MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS

By Qunhua Li[*], James B. Brown, Haiyan Huang and Peter J. Bickel

University of California at Berkeley

Reproducibility is essential to reliable scientific discovery in high-throughput experiments. In this work, we propose a unified approach to measure the reproducibility of findings identified from replicate experiments and identify putative discoveries using reproducibility. Unlike the usual scalar measures of reproducibility, our approach creates a curve, which quantitatively assesses when the findings are no longer consistent across replicates. Our curve is fitted by a copula mixture model, from which we derive a quantitative reproducibility score, which we call the "irreproducible discovery rate" (IDR) analogous to the FDR. This score can be computed at each set of paired replicate ranks and permits the principled setting of thresholds both for assessing reproducibility and combining replicates.

Since our approach permits an arbitrary scale for each replicate, it provides useful descriptive measures in a wide variety of situations to be explored. We study the performance of the algorithm using simulations and give a heuristic analysis of its theoretical properties. We demonstrate the effectiveness of our method in a ChIP-seq experiment.

1. introduction. High-throughput profiling technologies play an indispensable role in modern biology. By studying a large number of candidates in a single experiment and assessing their significance using data analytical tools, high-throughput technologies allow researchers to effectively select potential targets for further studies. Despite their ubiquitous presence in biological research, it is known that any single experimental output from a high-throughput assay is often subject to substantial variability. The importance of quality control, such as reproducibility between replicate samples processed by the same experimental or data analytic procedures (i.e. intra-platform) and consistency between different experimental or data analytic procedures on the same sample (i.e. inter-platform), has long been recognized in microarray experiments (e.g. (MAQC consortium, 2006) among

---

[*]Corresponding author
Keywords and phrases: reproducibility, association, mixture model, copula, iterative algorithm, irreproducible discovery rate, high-throughput experiment, genomics

many others), and recently also addressed in more recent sequencing-based profiling technology, e.g. ChIP-seq technology (Rozowsky et al., 2009; Park, 2009). Unified metrics and standards that can be used to objectively assess the reproducibility and consistency of experimental or data analytic methods are important for producing reliable scientific discoveries and monitoring the performances of data generating procedures.

An important criterion for assessing the reproducibility and consistency in high-throughput experiments is how reproducibly the top ranked signals are reported in intra- or inter-platform replicates. A common approach to assess this reproducibility is to compute the Spearman's pairwise rank correlation coefficient between the significance scores for signals that pass a prespecified significance threshold on each replicate. However, it actually is not entirely suitable for measuring the correlation between two rankings in this type of applications. First, this summary depends on the choice of significance thresholds and may render false assessment that reflects the effect of thresholds rather than the data generating procedure to be evaluated. For instance, with everything else being equal, stringent thresholds generally produce higher rank correlation than relaxed thresholds when applied to the same data. Although standardizing thresholds in principle can remove this confounding effect, calibration of scoring systems across replicate samples or different methods is challenging in practice, especially when the scores or their scales are incomparable on replicate outputs. Though this difficulty seemingly is only associated with heuristics-based scores, indeed, it is also present for probabilistic-based scores, such as p-values. For example, it has been reported in large-scale systematic analyses that strict reliance on p-values in reporting differentially expressed genes causes an apparent lack of inter-platform reproducibility in microarray experiments (MAQC consortium, 2006). Second, rank correlation treats all ranks equally, though the differences in the top ranks seem to be more critical for our purposes. Alternative measures of correlation that give more importance to higher ranks than lower ones, for instance, by weighing the difference of ranks differently, have been developed in more general settings (e.g. Blest (2000); Genest and Plante (2003); Da Costa and Soares (2005)) and applied to this application (Boulesteix and Slawski (2009) for a review). However, all these measures are also subject to prespecified thresholds and raise the question of how to select the optimal weighing scheme.

In this work, we take an alternative approach to measure the reproducibility of replicates. The proposed approach, indeed, is a general method that can be applied to any ranking systems, though we discuss it in the context of high-throughput experiments. Instead of depending on a prespeci-

fied threshold, reproducibility is described as the extent to which the ranks of the signals are no longer consistent across replicates in decreasing significance. We propose a copula-based graphical tool to visualize the loss of consistency and inspect the possible breakdown of consistency empirically, without prior model assumptions. We further quantify reproducibility by classifying the signals into a reproducible and an irreproducible group, using a copula mixture model. By jointly modeling the significance of scores on individual replicates and their consistency between replicates in this model, each signal is assigned a reproducibility score, which we refer to as the *local irreproducible discovery rate*, to infer its reliability. We then define the irreproducibility discovery rate (IDR) and a selection procedure, in a fashion analogous to their counterparts in multiple testing, to rank and select signals by this score. The overall reproducibility of the replicates is described as the amount of irreproducibility in the signals selected at various thresholds using IDR.

This approach not only produces a reproducibility measure that is independent of threshold choices and emphasizes the consistency between top ranks, but also, as we will illustrate, provides the potential for more accurate classification. In addition, because our approach does not make any parametric assumptions on the marginal distributions of scores, it is applicable to any ranking system that produces scores without ties, regardless if it is probabilistic- or heuristic-based. As a quality control measure, it is suited for assessing either reproducibility between replicate samples or consistency between different procedures. It also provides a principled solution for determining selection thresholds in scoring systems that are difficult to calibrate. It is easy to interpret both as a measure of reproducibility and as a criterion for threshold selection.

In the next section, we present the proposed graphical tool (Section 2.1), the copula mixture model and the basis of its estimation procedure (Section 2.2), and the reproducibility criterion (Section 2.3). In section 3, we use simulations to evaluate the accuracy of our model in classifying signals, and compare with some existing methods. In section 4, we apply our method to a data set that motivated this work. The data set was generated by the ENCODE consortium (ENCODE Project Consortium, 2004) from a ChIP-seq assay, a high-throughput technology for studying protein-binding regions on the genome. The primary interest is to assess the reproducibility of several commonly used and publicly available algorithms for identifying the protein-binding regions in ChIP-seq data. Using this data, we compare the reproducibility of these algorithms, infer the reliability of signals identified by each algorithm, and demonstrate how to use our method to identify

4 LI, BROWN, HUANG AND BICKEL

suboptimal results. Section 5 is a general discussion. Finally, in Appendix 2, we present a heuristic justifying our algorithm on optimality grounds.

2. Statistical methods. The data that we consider consist of a large number of putative signals measured on very few replicates of the same underlying stochastic process, for example, protein binding sites identified on the genomes of biological replicates in ChIP-seq experiments. We assume each putative signal has been assigned a score (e.g. p-value or fold ratio) that relates to the strength of the evidence for the signal to be real on the corresponding replicate by some data analysis method. We further assume all the signals are assigned distinct significance scores and the significance scores reasonably represent the relative ranking of signals. However, the distribution and the scale of the scores are unknown and can vary on different replicates. We assume without loss of generality that high scores represent strong evidence of being genuine signals and are ranked high. By convention, we take the "highest" rank to be 1 and so on. We shall use the scores as our data.

We assume n putative signals are measured and reported on each replicate. Then the data consist of n signals on each of the m replicates, with the corresponding vector of scores for signal i being $(x_{i,1}, \ldots, x_{i,m})$. Here $x_{i,j}$ is a scalar score for the signal on replicate j. Our goal is to measure the reproducibility of scores across replicates and select reliable signals by considering information on the replicates jointly. In what follows, we focus on the case of two replicates (i.e. $m = 2$), although the methods in this paper can be extended to the case with more replicates.

We resort to the scientific principle that real signals should be reproducible across replicates. If replicates measure the same underlying stochastic process, then for a reasonable scoring system, the significance scores of genuine signals are expected to be ranked not only higher but also more consistently on the replicates than those of spurious signals. When ranking signals by their significance scores, a (high) positive association is expected between the ranks of scores for genuine signals. A degradation or a breakdown of consistency may be observed when getting into the noise level. This change of concordance provides an internal indicator of the transition from real signal to noise. We will use this in measuring the reproducibility of signals.

2.1. *Displaying the heterogeneity of association.* In this section, we present our graphical method. As we mentioned, the bivariate association between the significance scores is expected to be positive for significant signals, then transits to zero when noise is called. By visualizing how association changes

in the decreasing order of significance, one may localize the transition of association and describe reproducibility in terms of how soon consistency breaks down and how much empirical consistency departs from perfect association before the breakdown.

Rank-based graphs have been studied previously for displaying association between variables, because they are invariant with respect to monotone transformations of the variables separately and are thus scale free. Earlier papers have proposed rank-based graphical tools, such as the Chi-plot (Fisher and Switzer, 1985, 2001) and the Kendall plot (Genest, 2003), for visualizing the presence of association in a sample of continuous bivariate distributions. Related to nonparametric tests of independence, these graphs are based on the null hypothesis of independence and primarily are designed for detecting departures from independence. These plots produce diagrams that have a definitive pattern under independence, e.g. approximately a horizontal (Chi-plot) or a diagonal line (Kendall plot); and the presence of association is shown as a corresponding deviation from the pattern at independence. The type and the level of simple bivariate association may be inferred by comparing the patterns of dependence observed in these plots with the prototypical patterns in Fisher and Switzer (1985, 2001); Genest (2003). When heterogeneity of association is present, such as the one described here, however, these graphs are less informative. See Figure 3 for an illustration on a real data set with mixed populations, considered by Kallenberg and Ledwina (1999); Fisher and Switzer (2001); Genest (2003).

We now present our rank-based graph. Its motivation and properties are different from the plots mentioned.

2.1.1. *Correspondence curves.* Throughout our discussion, we will suppose, for simplicity, that we are dealing with a sample of iid observations from a population. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of scores of n signals on a pair of replicates. Though this is in fact unrealistic in many applications, in particular for the signals from genome-wide profiling (e.g. ChIP-seq experiments), where observations are often dependent, the descriptive and graphical value of our method remains. Moreover, since we are concerned with first order effects, stationarity and mixing assumption will yield the same analysis. Define

(2.1)
$$\Psi_n(t, v) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \geq x_{(\lceil (1-t)n \rceil)}, Y_i \geq y_{(\lceil (1-v)n \rceil)}), \quad 0 < t \leq 1, 0 < v \leq 1,$$

where $x_{(\lceil (1-t)n \rceil)}$ and $y_{(\lceil (1-v)n \rceil)}$ denote the order statistics. $\Psi_n(t, v)$ essentially describes the proportion of the pairs that are ranked both on the upper

$t\%$ of X and on the upper $v\%$ of Y, i.e. the intersection of upper ranked identifications. As consistency usually is deemed a symmetric notion, we will just focus on the special case of $t = v$ and use the shorthand notation $\Psi_n(t)$ in what follows. In fact, $\Psi_n(t, v)$ is an empirical survival copula (Nelson, 1999), and $\Psi_n(t)$ is the diagonal section of $\Psi_n(t, v)$ (Nelson, 1999) (See section 2.2.1 for a brief introduction). Define the population version $\Psi(t) \equiv \lim_n \Psi_n(t)$. Then $\Psi(t)$ and its derivative $\Psi'(t)$, which represent the change of consistency, have the following properties. The same properties are approximately followed in the corresponding sample version Ψ_n and Ψ'_n with finite differences replacing derivatives.

Let $R(X)$ and $R(Y)$ be the ranks of X and Y, respectively.

1. If $R(X) = R(Y)$ on $[t_0, t]$ with $0 \leq t_0 \leq t \leq 1$, $\Psi(t) = \Psi(t_0) + t - t_0$ and $\Psi'(t) = 1$.
2. If $R(X) \perp R(Y)$ on $[0, t]$ with $0 \leq t \leq 1$, $\Psi(t) = t^2$ and $\Psi'(t) = 2t$.
3. If $R(X) = R(Y)$ on $[0, t_0]$ and $R(X) \perp R(Y)$ on $(t_0, 1]$ with $0 \leq t_0 \leq 1$, then for $t_0 \leq t \leq 1$, $\Psi(t) = \frac{t^2 - 2tt_0 + t_0}{1 - t_0}$ and $\Psi'(t) = \frac{2(t - t_0)}{1 - t_0}$.

The last case describes an idealized situation in our applications, where all the genuine signals are ranked higher than any spurious signals, and the ranks on the replicates are perfectly correlated for genuine signals but completely independent for spurious signals.

To visualize the change of consistency with the decrease of significance, a curve can be constructed by plotting the pairs $(t, \Psi_n(t))$ (or $(t, \Psi'_n(t))$) for $0 \leq t \leq 1$. The resulting graphs, which we will refer to as a correspondence curve (or a change of correspondence curve, respectively), depend on X and Y only through their ranks, and are invariant to both location and scale transformation on X and Y. Corresponding to the three special cases described earlier, the curves have the following patterns:

1. When $R(X)$ and $R(Y)$ are perfectly correlated, all points on the curve of Ψ_n will fall on a straight line of slope 1, and all points on the curve of Ψ'_n will fall on a straight line with slope 0.
2. When $R(X)$ and $R(Y)$ are independent, all points on the curve of Ψ_n will fall on a parabola t^2, and all points on the curve of Ψ'_n fall on a straigt line of slope of $2t$.
3. When $R(X)$ and $R(Y)$ are perfectly correlated for the top t_0 and independent for the rest $1 - t_0$, top t_0 points fall into a straight line of slope 1 on the curve of Ψ_n and slope 0 on the curve of Ψ'_n, and the rest $1 - t_0$ points fall into a parabola $\Psi_n(t) = \frac{t^2 - 2tt_0 + t_0}{1 - t_0}$ $(t > t_0)$ on the curve of Ψ_n and a straight line of slope $\frac{2(t - t_0)}{1 - t_0}$ on the curve of Ψ'_n.

a. Ψ_n b. Ψ_n'
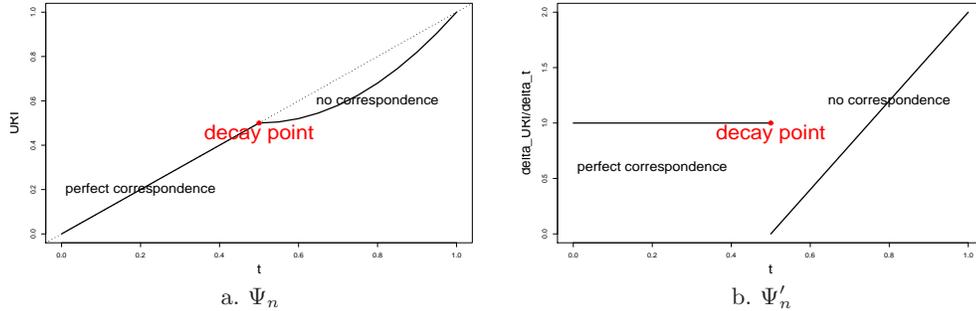
FIG 1. *An illustration of the correspondence profile in an idealized case, where top 50%
are genuine signals and bottom 50% are noise. In this case, all signals are called before
noise; two calling outputs have perfect correspondence for signals and no correspondence
for noise. a. Correspondence curve; b. Change of correspondence curve.*

These properties show that the level of positive association and the possible change of association can be read off these types of curves. For the curve of Ψ_n, strong association translates into a nearly straight line of slope 1, and lack of association shows as departures from the diagonal line, such as curvature bending towards x-axis (i.e. $\Psi_n(t) < t$); if almost no association is present, the curve shows a parabolic shape. Similarly, for the curve of Ψ_n', strong association translates into a nearly straight line of slope 0, and lack of association shows as a line with a positive slope. The transition of the shape of the curves, if present, indicates the breakdown of consistency, which provides guidance on when the signals should not be called any more.

2.1.2. *Illustration of the correspondence curves.* We first demonstrate the curves using an idealized case (Figure 1), where $R(X)$ and $R(Y)$ agree perfectly for the top 50% of observations and are independent for the rest 50% of observations. The curves display the pattern described in case 3 above. The transition of the shape of the curves occurs at 50%, which corresponds to the occurence of the breakdown of consistency. Transition can be seen more visibly on the curve of Ψ_n', as shown by the gap between the disjoint lines with 0 and positive slopes, which represent segments before and after the transition, respectively. This more distinct difference at the transition makes Ψ_n' a better choice for inspecting and localizing the transition than Ψ_n, especially when the transition is less sharp.

To show more realistic cases, we use simulated data to compare and contrast the curves in presence and absence of the aforementioned transition (Figure 2). The case where no transition occurs is illustrated using two

single-component bivariate Gaussian distributions with homogeneous association, $\rho = 0$ (Figure 2a) and $\rho = 0.8$ (Figure 2b), respectively. The presence of the transition is illustrated using two two-component bivariate Gaussian mixtures, whose lower ranked component has independent coordinates (i.e. $\rho_0 = 0$) and the higher ranked component has positively correlated coordinates with $\rho_1 = 1$ (Figure 2c) and $\rho_1 = 0.8$ (Figure 2d), respectively.

As in the idealized example (Figure 1), the characteristic transition of curves is observed when the transition of association is present (Figure 2c, d), but not seen when the data consists of only one component with homogeneous association. This shows that the transition of the shape of the curve may be used as an indicator for the presence of the transition of association.

We now compare the Ψ_n' plot with the Chi-plot and the K-plot using a real example considered by (Kallenberg and Ledwina, 1999; Fisher and Switzer, 2001; Genest, 2003). This dataset consists of 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catachable-sized fish in the Skeena River. It was speculated by (Fisher and Switzer, 2001) to contain a mixed populations with heterogeneous association. Though the dissimilarity of Chi-plot or K-plot to their prototypical plots (c.f. (Fisher and Switzer, 2001; Genest, 2003)) suggests the data may involve more than simple monotone association (Fisher and Switzer, 2001; Genest, 2003), neither of these plots manifest heterogeneity of association. In the Ψ_n' curve (Figure 3(d)), the characteristic pattern of transition is observed at about $t = 0.5$, which indicates that the data is likely to consist of two groups, with roughly the top 50% from a strongly associated group and the bottom 50% from a weakly associated group. It agrees with the speculation in (Fisher and Switzer, 2001).

2.2. *Inferring the reproducibility of signals.* In this section, we present a statistical model that quantifies the dependence structure and infers the reliability of signals by using information on both the significance of the scores and the consistency between replicates.

In general, genuine signals tend to be more reproducible and scored higher than spurious ones. The scores on replicates may be viewed as a mixture of two groups, which differ in both the strength of association and the level of significance. Recall that in these applications, the distributions and the scales of scores are usually unknown and may vary across data sets. To model such data, a semiparametric copula model is appropriate, in which the associations among the variables are represented by a simple parametric model but the marginal distributions are estimated nonparametrically using their ranks. Though using ranks, instead of the raw values of scores, generally causes some loss of information, the rank transformation is commonly used in
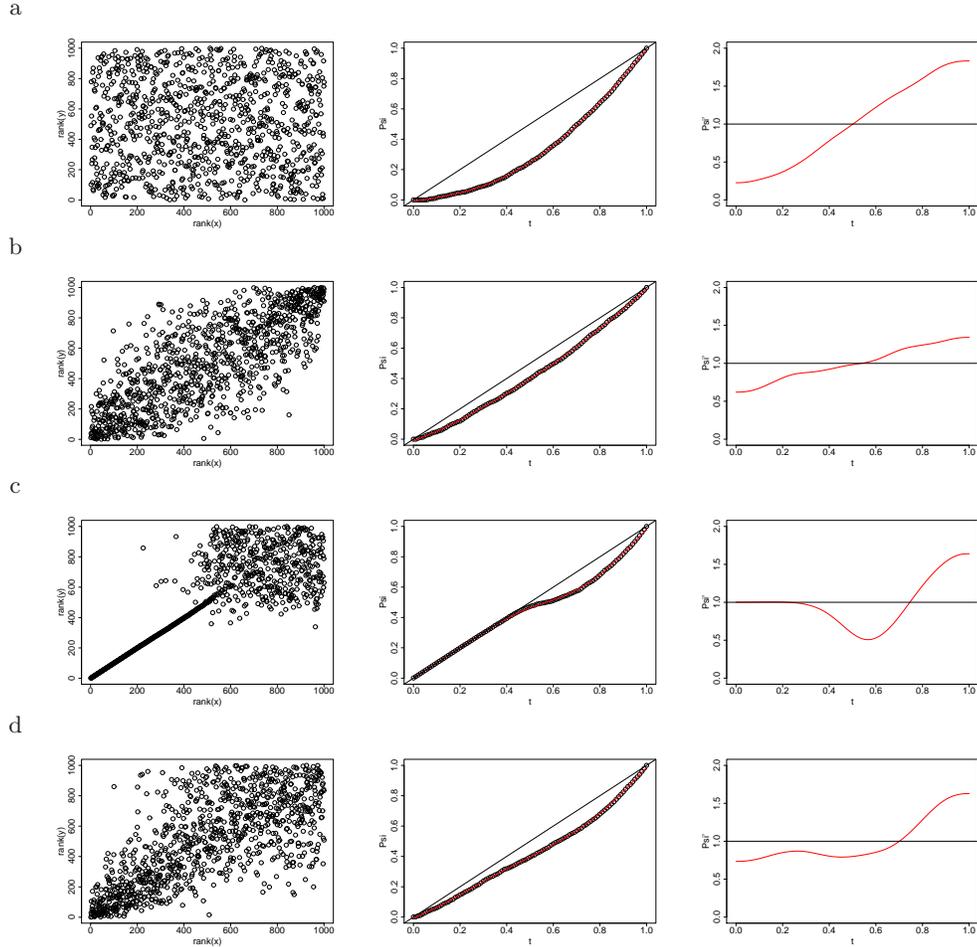
FIG 2. *Behavior of correspondence curves when data consists of homogeneous and heterogeneous association. From left to right, the three columns are the plot of ranks, the curve of Ψ and the curve of Ψ'. a. Bivariate Gaussian distribution with $\rho = 0$; b. Bivariate Gaussian distribution with $\rho = 0.8$; c. A mixture of two bivariate Gaussian distributions with marginals on both coordinates as $f_0 = N(0,1)$ and $f_1 = N(3,1)$, $\rho_0 = 0$ and $\rho_1 = 1$ and mixing proportion $\pi_1 = 0.5$; d. A mixture of two bivariate Gaussian distributions with marginals on both coordinates as $f_0 = N(0,1)$ and $f_1 = N(2,1)$, $\rho_0 = 0$ and $\rho_1 = 0.8$ and mixing proportion $\pi_1 = 0.5$. The curve of Ψ'_n is produced by taking derivative on the spline that fits Ψ_n with df=6.4.*
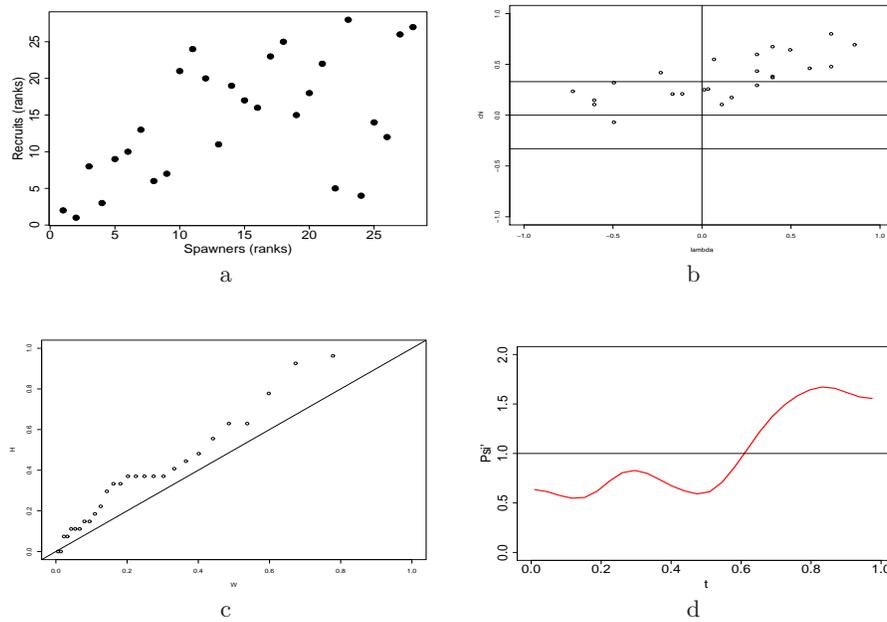
FIG 3. *Rank scatterplot (a), chi-plot (b), k-plot (c) and the change of correspondence curve (d) for 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-sized fish in the Skeena River. The curve of Ψ_n' is produced by taking derivative on the spline that fit Ψ_n with df=6.4.*

genomic data because scales are unknown and incomparable across replicates from many measurements. In view of the heterogeneous association in the genuine and spurious signals, we further model the heterogeneity of the dependence structure in the copula model using a mixture model framework.

Before proceeding to our model, we first provide a brief review of copula models, and refer to (Joe, 1997) and (Nelson, 1999) for a modern treatment of copula theory.

2.2.1. *Copulas.* The multivariate function $C = C(u_1, \ldots, u_p)$ is called a copula if it is a continuous distribution function and each marginal is a uniform distribution function on $[0, 1]$. That is, $C : [0, 1]^p \to [0, 1]$, with $C(u) = P(U_1 \leq u_1, ..., U_p \leq u_p)$, in which each $U_j \sim Unif[0, 1]$ and $u = (u_1, ..., u_p)$. By Sklar's theorem (Sklar, 1959), every continuous multivariate probability distribution can be represented by its univariate marginal distributions and a copula, described using a bivariate case as follows.

Let X_1 and X_2 be two random variables with continuous CDFs F_1 and F_2. The copula C of X_1 and X_2 can be found by making the marginal probability integral transforms on X_1 and X_2 so that

$$(2.2) \qquad C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)), \quad u_1, u_2 \in [0, 1]$$

where F is the joint distribution function of (X_1, X_2), F_1 and F_2 are the marginal distribution functions of X_1 and X_2, respectively, and F_1^{-1} and F_2^{-1} are the right-continuous inverses of F_1 and F_2, defined as $F_j^{-1}(u) = \inf\{z : F_j(z) \geq u\}$. That is, the copula is the joint distribution of $F_1(X_1), F_2(X_2)$. These variables are unobservable but estimable by the normalized ranks $F_{n1}(X_1), F_{n2}(X_2)$ where F_{n1}, F_{n2} are the empirical distribution functions of the sample. The function $\delta_C(t, t) = C(t, t)$ is usually referred to as the diagonal section of a copula C. We will use the survival function of the copula C, $\bar{C}(u_1, u_2) = P(U_1 > 1 - u_1, U_2 > 1 - u_2)$, which describes the relationship between the joint survival function $(\bar{F}(x_1, x_2) = P(X_1 > x_1, X_2 > x_2))$ and its univariate margins $(\bar{F}_j = 1 - F_j)$ in a manner completely analogous to the relationship between univariate and joint functions, as $\bar{C}(u_1, u_2) = \bar{F}(\bar{F}_1^{-1}(u_1), \bar{F}_2^{-1}(u_2))$. The sample version of (2.2) is called an *empirical copula* (Deheuvels, 1979; Nelson, 1999), defined as

$$(2.3) \qquad C_n(\frac{i}{n}, \frac{j}{n}) = \frac{1}{n} \sum_{k=1}^{n} 1(x_{k,1} \leq x_{(i),1}, x_{k,2} \leq x_{(j),2}), \quad 1 \leq i, j \leq n$$

for a sample of size n, where $x_{(i),1}$ and $x_{(j),2}$ denotes order statistics on each coordinate from the sample. The sample version of survival copulas follows similarly.

This representation provides a way to parametrize the dependence structure between random variables separately from the marginal distributions, for example, a parametric model for the joint distribution of u_1 and u_2 and a nonparametric model for marginals. Copula-based models are natural in situations where learning about the association between the variables is important, but the marginal distributions are assumably unknown. For example, the 2-dimensional Gaussian copula C is defined as

$$(2.4) \qquad C(u_1, u_2 | \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2) | \rho)$$

where Φ is the standard normal cumulative distribution function, $\Phi_2(\cdot, \cdot | \rho)$ is the cumulative distribution function for a bivariate normal vector $(z_1, z_2) \sim N(\binom{0}{0}, \binom{1 \ \rho}{\rho \ 1})$, and ρ is the correlation coefficient. Modeling dependence with arbitrary marginals F_1 and F_2 using the Gaussian copula (2.4) amounts to assuming data is generated from latent variables (z_1, z_2) by setting $x_1 = F_1^{-1}(\Phi(z_1))$ and $x_2 = F_2^{-1}(\Phi(z_2))$. Note that if F_1 and F_2 are not continuous, u_1 and u_2 are not uniform. For convenience, we assume that F_1 and F_2 are continuous throughout the text.

2.2.2. *A copula mixture model.* We now present our model for quantifying the dependence structure and inferring the reproducibility of signals. We assume throughout this part that our data is a sample of independent identically distributed bivariate vectors $(x_{i,1}, x_{i,2})$.

We assume the data consists of genuine signals and spurious signals, which in general correspond to a more reproducible group and a less reproducible group. We use the indicator K_i to represent whether a signal i is genuine ($K_i = 1$) or spurious ($K_i = 0$). Let π_1 and $\pi_0 = 1 - \pi_1$ denote the proportion of genuine and spurious signals, respectively. Given $K = 1$, we assume the pairs of scores for genuine (respectively, spurious) signals are independent draws from a continuous bivariate distribution with density $f_1(\cdot, \cdot)$ (respectively, $f_0(\cdot, \cdot)$, given $K = 0$) with joint distribution $F_1(\cdot, \cdot)$ (respectively, $F_0(\cdot, \cdot)$). Note, however, that even if the marginal scales are known, K_i would be unobservable so that the copula is generated by the marginal mixture (with respect to K), $F_j = \pi_0 F_j^0 + \pi_1 F_j^1$, where F_j is the marginal distribution of the jth coordinate and F_j^k is the marginal distribution of the corresponding k^{th} component.

Because genuine signals are more reproducible than spurious signals, we expect the two groups to have different dependence structures between replicates. We assume that, given the indicator K_i, the dependence between replicates for genuine (respectively, spurious) signals is induced by a bivariate Gaussian distribution $\mathbf{z}_1 = (z_{1,1}, z_{1,2})$ (or respectively, $\mathbf{z}_0 = (z_{0,1}, z_{0,2})$).

The choice of Gaussian distribution for inducing the dependence structure in each component is made based on the observation that the dependence within a component in the data we consider generally is symmetric and that an association parameter with a simple interpretation, such as the correlation coefficient for a Gaussian distribution, is natural.

Since spurious signals are presumably less reproducible, we assume corresponding signals on the replicates to be independent, i.e. $\rho_0 = 0$; whereas, since we expect genuine signals be positively associated between replicates, we assume $\rho_1 > 0$. We expect that a pair of scores from $F_1(\cdot, \cdot)$ should tend to be larger than under $F_0(\cdot, \cdot)$. It also seems natural to assume that the underlying latent variables, reflecting biological replicates, have the same marginal distributions. Finally we note that if the marginal scales are unknown we can only identify the difference in means of the two latent variables and the ratio of their variances. Thus, the parametric model generating our copula can be described as follows:

Let $K_i \sim Bernoulli(\pi_1)$ and $(z_{i,1}, z_{i,2})$ be distributed as

$$(2.5\text{a}) \qquad \begin{pmatrix} z_{i,1} \\ z_{i,2} \end{pmatrix} \mid K_i = k \sim N\left(\begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & \rho_k \sigma_k^2 \\ \rho_k \sigma_k^2 & \sigma_k^2 \end{pmatrix} \right), \quad k = 0, 1$$

where $\mu_0 = 0$, $\mu_1 > 0$, $\rho_0 = 0$, $0 < \rho_1 \leq 1$.
Let

$$(2.5\text{b}) \qquad u_{i,1} \equiv G(z_{i,1}) = \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,1} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,1})$$

$$u_{i,2} \equiv G(z_{i,2}) = \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,2} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,2})$$

Our actual observations are

$$(2.5\text{c}) \qquad\qquad x_{i,1} = F_1^{-1}(u_{i,1})$$

$$x_{i,2} = F_2^{-1}(u_{i,2})$$

where F_1 and F_2 are the marginal distributions of the two coordinates, which are assumed continuous but otherwise unknown.

Thus, our model, which we shall call a copula mixture model, is a semiparametric model parametrized by $\theta = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$ and (F_1, F_2). The

corresponding mixture likelihood for the data is

(2.6a)
$$L(\theta) = \prod_{i=1}^{n} [\pi_0 h_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2}))) + \pi_1 h_1(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))]$$

(2.6b)
$$= \prod_{i=1}^{n} [c(F_1(x_{i,1}), F_2(x_{i,2})) g(G^{-1}(F_1(x_{i,1}))) g(G^{-1}(F_2(x_{i,2})))]$$

where

$$(2.7) \qquad c(u_1, u_2) = \frac{\pi_0 h_0(G^{-1}(u_1), G^{-1}(u_2)) + \pi_1 h_1(G^{-1}(u_1), G^{-1}(u_2))}{g(G^{-1}(u_1)) g(G^{-1}(u_2))}$$

is a copula density function with $h_0 \sim N\left(\binom{0}{0}, \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)\right)$ and $h_1 \sim N\left(\binom{\mu_1}{\mu_1}, \left(\begin{smallmatrix} \sigma_1^2 & \rho_1 \sigma_1^2 \\ \rho_1 \sigma_1^2 & \sigma_1^2 \end{smallmatrix}\right)\right)$ G is defined in (2.5b) and g is the density function of G. Note that G depends on θ.

Given the parameters θ, the posterior probability that a signal i is in the irreproducible group can be computed as
(2.8)
$$Pr(K_i = 0 \mid (x_{i1}, x_{i2}); \theta) = \frac{\pi_1 h_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}{\sum_{k=0,1} \pi_k h_k(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}$$

We estimate values for these classification probabilities by estimating the parameters θ using an estimation procedure described in section 2.2.3, and substituting these estimates into the above formulas.

The idea of using a mixture of copulas to describe complex dependence structures is not entirely new. For example, the mixed copula model (Hu, 2006) in economics uses a mixture of copulas $(C_{mix}(u_1, u_2 \mid (\theta_1, \ldots, \theta_k)) = \sum_{i=1}^{k} C(u_1, u_2 \mid \theta_i))$ to generate flexible fits to the dependence structures that do not follow any standard copula families. All the copulas in C_{mix} are induced from latent variables with the *same* marginal distribution. In contrast, the copula in our model is induced from latent variables with *different* marginal distributions; and our modeling goal is to cluster the observations into groups with homogeneous associations. A nonstandard estimation procedure turns out to be convenient and we expect to be efficient, as we shall see in section 2.2.3.

2.2.3. *Estimation of the copula mixture model.* In this section, we describe an estimation procedure that estimates the parameters θ in (2.6) and the membership K_i of each observation.

A common strategy to estimate the association parameters in semiparametric copula model is a "pseudo-likelihood" approach, which is described in broad, nontechnical terms by (Oakes, 1994). In this approach, the empirical marginal distribution functions $\hat{F}_j$, after rescaling by multiplying by $(\frac{n}{n+1})$ to avoid infinities, are plugged into the copula density in (2.6b), ignoring the terms involving g. The association parameters are then estimated by maximizing the pseudo copula likelihood. Genest, Ghoudi and Rivest (Genest et al., 1995) showed, without specifying the algorithms to compute them, that under certain technical conditions, the estimators obtained from this approach are consistent, asymptotically normal and fully efficient only if the coordinates of the copula are independent.

We adopt a different approach which, in principle, leads to efficient estimators under any choice of parameters and F_1, F_2. Note that the estimation of the association parameter ρ_1 depends on the estimation of μ_1, σ_1^2 and π_1 due to the presence of the mixture structure, which make the log-likelihood (2.6) difficult to maximize directly. Our approach is to estimate the parameters $\hat{\theta}$ by maximizing the log-likelihood (2.6) of pseudo-data $G^{-1}(\frac{n}{n+1}\hat{F}_{i,j}; \theta)$, where $\hat{F}_{i,j} \equiv \hat{F}_j(x_{i,j})$.

As the latent variables $z_{0,j}$ and $z_{1,j}$ in our model form a mixture distribution, it is natural to use an expectation-maximization (EM) algorithm (Dempster et al., 1977) to estimate the parameters $\hat{\theta}$ and infer the status of each putative signal for pseudo-data. In our approach, we first compute the pseudo-data $G^{-1}(\frac{n}{n+1}\hat{F}_{i,j}; \theta_0)$ from some initialization parameters $\theta^{(0)}$, then iterate between two stages: (1) maximizing θ based on the pseudo-data using EM and (2) updating the pseudo data. The detailed procedure is given in Appendix 1. The EM stage may be trapped in local maxima, and the stage of updating pseudo data may not converge from all starting points. However, in the simulations we performed (Section 3), it behaves well and finds the global maxima, when started from a number of initial points.

We sketch in Appendix 2, a heuristic argument that a limit point of our algorithm close to the true value satisfies an equation whose solution is asymptotically efficient. Although our algorithm converges in practice, we have yet to show its convergence in theory. However, a modification which we are investigating does converge to the fixed point mentioned above. This work will appear elsewhere.

2.3. *Irreproducible identification rate.*   In this section, we derive a reproducibility criterion from the copula mixture model in section 2.2.2 based on an analogy between our method and the multiple hypothesis testing problem. The reproducibility criterion can be used to assess the individual

reliable level and measure the overall reproducibility of the replicate outputs.

In the multiple hypothesis testing literature, the false discovery rate (FDR) and its variants, including positive false discovery rate (pFDR) and marginal false discovery rate (mFDR), are introduced to control the number of false positives in the rejected hypotheses (Benjamini and Hochberg, 1995; Storey, 2002; Genovese and Wasserman, 2002). In the FDR context, when hypotheses are independent and identical, the test statistics can be viewed as following a mixture distribution of two classes, corresponding to whether or not the statistic is generated according to the null hypothesis (e.g. (Efron, 2004b; Storey, 2002)). Based on this mixture model, the local false discovery rate, which is the posterior probability of being in the null component $Lfdr(\cdot) = (1-\pi)f_0(\cdot)/f(\cdot)$, was introduced to compute the individual significance level (Efron, 2004b). Sun and Cai (2007) show, again for the iid case, that Lfdr is also an optimal statistic in the sense that the thresholding rule based on Lfdr controls the marginal false discovery rate with the minimum marginal false nondiscovery rate.

As in multiple hypothesis testing, we also build our approach on a mixture model and classify the observations into two classes, though the two classes have different interpretation and representation. In our model, the two classes represent irreproducible measurements and reproducible measurements, in contrast to nulls and nonnulls in the multiple testing context, respectively.

In analogy to the local false discovery rate, we define a quantity, which we call the *local irreproducible discovery rate*, to be

$$(2.9) \qquad idr(x_{i,1}, x_{i,2}) = \frac{\pi_0 h_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}{\sum_{k=0,1} \pi_k h_k(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}.$$

This quantity can be thought as the *a posteriori* probability that a signal is not reproducible on a pair of replicates (i.e. (2.8)), and can be estimated from the copula mixture model.

Similarly, we define the *irreproducible discovery rate* (IDR) in analogy to the mFDR,

(2.10)

$$IDR(\gamma) = P(irreproducible \mid i \in I_\gamma) = \frac{\pi_0 \int_{I_\gamma} dH_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}{\int_{I_\gamma} dH(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}$$

where $I_\gamma = \{(x_{i,1}, x_{i,2}) : idr(x_{i,1}, x_{i,2}) < \gamma\}$, H_0 and H are the CDF of density functions h_0 and $h = \pi_0 h_0 + \pi_1 h_1$, respectively. For a desired control level α, if $(x_{(i),1}, x_{(i),2})$ are the pairs ranked by idr values, define $l = \max\{i : \frac{1}{i} \sum_{j=1}^{i} idr_j \le \alpha\}$. By selecting all $(x_{(i),1}, x_{(i),2})$ $(i = 1, \ldots, l)$, we can think

of this procedure as giving an expected rate of irreproducible discoveries no greater than α. It is analogous to the adaptive step-up procedure of Sun and Cai (2007) for the multiple testing case.

This procedure essentially amounts to re-ranking the identifications according to the likelihood ratio of the joint distribution of the two replicates. The resulting rankings are generally different from the ranking of the original significance scores on either replicate.

Unlike the multiple testing procedure, our procedure does not require $x_{i,j}$ to be p-values; instead, $x_{i,j}$ can be any scores with continuous marginal distributions. When p-values are used as scores, our method can also be viewed as a method to combine p-values. We compare our method and two commonly-used p-value combinations through simulations in Section 3.

3. Simulation studies. We first use simulation studies to examine the performance of our approach. In particular, we aim to assess the accuracy of our classification, to evaluate the benefit of combining information between replicates over using only information on one replicate, and to compare with two existing methods for combining information across replicates. Our simulations are generated from a model commonly used for modeling high-throughput data (e.g. (Lee et al., 2000; Efron, 2004a)). As the model in fact is also a reparameterization of our model (2.5), the comparison with other combination methods is not to provide evidence that our approach is actually superior in practice, but rather, to provide insight into the kind of gains in performance that might be achievable in practice. In addition, they provide a helpful check on the convergence of our estimation procedure.

In the simulation studies, a sample of n pairs of signals is generated on two replicates. Each pair of observed signal $(Z_{i1}, Z_{i,2})$ $(i = 1, \ldots, n)$ is a noisy realization of a latent signal Z_i, which is independently and identically generated from the following normal mixture model:

$$(3.1) \qquad\qquad K_i \sim Bernoulli(\pi_1)$$
$$Z_i \mid K_i = k \sim N(\mu_k, \tau_k^2), \quad k = 0, 1$$
$$Z_{ij} \mid K_i = k, Z_i = Z_i + \epsilon_{ijk}, \quad j = 1, 2$$
$$\epsilon_{ijk} \sim N(0, \omega_k^2)$$

where $\mu_0 = 0$ and $\mu_1 > 0$. As can easily be seen from the joint distribution of $(Z_{i,1}, Z_{i,2}) \mid K_i$,

$$\begin{pmatrix} Z_{i,1} \\ Z_{i,2} \end{pmatrix} \mid K_i = k \sim N \left(\begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \tau_k^2 + \omega_k^2 & \rho_k(\tau_k^2 + \omega_k^2) \\ \rho_k(\tau_k^2 + \omega_k^2) & \tau_k^2 + \omega_k^2 \end{pmatrix} \right), \quad k = 0, 1$$

where $\rho_0 = 0$ and $\rho_1 = \frac{\tau_1^2}{\tau_1^2+\omega_1^2}$, $\tau_k^2 + \omega_k^2 = \sigma_k^2$ in (2.5a). Thus by setting $\tau_k^2 + \omega_k^2 = 1$, $(Z_{i,1}, Z_{i,2})$ directly corresponds to the latent Gaussian variables in (2.5a). We use the significance score in this setting to be the p-value from a one-sided z-test for $H_0 : \mu = 0$ vs $H_1 : \mu > 0$. Using p-values as the scores $X_{i,j}$ is equivalent to letting $F_j = 1 - \Phi(G^{-1}(\cdot))$ in (2.5c).

With this choice, a multiple hypothesis testing procedure for independent hypotheses is a natural approach for selecting signals on a single replicate. In the multiple hypothesis testing literature, selecting signals can be done either by ranking individual p-values and choosing a cutoff along the rankings, such as (Benjamini and Hochberg, 1995, 2000; Genovese and Wasserman, 2004; Storey, 2003), or by converting p-values to z-values and thresholding based on the likelihood ratio of z-values, such as (Sun and Cai, 2007). In the given setting of our simulations (i.e. one-sided test with $\sigma_0^2 = \sigma_1^2$), both types of methods rank the signals in the same order. For operational simplicity, we select signals on individual replicates by thresholding p-values, though our approach is conceptually closer to the z-value thresholding method.

With the p-value as the significance score, our method can also be viewed as a way to combine p-values for ranking signals by their consensus. The two most commonly-used methods for combining p-values of a set of independent tests are Fisher's combined test (Fisher, 1932) and Stouffer's z method (Stouffer et al., 1949). In Fisher's combination for the given one-sided test, the test statistic $Q_i = -2 \sum_{j=1}^m \log(p_{i,j})$ for each pair of signal has the χ_{2m}^2 distribution under H_0, where $p_{i,j}$ is the p-value for the ith signal on the jth replicate, m is the number of studies and $m = 2$ here. In Stouffer's method, the test statistic $S_i = \frac{1}{\sqrt{m}} \sum_{j=1}^m \Phi^{-1}(1 - p_{i,j})$ has $N(0,1)$ under H_0, where Φ is the standard normal CDF.

In our setting, the classification results from our method and from the three hypothesis testing methods are directly comparable: the reproducible signals correspond to genuine signals and the irreproducible signals correspond to spurious signals. For a given threshold, we classify a call as correct (or incorrect), when a genuine (or spurious) signal is assigned an idr value smaller than an idr threshold for our method. Correspondingly, for a call from individual replicates, Fisher's method or Stouffer's method, the same classification applies, when its p-value is smaller than the threshold, or a value of the test statistic in Fisher's method or Stouffer's method exceeds their respective thresholds. We compare the discriminative power of these methods by assessing the tradeoff between the number of correct and incorrect calls made at various thresholds.

We simulated data from three sets of parameters with different levels of association between replicates for genuine signals (Table 1). For each

TABLE 1

Simulation parameters and parameter estimation in the simulation studies of 100 datasets. Each data set consists of 10000 pairs of observations. The simulation parameters are estimated from a ChIP-seq dataset. In all simulations, $\mu_0 = 0$, $\sigma_0^2 = 1$ and $\rho_0 = 0$. The table shows the mean and the standard deviation of the estimated parameters over the 100 data sets.

		π_1	ρ_1	μ_1	σ_1^2
S1	True parameter	0.650	0.840	2.500	1.000
	Estimated values	0.648 (0.005)	0.839 (0.005)	2.524 (0.033)	1.003 (0.024)
S2	True parameter	0.300	0.400	2.500	1.000
	Estimated values	0.302 (0.004)	0.398 (0.024)	2.549 (0.037)	1.048 (0.032)
S3	True parameter	0.050	0.840	2.500	1.000
	Estimated values	0.047 (0.004)	0.824 (0.026)	2.536 (0.110)	0.876 (0.087)

parameter set, we simulated 100 datasets, each of which consists of two replicates with 10000 signals on each replicate. In an attempt to generate realistic simulations, we first estimated parameters from a ChIP-seq dataset (described in section 4) using the model in section 2.2, then simulated the signals on a pair of replicates from similar parameters using the sampling model (3.1). To illustrate the behavior of our method when genuine signals have low correlation, we include a simulation with $\rho = 0.4$. We also consider a simulation with $\pi = 0.05$ and $\rho = 0.90$ to illustrate the case when only a small proportion of real but highly correlated signals are present. In each simulation, we ran the estimation procedure from 10 random initializations, and stopped the procedure when the increment of log-likelihood is < 0.01 in an iteration or the number of iterations exceeds 100. The results that converges to the highest likelihood are reported.

3.1. *Parameter Estimation and calibration of IDR.* As shown in Table 1, the parameters $\theta \equiv (\pi_1, \rho_1, \mu_1, \sigma_1^2)$ are estimated with reasonable accuracy in most of the cases studied here. The only exception is that σ_1 was underestimated when the proportion of true signals is small, $\pi_1 = 0.05$, a case hard to distinguish from that of a single component. All the simulations converge, when starting points are close to the true parameters.

The irreproducible discovery rate as a guide for the selection of the signals needs to be well-calibrated. To check the calibration of our method, we compare the estimated IDR with the empirical FDR (Figure 4 Left column). As shown in Figure 4, our method is reasonably well calibrated for all the situations studied. When correlation between genuine signals is weak ($\rho = 0.4$), estimated IDR slightly underestimates FDR. However, the correlation usually are stronger than this case in practice for reasonable replicates.

3.2. *Comparison of discriminative power.* To assess the benefit of combining information on replicates and compare with existing methods of combining p-values, we compared our method with the p-value thresholding method, Fisher's method and Stouffer's method, by assessing the tradeoff between the numbers of correct and incorrect calls made at various thresholds. As a small number of false calls is desired in practice, the comparison focuses on the performance in this region.

As shown in Figure 4, our method and the two combination methods consistently identify substantially more true signals than using only information on one replicate in all the simulations we experimented. Though our method only uses the rank of p-values, it consistently outperforms the two p-value combination methods, when genuine signals in replicates are moderately or highly correlated. When replicates are weakly correlated (Figure 4b), our method still outperforms Fisher's method and performs similarly as Stouffer's method. This gain illustrates that combining information on replicates using our method improves accuracy of identification. The superiority over the other two combinations is not surprising since the simulations are generated from our own model. However, the simulations provide insight into the kind of gains in performance that might be achievable in practice.

4. Applications on real data.

4.1. *Comparing the reproducibility of multiple peak callers for ChIP-seq experiments.* We now consider an application arising from a collaborative project with the ENCODE consortium (ENCODE Project Consortium, 2004). This project has three primary goals: comparing the performance of multiple algorithms for identifying protein-binding regions in ChIP-seq data (described below), selecting reliable binding regions using a uniform criterion for data from different sources, and identifying inconsistent experiments. A detailed analyses of a large ENCODE data set will appear elsewhere. Here we only use one subset of the data to illustrate how our method is used for assessing and comparing the reproducibility of algorithms between biological replicates, setting a uniform criterion for selecting binding regions, and identifying suboptimal results.

We now state the background of the data in more detail and refer to (Park, 2009) for a recent review. A ChIP-seq experiment is a high-throughput assay to study the protein binding sites on DNA sequences in the genome. In a typical ChIP-seq experiment, the DNA regions that are specifically bound by the protein of interest are first enriched by immunoprecipitation, and then the enriched DNA regions are sequenced by high-throughput sequencing, which generates a genome-wide scan of tag counts that correspond to

a-1. calibration for $\pi = 0.65, \rho = 0.84$

a-2. tradeoff for $\pi = 0.65, \rho = 0.84$

b-1. calibration for $\pi = 0.30, \rho = 0.4$

b-2. tradeoff for $\pi = 0.30, \rho = 0.4$

c-1. calibration for $\pi = 0.05, \rho = 0.9$

c-2. tradeoff for $\pi = 0.05, \rho = 0.9$
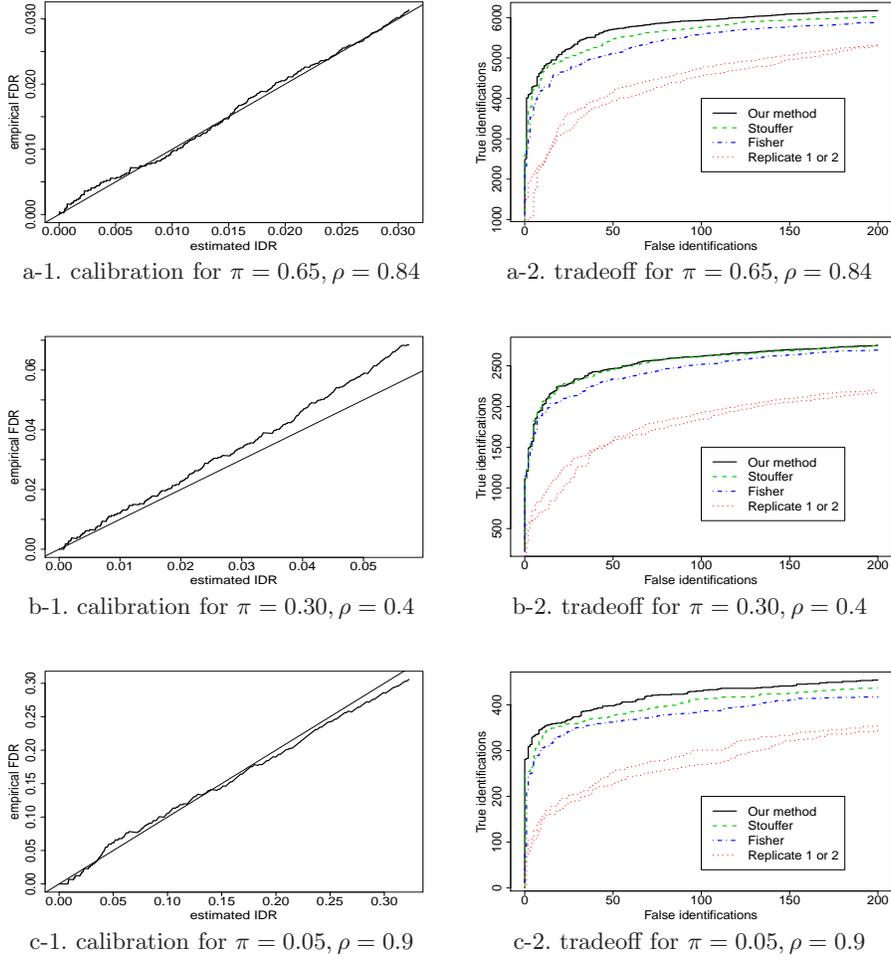
FIG 4. *The number of correct and incorrect calls made at various thresholds in simulation studies. Incorrect calls: the number of spurious signals assigned idr value smaller than the thresholds (our method) or with p-values smaller than the cutoffs (individual replicates) or with Q (Fisher's method) or S (Stouffer's method) values exceeding thresholds. Correct calls: the number of genuine signals assigned idr value smaller than the thresholds (our method) or with p-values smaller than the cutoffs (individual replicates) or with Q (Fisher's method) or S (Stouffer's method) values exceeding thresholds.*

the level of enrichment at each region. The relative significance of the regions are determined by a computational algorithm (usually referred to as a peak caller) largely according to the local tag counts, based on either heuristics or some probabilistic models. The regions whose significance are above some prespecified threshold then are identified for downstream analyses. To date, more than a dozen of peak callers have been published. Some common measures of significance are fold of enrichment, p-value or q-value (Storey, 2003).

As no genome-wide ground truth is available for comparing the identified regions, direct assessment of the accuracy of the peak callers is not possible. We assess the reproducibility between biological replicates for each peak caller using the proposed approach on the CTCF ChIP-seq data described below.

4.1.1. *Description of the data.* In this comparison, the ChIP-seq experiments of a transcription factor CTCF from two biological replicates were generated from the Berstein Laboratory at the Broad Institute on human K526 cells. Peaks were identified using nine commonly used and publicly available peak callers, namely, Peakseq (Rozowsky et al., 2009), MACS (Zhang et al., 2008), SPP (Kharchenko et al., 2008), Fseq (Boyle et al., 2008), Hotspot (Thurman et al., In preparation), Erange (Mortazavi et al., 2008), Cisgenome (Ji et al., 2008), Quest (Valouev et al., 2008), and SISSRS (Jothi et al., 2008), using their default significance measures and default parameter settings with either default thresholds or more relaxed thresholds. Among them, Peakseq and SPP use q-value, MACS, HotSpot and SISSRS use p-value, and the rest use fold of enrichment, as their significance measures. The peaks generated from different algorithms have substantially different peak widths. SPP and SISSRS generate peaks with fixed width of 100bp and 40bp, respectively; all other algorithms generate peaks with varying peak width (median=130bp-760bp).

To standardize the comparison, the peak width were normalized by truncating the peaks wider than 40bp down to intervals of 40bp centered at the reported summits of peaks. Prior to applying our method, peaks on different replicates are paired up if their coverage regions overlap $\geq$ 1bp. Only paired peaks are kept for assessing reproducibility.

4.2. *Results.*

4.2.1. *Correspondence profiles.* Figure 5 shows the correspondence profiles for the nine peak callers. By referring to the prototypical plots in Figure

2, five peak callers (Peakseq, MACS, SPP, Fseq and Hotspot) show the characteristic transition from strong association to near independence (Figure 5b). As described in Section 2.1, a high reproducibility translates to late occurence of the transition to a segment with a positive slope. According to how much down the rank list the transition is observed, the three peak callers that show the highest reproducibility on this dataset are Peakseq, MACS and SPP (Figure 5). For the other four peak callers (Erange, Cisgenome, Quest and SISSRS), the presence of components with distinct association seems to be less clear and low consistency is shown. It later became known that the default thresholds were overly stringent for three of these peak callers (Cisgenome, Quest and SISSRS) on this data, which caused too few peaks to be reported and prevented us from seeing the possibility of two components.

4.2.2. *Inference from the copula mixture model.* We applied the copula mixture model to the peaks identified on the replicates for each peak caller. As data may consist of only one group with homogeneous association, we also estimated the fit using a one-component model that corresponds to setting $\pi_1 = 1, \mu_1 = 0$ and $\sigma_1^2 = 1$ in (2.5). We then tested for the smallest number of components compatible with the data, using a likelihood ratio test statistic ($\lambda = \frac{L_2}{L_1}$), where L_2 and L_1 are the likelihood of two-component and single-component models, respectively. With mixture models, it is well known that the regularity conditions do not hold for $2\log(\lambda)$ to have its usual asymptotic Chi-square null distribution. We therefore used a parametric bootstrap procedure to obtain appropriate p-values (McLachlan, 1987). In our procedure, 100 bootstrap samples were sampled from the null distribution under the one component hypothesis using the parametric bootstrap, where the parameter estimate was obtained by maximizing the pseudo-likelihood of the data under the null hypothesis of the one-component model. Then p-values were obtained by referring to the distribution of the likelihood ratio computed from the bootstrap samples. Table 2 summarizes the parameter estimation from both models and the bootstrap results.

Based on the likelihood ratio test, it seems that the one-component model fits the results from SISSRS, Quest and Cisgenome better, and the two-component model fits the results from other peak callers. This is consistent with the correspondence profile (Figure 5b).

We compute the irreproducible discovery rate (IDR) for peaks selected at various local idr cutoffs for all the peak callers using (2.10) and illustrate it in Figure 6. For a given IDR level, one can call peaks by the values of local idr and determine the number of peaks to be called from this plot, as

Parameters estimated from the copula mixture model and the single-component model, and model selection for determining the number of components. $(\pi_1, \rho_1, \mu_1, \sigma_1)$ are parameters estimated from the copula mixture model; ρ is estimated from the single-component model. The number of components is selected using a likelihood ratio test and the p-value of the test statistics is determined using a parametric bootstrap approach based on 100 bootstrap samples.

	Peakseq	MACS	SPP	Fseq	HotSpot	cisgenome	erange	quest	sissrs
π_1	0.69	0.84	0.77	0.74	0.69	0.85	0.72	0.72	1
ρ_1	0.89	0.89	0.88	0.82	0.88	0.65	0.81	0.67	0.24
μ_1	2.27	2.07	2.28	2.12	1.62	2.05	2.04	2.01	7.27
σ_1	0.87	1.34	1.05	0.86	0.64	1.35	0.90	1.39	0.03
ρ	0.87	0.87	0.86	0.83	0.78	0.66	0.80	0.66	0.23
p value	0	0	0	0	0	1	0	1	1

described in Section 2.3. For example, at 5% IDR, the top 27500 peaks can be called using MACS. It can also be used to compare the overall reproducibility of different peak callers. For example, while Peakseq, MACS and SPP on average have about 3% irreproducible peaks when selecting the top 25000 peaks, most of other peak callers have already reached a much higher IDR before identifying the top 10000 peaks. On the basis of the number of peaks before reaching high IDR, the three most reproducible peak callers on this dataset are Peakseq, MACS and SPP, then followed by Fseq, then others. This result is consistent with the reproducibility comparison from the correspondence profile (Figure 5). Note that though reproducbility can be compared using Figure 5, it does not provide clear information on how many peaks should be selected.

4.2.3. *Evaluating the biological relevance of the reproducibility assessment.* The results of our method have shown that four peak callers produced peak calls with low reproducibility on the examined data set. To evaluate if the reproducibility analysis correctly identifies suboptimal results, we check the accuracy of of peak identifications by comparing the identified peaks with the regions predicted by a motif prediction method (Kheradpour et al., 2007), which is a computational method not based on ChIP-seq experiments. Although the computational prediction is not ground truth, it serves as an independent source for a crude evaluation of the accuracy of peak callers. A common strategy to evaluate accuracy in this type of high-throughput experiments using an external standard, is to compute the proportion of identified signals that overlap with the external standard as a function of decreasing significance. We adopt this strategy to verify the reproducibility
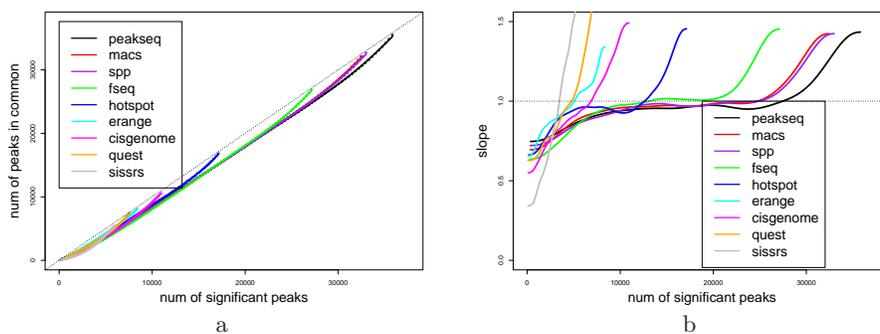
FIG 5. *The correspondence profiles along the decreasing order of significance, plotted for 9 peak callers on a CTCF Chip-seq experiment from ENCODE. a. Correspondence curve (Ψ_n). X-axis: the number of peaks identified on a replicate. Y-axis: the number of peaks that are commonly identified on both replicates. b. Change of correspondence curve (Ψ'_n).*
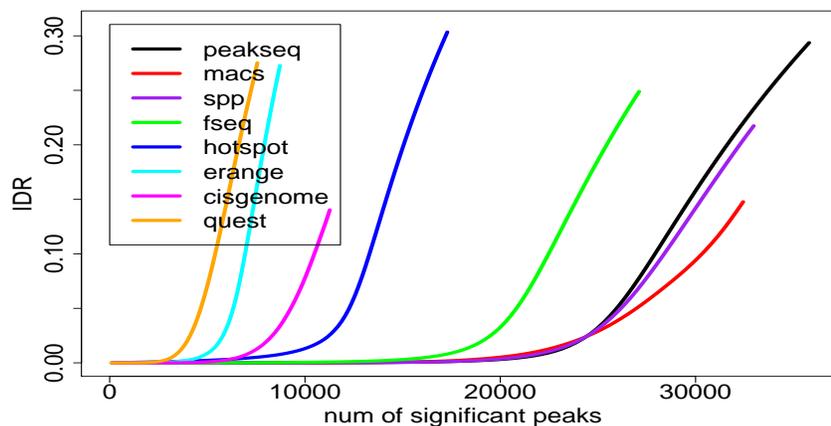
FIG 6. *Irreproducible discovery rate (IDR) at different numbers of selected peaks, plotted at various idr cutoffs for eight peak callers on a CTCF Chip-seq experiment from ENCODE. Peaks are selected using local idr. X-axis: the number of selected significant peaks, Y-axis: Irreproducible discovery rate (IDR). SISSRS is not shown because its results are highly inconsistent and all peaks are grouped into a low correlation group.*
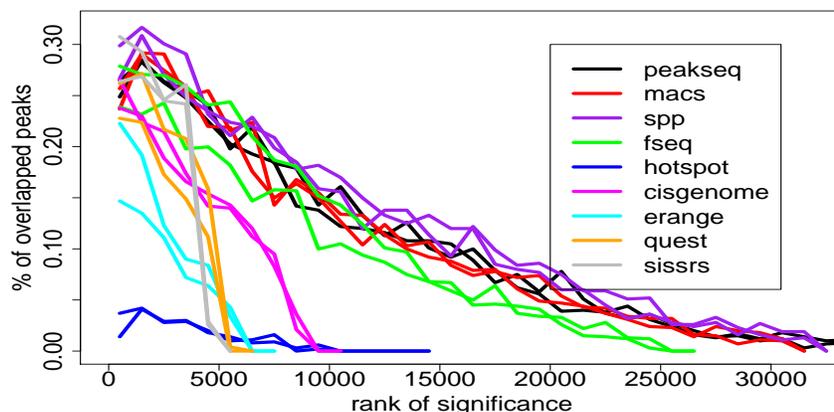
FIG 7. *The proportion of identified peaks that are overlapped with the binding regions predicted by a computational method, plotted for peaks identified from the nine peak callers in the decreasing order of significance.*

assessment.

As shown in Figure 7, though in principle reproducibility by no means implies accuracy, the reproducibility assessment empirically agrees with the accuracy assessment reasonably well: the three peak calling results with the highest reproducibility in Figure 6 also show the highest accuracy; the ones that are reported to be less reproducible do show unsatisfactory accuracy, as was later confirmed by hand curation. This illustrates the potential of our method as a quality measure.

5. Discussion. We have presented a new statistical method for measuring the reproducibility of high-throughput experiments and improving signal identification using replicates. Using simulated and real data, we have illustrated the potential of our method for improving the accuracy of signal identification and demonstrated its use as a quality measure to identify suboptimal results.

As no assumption is made on the scale of the scores, this model offers great flexibility both as a reproducibility measure and as a method for pooling replicate experiments. It is applicable for any scoring system that produces continuous ranking to reflect the relative ordering of the signals, including both probabilistic-based scores, such as p-values, and nonprobabilistic-based scores, such as fold of enrichment. In addition, this flexibility provides a principled way to make calls for signals that are scored on nonprobabilistic

measures, where arbitrary judgment often is involved for setting thresholds because signal selection can not be done using classical multiple testing methods. Moreover, because consistency between replicates is an internal standard that is independent of the scoring schemes and is comparable across datasets, the proposed reproducibility criterion is suited for setting uniform standards for selecting signals for data from multiple sources. Because our measure of consistency is not confounded by platform-dependent thresholds, inter-platform consistency can be objectively assessed.

As a method for selecting signals, our method uses the reproducibility between replicates, in conjunction with the significance of the scores, to classify the signals. Of course, reproducibility is only a necessary but not sufficient condition to accuracy. If the replicates are generated in presence of a systematic bias that introduces false association, the selection from this procedure, as in any other methods for pooling information, can be misleading. On the other hand, lack of reproducibility on replicates indeed is a flag on the quality of data or data processing procedure. We have demonstrated that the proposed graphical and quantitative reproducibility measures can be used for identifying inconsistent results and monitoring the performance of the experimental and data analytical procedures.

6. Appendix 1: Estimation algorithm for the copula mixture model. Here we describe the iterative procedure to estimate the parameter $\theta = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$ in detail.

1. Compute the empirical marginal CDF $\hat{F}_j(x_{i,j}) = \frac{r_{i,j}}{n}$ where $r_{i,j}$ is the rank of $x_{i,j}$ on replicate j and n is the number of pairs.
2. Rescale $\hat{F}_j(x_{i,j})$ by $u_{i,j} \equiv \frac{n}{n+1}\hat{F}_j(x_{i,j})$ to avoid potential unboundedness of $G^{-1}(u_{i,j})$ if $u_{i,j}$'s tend to one.
3. Initialize $\theta = \theta_0$.
4. Compute the pseudo-data $z_{i,j} = G^{-1}(u_{i,j} \mid \theta)$. As G^{-1} does not have a closed form, G is first computed on a grid of 1000 points for $u \in [\min(-3, \mu_1 - 3), \max(3, \mu_1 + 3)]$, then $z_{i,j}$ is obtained by linear interpolation on the grid.

5. Run EM to maximize the log-likelihood of pseudo data,

$$l(\theta) = \sum_{i=1}^{n} \log((1 - \pi_1)h_0(z_{i,1}, z_{i,2}) + \pi_1 h_1(z_{i,1}, z_{i,2}; \mu_1, \sigma_1^2, \rho_1)),$$

to get $\theta^{(t)} = \text{argmax}_\theta l$. The E-step and M-step are described below.
6. Set $\theta = \theta^{(t)}$ and go to step 4 until convergence.

Here we describe the EM algorithm in step 5 above. To proceed, we denote K_i as the latent variable, then the complete pseudo log-likelihood for the augmented data $Y_i \equiv (Z_i, K_i)$ is

(6.1)
$$l^c = \sum_{i=1}^{n} [(1 - K_i)(\log(1 - \pi_1) + \log(h_0(z_{i,1}, z_{i,2}))) + K_i(\log \pi_1 + \log(h_1(z_{i,1}, z_{i,2}; \mu_1, \sigma_1^2, \rho_1)))]$$

E-step:

$$(6.2) \quad Q(\theta, \theta^{(t)}) \equiv E(l^C(\Psi) \mid \mathbf{z}, \theta^{(t)})$$
$$= \sum_{i=1}^{n} \{ P(K_i = 0)[\log(1 - \pi_1^{(t)}) + \log(h_0(z_{i,1}^{(t)}, z_{i,2}^{(t)}))]$$
$$+ P(K_i = 1)[\log \pi_1^{(t)} + \log(h_1(z_{i,1}^{(t)}, z_{i,2}^{(t)}; \mu_1^{(t)}, (\sigma_1^2)^{(t)}, \rho_1^{(t)}))] \}$$

Then

$$(6.3) \quad K_i^{(t+1)} \equiv E(K_i \mid \mathbf{z}_i, \theta^{(t)})$$
$$= \frac{P(K_i = 1, \mathbf{z}_i \mid \theta^{(t)})}{P(\mathbf{z}_i \mid \theta^{(t)})}$$
$$= \frac{\pi_1^{(t)} h_0(z_{i,1}^{(t)}, z_{i,2}^{(t)})}{(1 - \pi_1^{(t)})h_0(z_{i,1}^{(t)}, z_{i,2}^{(t)}) + \pi_1^{(t)} h_1(z_{i,1}^{(t)}, z_{i,2}^{(t)}; \mu_1^{(t)}, (\sigma_1^2)^{(t)}, \rho_1^{(t)})}$$

M-step:
Now we need maximize $Q(\theta, \theta^{(t)})$. The MLE of the mixing proportion is:

$$(6.4) \qquad\qquad \pi_1^{(t+1)} = \frac{\sum_{i=1}^{n} K_i^{(t+1)}}{n}$$

Only the 2nd term of $Q(\theta, \theta^{(t)})$ is relevant to $(\mu_1, \sigma_1^2, \rho_1)$, so this is equivalent to maximizing the following:

(6.5)

$$
El_z = \sum_{i=1}^{n} EK_i \{ \log(\frac{1}{2\pi\sigma_1^2\sqrt{1-\rho_1^2}})
$$
$$
- \frac{1}{2(1-\rho_1^2)} \left[\frac{(z_{i,1}-\mu_1)^2 - 2\rho_1(z_{i,1}-\mu_1)(z_{i,2}-\mu_1) + (z_{i,2}-\mu_1)^2}{\sigma_1^2} \right] \}
$$

Taking derivatives w.r.t. each term, we have the following:

(6.6)

$$
\frac{\partial l_z}{\partial \mu_1} = \sum_{i=1}^{n} \frac{K_i^{(t+1)}}{2(1-\rho_1^2)} \frac{2(z_{i,1}-\mu_1) + 2(z_{i,2}-\mu_1) - 2\rho(z_{i,1}+z_{i,2}-2\mu_1)}{\sigma_1^2}
$$

So

(6.7)
$$
\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} K_i^{(t+1)}(z_{i,1}+z_{i,2})}{2\sum_{i=1}^{n} K_i}
$$

(6.8)
$$
\frac{\partial l_z}{\partial \sigma_1^2} = \sum_{i=1}^{n} K^{(t+1)} \left[-\frac{1}{\sigma_1^2} + \frac{(z_{i,1}-\mu_1)^2 - 2\rho(z_{i,1}-\mu_1)(z_{i,2}-\mu_1) + (z_{i,2}-\mu_1)^2}{2\sigma_1^4(1-\rho_1^2)} \right]
$$

and

(6.9) $\frac{\partial l_z}{\partial \rho_1} = \sum_{i=1}^{n} K^{(t+1)} [\frac{\rho_1}{1-\rho_1^2} + \frac{1}{1-\rho_1^2} \frac{(z_{i,1}-\mu_1)(z_{i,2}-\mu_1)}{\sigma_1^2}$

$$
- \frac{\rho_1}{(1-\rho_1^2)^2} \frac{(z_{i,1}-\mu_1)^2 - 2\rho(z_{i,1}-\mu_1)(z_{i,2}-\mu_1) + (z_{i,2}-\mu_1)^2}{\sigma_1^2}]
$$

Solving the above together, we get

(6.10)
$$
(\sigma_1^2)^{(t+1)} = \frac{\sum_{i=1}^{n} K_i((z_{i,1}-\mu_1)^2 + (z_{i,2}-\mu_1)^2)}{2\sum_{i=1}^{n} K_i}
$$
$$
\rho_1^{(t+1)} = \frac{2\sum_{i=1}^{n} K_i(z_{i,1}-\mu_1)(z_{i,2}-\mu_1)}{\sum_{i=1}^{n} K_i[(z_{i,1}-\mu_1)^2 + (z_{i,2}-\mu_1)^2]}
$$

7. Appendix 2: Behavior of the limit of algorithm for n large.
We give a heuristic argument for the asymptotic behavior of the limit of our
algorithm when it converges. In another paper, we shall fill in the details of
this argument. Although we are unable to prove convergence of our present
algorithm with probability tending to 1, we shall exhibit another algorithm
converging with probability tending to 1 to the same limit as the present
one when the latter converges.

We begin with some notation. Let (U, V) be distributed according to the
copula density,

$$(7.1) \qquad c(u, v, \theta) = \frac{h(F^{-1}(u, \theta), G^{-1}(v, \theta), \theta)}{f(F^{-1}(u, \theta))g(G^{-1}(v, \theta))}$$

where $h(x, y, \theta)$ is a parametric family in $\theta \in \Theta \subset R^p$ open, $F(\cdot, \theta)$, $G(\cdot, \theta)$
are the cdf of X', Y', respectively under $h(\cdot, \cdot, \theta)$, and $f(\cdot, \theta)$, $g(\cdot, \theta)$ are the
marginal densities, if $(X', Y') \sim h(\cdot, \cdot, \theta)$.

In our case, $\theta = (\epsilon, \mu, \sigma, \rho)$,

$$(7.2) \qquad h(x, y, \theta) = (1 - \epsilon)\phi(x, y, \mu, \mu, \sigma^2, \sigma^2, \rho) + \epsilon\phi(x, y, 0, 0, 1, 1, 0)$$

where $\phi(\cdot, \cdot, \mu, \sigma^2, \rho)$ is the density of bivariate normal with mean μ, variance
σ^2 and correlation coefficient ρ.

Let x_i, y_i, $i = 1, \ldots, n$ be iid with density f where

$$(7.3) \qquad h(x, y) \equiv h(x, y, \theta, F, G) \equiv c(F(x), G(y), \theta)f(x)g(y),$$

where F, G are absolutely continuous marginal cdfs and $f(\cdot)$, $g(\cdot)$ are the
corresponding densities. Let θ_0 denote the true value of θ and F and G
be as above. Let $\hat{F}_n$, $\hat{G}_n$ be the empirical distribution functions of X, Y,
respectively.

The ranks we use are $(\hat{F}_n(X_i), \hat{G}_n(Y_i))$, $i = 1, \ldots, n$, where $\hat{F}_n$, $\hat{G}_n$ are
the empirical cdfs of (X, Y).

Our algorithm for finding θ is,

1. Initialize $\theta_1 = \hat{\theta}_0$
2. Let $X_i(\theta) \equiv F^{-1}(\hat{F}_n(X_i), \theta)$ and $Y_i(\theta) \equiv G^{-1}(\hat{G}_n(Y_i), \theta)$
3. Run EM to get, assuming the initial point is close enough to the arg
   max below,

$$(7.4) \qquad \hat{\theta}_1 = \text{argmax}_\theta \int \log h(x, y, \theta)dP_n(x, y, \theta_1)$$

where $P_n(x, y, \theta) \equiv \frac{1}{n}\sum_{i=1}^n 1(X_i(\theta) \leq x, Y_i(\theta) \leq y)$.

4. Set $\theta_1 = \hat{\theta}_1$ and return to (2).

Define

$$(7.5) \qquad T_n(t) \equiv \text{argmax}_\theta \int \log h(x, y, \theta) dP_n(x, y, t)$$

It is clear that if our algorithm converges to $\hat{\theta}$, then $\hat{\theta}$ is a fixed point of T_n, i.e. $T_n(\hat{\theta}) = \hat{\theta}$.

Let

$$T(t) = \text{argmax}_\theta \int \log h(x, y, \theta) dP(x, y, t)$$

where $P(x, y, t)$ is the cdf of $X(t)$, $Y(t)$. Evidently, θ_0 is the unique fixed point of T since $dP(x, y, \theta_0) = f(x, y, \theta_0)dxdy$.

Define $W_n(t)$ as the solution of

$$S_n(\theta, t) \equiv \int \nabla_\theta \log h(x, y, \theta) dP_n(x, y, t) = 0$$

which is nearest to t and a local maximum. Then, $\hat{\theta}_1 = W_n(\hat{\theta}_0)$ and $\hat{\theta}$ is a fixed point of W_n.

Similarly, define $W(t)$ as the solution of the population version of S_n

$$S(\theta, t) \equiv \int \nabla_\theta \log h(x, y, \theta) dP(x, y, t) = 0$$

which is closest to t and a local maximum. Clearly θ_0 is a fixed point of W.

Let $\hat{H}_n(\cdot, \cdot)$ be the empirical distribution function of (X_i, Y_i). We can rewrite,

$$(7.6) \qquad S_n(\theta, t) = \int \dot{l}(F^{-1}(\hat{F}_n(x), t), G^{-1}(\hat{G}_n(y), t), \theta) d\hat{H}_n(x, y)$$

where $\dot{l} = \nabla_\theta \log f$, and similarly

$$(7.7) \qquad S(\theta, t) = \int \dot{l}(F^{-1}(F(x), t), G^{-1}(G(y), t), \theta) dH(x, y)$$
$$= \int \dot{l}(x, y, \theta) h(x, y, t) dxdy$$

It is easy to see that θ_0 is a fixed point of W and is unique.

We will now study $\hat{\theta}_n$. We assume that

1. $\hat{\theta}_n$ exists
2. $\hat{\theta}_n$ corresponds to the unique fixed point of (7.5)
3. $\hat{\theta}_n$ is in $\{\theta : |\theta - \theta_0| < \epsilon\}$ for $\epsilon > 0$ to be specified.

For simplicity, we take $p = 1$ but this is inessential. As usual our point of departure is the equation

$$(7.8) \qquad - S(\theta_0, \hat{\theta}_n) = S_n(\hat{\theta}_n, \hat{\theta}_n) - S(\theta_0, \hat{\theta}_n)$$

We suppose that, as we shall show, under suitable conditions elsewhere, we can expand the right hand side of (7.8) as

$$(7.9) \qquad - \frac{\partial}{\partial \theta} S(\theta_0, \theta_0)(\hat{\theta}_n - \theta_0) + O_p(|\hat{\theta}_n - \theta_0|^2)$$

where

$$\frac{\partial S(\theta_0, \theta_0)}{\partial t} \mid_{t=\theta_0} = \int \dot{l}(x, y, \theta_0) \frac{\partial h(x, y, \theta_0)}{\partial t} \mid_{t=\theta_0} dxdy$$
$$= \int \dot{l}^2(x, y, \theta_0) h(x, y, \theta_0) dxdy$$

We now analyze the second term in (7.8). Expand formally the integrand in (7.6),

$$(7.10)$$

$$\dot{l}(F^{-1}(\hat{F}_n(x), \hat{\theta}_n), G^{-1}(\hat{G}_n(y), \hat{\theta}_n), \hat{\theta}_n)$$
$$= \dot{l}(F^{-1}(F(x), \hat{\theta}_n), G^{-1}(G(y), \hat{\theta}_n), \hat{\theta}_n)$$
$$+ \frac{1}{f(F^{-1}(F(x), \hat{\theta}_n))} \frac{\partial \dot{l}}{\partial x}(F^{-1}(F(x), \hat{\theta}_n), G^{-1}(G(y), \hat{\theta}_n), \hat{\theta}_n)(\hat{F}_n(x) - F(x))$$
$$+ \frac{1}{g(G^{-1}(G(y), \hat{\theta}_n))} \frac{\partial \dot{l}}{\partial y}(F^{-1}(F(x), \hat{\theta}_n), G^{-1}(G(y), \hat{\theta}_n), \hat{\theta}_n)(\hat{G}_n(y) - G(y)) + O_p(n^{-1})$$

Further, formally

$$(7.11)$$

$$\frac{1}{f(F^{-1}(F(x), \hat{\theta}_n))} \frac{\partial \dot{l}}{\partial x}(F^{-1}(F(x), \hat{\theta}_n), G^{-1}(G(y), \hat{\theta}_n), \hat{\theta}_n)(\hat{F}_n(x) - F(x))$$
$$= \frac{1}{f(F^{-1}(F(x), \theta_0))} \frac{\partial \dot{l}}{\partial x}(F^{-1}(F(x), \theta_0), G^{-1}(G(y), \theta_0), \theta_0)(\hat{F}_n(x) - F(x)) + O_p(|\hat{\theta}_n - \theta_0|n^{-1/2})$$

and a similar approximation holds for the third term in (7.10). Next note that,

$$(7.12) \qquad \dot{l}(F^{-1}(F(x), \hat{\theta}_n), G^{-1}(G(y), \hat{\theta}_n), \theta_n)$$
$$= \dot{l}(F^{-1}(F(x), \theta_0), G^{-1}(G(y), \theta_0), \theta_0) + O_p(|\hat{\theta}_n - \theta_0|)$$

Now expand

(7.13)

$$
\dot{l}(F^{-1}(F(x),\hat{\theta}_n),G^{-1}(G(y),\hat{\theta}_n),\theta_0)
$$
$$
=\dot{l}(F^{-1}(F(x),\theta_0),G^{-1}(G(y),\theta_0),\theta_0)+\frac{\partial\dot{l}}{\partial t}(F^{-1}(F(x),t),G^{-1}(G(y),t),\theta_0)\mid_{t=\theta_0}(\hat{\theta}_n-\theta_0)
$$
$$
+O_p(|\hat{\theta}_n-\theta_0|^2)
$$

Combining (7.10), (7.11), (7.12), and (7.13) we get,

(7.14)

$$
S_n(\hat{\theta}_n,\hat{\theta}_n)-S(\theta_0,\hat{\theta}_n)
$$
$$
=\int Q_1(x,\theta_0)(\hat{F}_n(x)-F(x))d\hat{H}_n(x,y)+\int Q_2(y,\theta_0)(\hat{G}_n(y)-G(y))d\hat{H}_n(x,y)
$$
$$
+\int\frac{\partial\dot{l}}{\partial t}(F^{-1}(F(x),t),G^{-1}(G(y),t),\theta_0)\mid_{t=\theta_0}dH(x,y)(\hat{\theta}_n-\theta_0)
$$
$$
+\int\dot{l}(F^{-1}(F(x),\theta_0),G^{-1}(G(y),\theta_0),\theta_0)d(\hat{H}_n-H)(x,y)+O_p(n^{-1}+|\hat{\theta}_n-\theta_0|^2)
$$

where $Q_1(x,\theta_0)$ is given in (7.10) and (7.11), and similarly for $Q_2(y,\theta_0)$. All terms but the last come from the expansions (7.11), (7.13) of $\dot{l}$ integrated with respect to H. The last comes from the difference between S_n as given and the expression in which the measure $\hat{H}_n(x,y)$ is replaced by $H(x,y)$.

Combining (7.14) and (7.9), we get finally,

(7.15)
$$
M(\theta_0)(\hat{\theta}_n-\theta_0)=\int\dot{l}(x,y,\theta_0)d\hat{H}(x,y)+\int\alpha(x,\theta_0)d\hat{F}_n(x)+\int\beta(y,\theta_0)d\hat{G}_n(y)+o_p(n^{-1/2})
$$

for suitable M,α,β, depending on $F(\cdot),G(\cdot)$ as well as θ_0. The heuristics yield asymptotic normality for $\hat{\theta}_n$. More than that $\hat{\theta}_n$ is asymptotically linear and its influence function is of the form $(\frac{\partial l}{\partial\theta}(x,y,\theta_0)+a(x,\theta_0,F,G)+b(y,\theta_0,F,G))M(\theta_0,F,G)$. By Proposition 1B (cf Bickel et al. (1993) pp. 65-66) it follows that $\hat{\theta}_n$ is always efficient, since its influence function lies in the targent space of the model $h(\cdot,\cdot,\theta,F,G)$, where $\theta\in\Theta$, F, G are positive densities, and f, g arbitrary.

Thus in principle, these estimates are preferable to those of Genest et al (Genest et al., 1995). In a paper making these heuristics rigorous and constructing an algorithm for $\hat{\theta}_n$ which will converge with probability tending to 1 under suitable conditions, we will also do further extensive simulations

for our Gaussian mixture and other models, comparing this procedure with that of Genest et al. We note that, in fact, no algorithm is given for the realization of their procedure except in a special case quite different from ours. We also note that our algorithm, just as that of Genest et al, depends on finding the "correct" local maximum of a multimodal function and thus depends crucially on finding appropriate starting points, in particular, running from a number of possible starting points.

References.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser. B 57*, 289–300.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics 25*, 60–83.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models* (1 ed.). Johns Hopkins University Press.

Blest, D. (2000). Rank correlation - An alternative measure. *Australian & New Zealand Journal of Statistics 42*(1), 101–111.

Boulesteix, A.-L. and M. Slawski (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics 10*(5), 556–568.

Boyle, A. P., J. Guinney, G. E. Crawford, and T. S. Furey (2008). F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics 24*(21), 2537–2538.

Da Costa, J. and C. Soares (2005). A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics 47*(4), 515–529.

Deheuvels, P. (1979). La fonction de dépendence empitique et ses propriétés. un test non paramétrique d'indépendence. *Acad Roy Belg Bul Cl Sci 5*(65), 274–292.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 39*(1), 1–38.

Efron, B. (2004a). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association 99*(1), 96–104.

Efron, B. (2004b). Local false discovery rate. Technical report, Stanford University, Dept of Statistics.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) Project. *Science 306*(5696), 636–640.

Fisher, N. I. and P. Switzer (1985). Chi-plots for assessing dependence. *Biometrika 72*(2), 253–265.

Fisher, N. I. and P. Switzer (2001). Graphical assessment of dependence: Is a picture worth 100 tests? *The American Statistician 55*(3), 233–239.

Fisher, R. A. (1932). *Statistical methods for research workers* (1 ed.). Oliver and Boyd.

Genest, C. (2003). Detecting dependence with kendall plots. *The American Statistician 57*(4), 275–284.

Genest, C., K. Ghoudi, and L. P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika 82*(3), 543–552.

Genest, C. and J. Plante (2003). On Blest's measure of rank correlation. *Canadian Journal of Statistics 31*(1), 35–52.

Genovese, C. and L. Wasserman (2002). Operating characteristic and extensions of the

false discovery rate procedure. *Journal of the Royal Statistical Society, Ser. B 64*, 499–517.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Annals of Statistics 32*, 1035–1061.

Hu, L. (2006). Dependence patterns across financial markets: a mixed copula approach. *Applied financial economics 16*(10), 717–729.

Ji, H., H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology 26*(11), 1293–1300.

Joe, H. (1997). *Multivariate models and dependence concepts* (1 ed.). Chapman & Hall/CRC.

Jothi, R., S. Cuddapah, A. Barski, K. Cui, and K. Zhao (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research 36*(16), 5221–5231.

Kallenberg, W. C. M. and T. Ledwina (1999). Data-driven rank tests for independence. *Journal of the American Statistical Association 94*, 285–301.

Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology 26*(12), 1351–1359.

Kheradpour, P., A. Stark, S. Roy, and M. Kellis (2007). Reliable prediction of regulator targets using 12 drosophila genomes. *Genome Res. 17*(12), 1919–1931.

Lee, M., F. Kuo, G. Whitmore, and J. Sklar (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *The Proceedings of the National Academy of Sciences of the United States of America 97*(18), 9834–9839.

MAQC consortium (2006). The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology 24*(9), 1151–1161.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics 36*(3), 318–324.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods 5*(7), 621–628.

Nelson, R. B. (1999). *An introduction to copula* (2 ed.). Springer Verlag.

Oakes, D. (1994). Multivariate survival distribution. *J. Nonparametr. Statist. 3*, 343–354.

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics 10*, 669–680.

Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology 27*(1), 66–75.

Sklar, A. (1959). Fonctions de rpartition n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universit de Paris 8*, 229–231.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Ser. B 64*, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics 31*, 2013–2035.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and J. Williams, R. M. (1949). *The American soldier: Vol. 1. Adjustment during army life.* Princeton University Press.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of American Statistical Association 102*(479), 901–912.

Thurman, R., M. Hawrylycz, S. Kuehn, E. Haugen, and S. Stamatoyannopoulos. Hotspot:

a scan statistic for identifying enriched regions of short-read sequence tags. In preparation.

Valouev, A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods 5*(9), 829–834.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology 9*(9), R137.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
367 EVANS HALL, MAIL STOP 3860
BERKELEY, CA 94720, USA
E-MAIL: qli@stat.berkeley.edu; ben@newton.berkeley.edu; hhuang@stat.berkeley.edu; bickel@stat.berkeley.edu