Med. Decis. Making*, vol. 20, no. 3, pp. 323–331, 2000.

[18] D. M. Gatti, A. A. Shabalin, T.-C. Lam, F. A. Wright, I. Rusyn, and A. B. Nobel, "FastMap: Fast eQTL mapping in homozygous populations," *Bioinformatics*, vol. 25, no. 4, pp. 482–489, 2009.

[19] W. Xu, C. Chang, Y. Hung, S. Kwan, and P. Fung, "Order statistic correlation coefficient and its application to association measurement of biosignals," *Proc. Int. Conf. Acoustics, Speech, Signal Process. (ICASSP) 2006*, vol. 2, pp. II–1068–II–1071, May 2006.

[20] W. Xu, C. Chang, Y. Hung, S. Kwan, and P. Chin Wan Fung, "Order statistics correlation coefficient as a novel association measurement with applications to biosignal analysis," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5552–5563, dec. 2007.

[21] W. Xu, C. Chang, Y. Hung, and P. Fung, "Asymptotic properties of order statistics correlation coefficient in the normal cases," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2239–2248, Jun. 2008.

[22] W. Xu, Y. S. Hung, M. Niranjan, and M. Shen, "Asymptotic mean and variance of Gini correlation for bivariate normal samples," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 522–534, Feb. 2010.