MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS[1]

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. Reproducibility of high-throughput assays, such as the level of agreement between results from replicate experiments across (biological or technical) replicate samples, test sites or experimental or data analytical platforms, is a constant concern in their scientific applications [e.g., MAQC consortium (2006) in microarray experiments, Park (2009) in ChIP-seq technology]. Metrics that objectively assess the reproducibility of high-thoughput assays are important for producing reliable scientific discoveries and monitoring the performances of data generating procedures.

An important criterion for assessing reproducibility of results from high-throughput experiments is how reproducibly top ranked signals are reported in

replicate experiments. These signals and their significance scores, often presented as the primary results to be accessed by downstream steps, are critical for prioritizing follow-up studies. 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 [see MAQC consortium (2006) and Kuo et al. (2006) for examples in microarray studies]. However, the Spearman's correlation coefficient actually is not entirely suitable for measuring the reproducibility between two rankings in this type of application. First, this summary depends on the choice of significance thresholds and may render false assessments that reflect 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 correlations 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 associated only with heuristics-based scores, indeed, it is also present for probabilistic-based scores, such as p-values, if the probabilistic model is ill-defined. 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, the rank correlation treats all ranks equally, though the differences in the top ranks seem to be more critical for judging the reproducibility of findings from high-throughput experiments. 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 [see 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 results in replicate experiments. Instead of depending on a prespecified threshold, we describe reproducibility 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 localize the possible breakdown of consistency empirically. We further quantify reproducibility by classifying 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, our model assigns each signal a reproducibility index, which estimates its probability to be reproducible. Based on this index, we then define the irreproducible discovery rate (IDR) and a selection procedure, in a fashion analogous to their counterparts in multiple testing, to

rank and select signals. As we will illustrate, the selection by this reproducibility criterion provides the potential for more accurate classification. The overall reproducibility of the replicates is described using IDR as the average amount of irreproducibility in the signals selected.

The proposed approach, indeed, is a general method that can be applied to any ranking systems that produce scores without ties, though we discuss it in the context of high-throughput experiments. Because our copula-based approach does not make any parametric assumptions on the marginal distributions of scores, it is applicable to both probabilistic- and heuristic-based scores. When a threshold is difficult to determine in a scoring system, for example, heuristic-based scores, it provides a reproducibility-based criterion for setting selection thresholds.

In the next section we present the proposed graphical tool (Section 2.1), the copula mixture model and its estimation procedure (Section 2.2), and the reproducibility criterion (Section 2.3). In Section 3 we use simulations to evaluate the performance of our model, 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 in replicate experiments, infer the reliability of signals identified by each algorithm, and demonstrate how to use our method to identify suboptimal results. Section 5 is a general discussion. Finally, we present a heuristic justification of our algorithm on optimality grounds in the supplementary materials [Li et al. (2011)].

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 that each putative signal has been assigned a score that relates to the strength of the evidence for the signal to be real on the corresponding replicate by some data analysis method. The score can be either heuristic based (e.g., fold enrichment) or probabilistic based (e.g., p-value). We further assume that all the signals are assigned distinct significance scores and that 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 (see supplementary materials [Li et al. (2011)]).

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 between ranks may be observed when getting into the noise level. This change of association provides an internal indicator of the transition from real signal to noise. We will use this information in measuring the reproducibility of signals.

In this section we first present a graphical tool (Section 2.1) for visualizing the change of association and localizing the possible breakdown of association, empirically without model assumptions. We then present a model-based approach (Section 2.2), which quantifies the heterogeneity of association and leads to a reproducibility criterion for threshold selection.

2.1. *Displaying the change of association.* As we mentioned, the bivariate association between the significance scores is expected to be positive for significant signals, then transits to zero when getting into noise level. 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 are useful tools for displaying bivariate dependence structure, because they are invariant with respect to monotone transformations of the variables 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 and Boies (2003)], for visualizing the presence of association in samples from continuous bivariate distributions. Related to nonparametric tests of independence, these graphs primarily are designed for detecting bivariate dependence by representing the presence of association as departures from the pattern under 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 and Boies (2003). However, these graphs are less informative, when heterogeneity of association, such as the one described here, is present. (See Figure 2 in Section 2.1.2 for an illustration on a real data set with mixed populations.)

We now present our rank-based graph, which we refer to as a correspondence curve, intended to explicitly display the aforementioned change of association.

2.1.1. *Correspondence curves.*   Let $(X_{1,1}, X_{1,2}), \dots, (X_{n,1}, X_{n,2})$ be a sample of scores of n signals on a pair of replicates. Define

$$(2.1) \qquad \Psi_n(t, v) = \frac{1}{n} \sum_{i=1}^{n} 1\big(X_{i,1} > x_{(\lceil(1-t)n\rceil),1}, X_{i,2} > x_{(\lceil(1-v)n\rceil),2}\big),$$

$$0 < t \le 1, 0 < v \le 1,$$

where $x_{(\lceil(1-t)n\rceil),1}$ and $x_{(\lceil(1-v)n\rceil),2}$ denote the order statistics of X_1 and X_2, respectively. $\Psi_n(t, v)$ essentially describes the proportion of the pairs that are ranked both on the upper $t\%$ of X_1 and on the upper $v\%$ of X_2, that is, 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 of copulas.) 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. (See supplementary materials [Li et al. (2011)] for derivation.)

Let $R(X_{i,1})$ and $R(X_{i,2})$ be the ranks of $X_{i,1}$ and $X_{i,2}$, respectively.

(1) If $R(X_{i,1}) = R(X_{i,2})$ for $X_{i,j} \in (F_j^{-1}(1 - t), F_j^{-1}(1 - t_0)]$, $j = 1, 2$, with $0 \le t_0 \le t \le 1$, $\Psi(t) = \Psi(t_0) + t - t_0$ and $\Psi'(t) = 1$.

(2) If $R(X_{i,1}) \perp R(X_{i,2})$ for $X_{i,j} \in (F_j^{-1}(1 - t), F_j^{-1}(1)]$, $j = 1, 2$, with $0 \le t \le 1$, $\Psi(t) = t^2$ and $\Psi'(t) = 2t$.

(3) If $R(X_{i,1}) = R(X_{i,2})$ for $X_{i,j} \in (F_j^{-1}(1 - t_0), F_j^{-1}(1)]$ and $R(X_1) \perp R(X_2)$ for $X_{i,j} \in (F_j^{-1}(0), F_j^{-1}(1 - t_0)]$, $j = 1, 2$, with $0 \le t_0 \le 1$, then for $t_0 \le t \le 1$, $\Psi(t) = \frac{t^2 - 2tt_0 + t_0}{1 - t_0}$ 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. The same properties are approximately followed in the corresponding sample version Ψ_n and Ψ'_n with finite differences replacing derivatives.

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 $0 \le t \le 1$. The resulting graphs, which we will refer to as a correspondence curve (or a change of correspondence curve, resp.), depend on $X_{i,1}$ and $X_{i,2}$ only through their ranks, and are invariant to both location and scale transformation on $X_{i,1}$ and $X_{i,2}$. Corresponding to the three special cases described earlier, the curves have the following patterns:

(1) When $R(X_{i,1})$ and $R(X_{i,2})$ are perfectly correlated for $i = 1, \dots, n$, 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 $R(X_{i,1})$ and $R(X_{i,2})$ are independent for $i = 1, \ldots, n$, 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 straight line of slope of $2t$.

(3) When $R(X_{i,1})$ and $R(X_{i,2})$ are perfectly correlated for the top t_0n observations and independent for the remaining $(1 - t_0)n$, top t_0n 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 $(1 - t_0)n$ 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.

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 toward the 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 become spurious.

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 remaining 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 by the gap between the disjoint lines with 0 and positive slopes, which makes Ψ'_n a
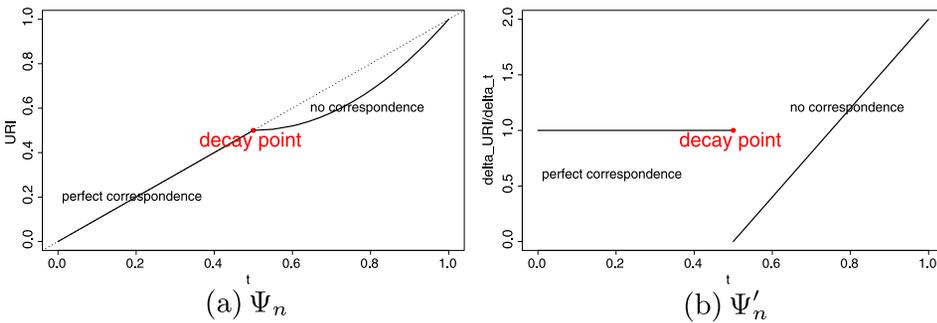

FIG. 1. *An illustration of the correspondence profile in an idealized case, where top* 50% *are genuine signals and bottom* 50% *are noise. In this case, all signals are ranked higher than noise; two rank lists have perfect correspondence for signals and no correspondence for noise.* (a) *Correspondence curve.* (b) *Change of correspondence curve.*

better choice for inspecting and localizing the transition than Ψ_n, especially when the transition is less sharp. More simulated examples are presented in Section 3 to illustrate the curves in the presence and absence of the transition of association in more realistic settings.

We now compare the Ψ_n' plot with the Chi-plot and the K-plot using a real example considered in Kallenberg and Ledwina (1999), Fisher and Switzer (2001), Genest and Boies (2003). This data set consists of 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-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 [cf. Fisher and Switzer (2001); Genest and Boies (2003)] suggests the data may involve more than simple monotone association [Fisher and Switzer (2001); Genest and Boies (2003)], neither of these plots manifest heterogeneity of association. In the Ψ_n' curve [Figure 2(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).
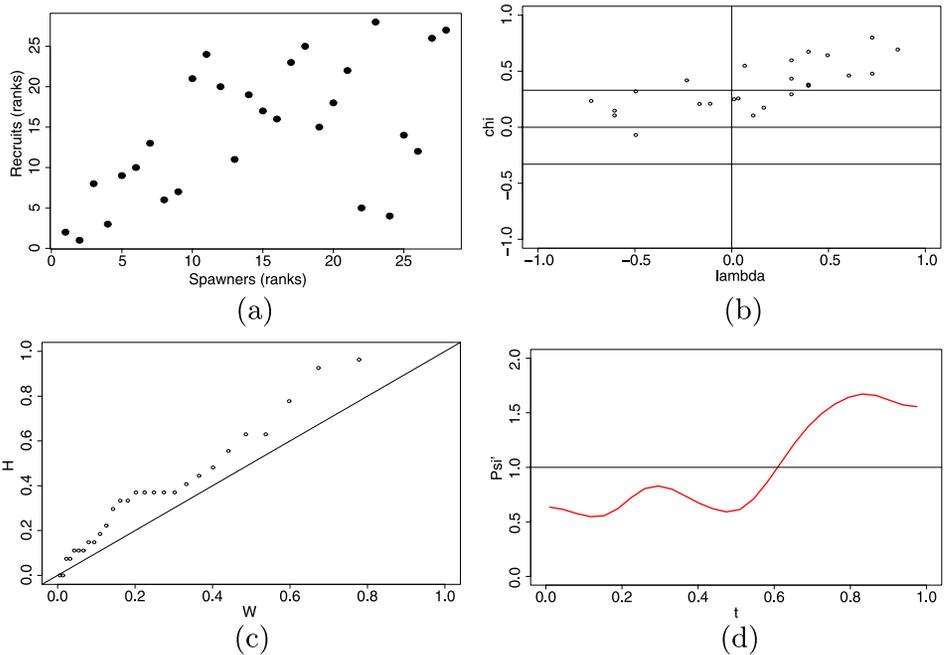

FIG. 2. *Rank scatterplot* (a), *Chi-plot* (b), *K-plot* (c) *and the change of correspondence curve* (d) *for salmon data, which consists of* 28 *measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-sized fish in the Skeena River. The curve of* Ψ_n' *is produced by taking derivative on the spline that fits* Ψ_n *with df* = 6.4.

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. Throughout this section, we will suppose, for simplicity, that we are dealing with a sample of i.i.d. observations from a population. 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, as we are concerned with first order effects.

In general, genuine signals tend to be more reproducible and score 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 to permit arbitrary scales. Though using ranks, instead of the raw values of scores, generally causes some loss of information, this loss is known to be asymptotically negligible [Lehmann (2006)]. 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 $C(u) = P(U_1 \leq u_1, \ldots, U_p \leq u_p)$, in which each $U_j \sim \text{Unif}[0, 1]$ and $u = (u_1, \ldots, u_p)$. 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

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

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

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

for a sample of size n, where $x_{(i),1}$ and $x_{(j),2}$ denote 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 $(z_1, z_2) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, 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 i.i.d. 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 $K_i = 1$, we assume the pairs of scores for genuine (resp., spurious) signals are independent draws from a continuous bivariate distribution with density $f_1(\cdot, \cdot)$ [resp., $f_0(\cdot, \cdot)$, given $K_i = 0$] with joint distribution $F_1(\cdot, \cdot)$ [resp., $F_0(\cdot, \cdot)$]. 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_i), $F_j = \pi_0 F_j^0 + \pi_1 F_j^1$, where F_j is the marginal distribution of the jth coordinate and F_j^k is the marginal distribution of the corresponding kth component.

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

As the scores from $F_1(\cdot, \cdot)$ are expected to be higher than the scores from $F_0(\cdot, \cdot)$, we assume $\mathbf{z}_1$ has a higher mean than $\mathbf{z}_0$. Since spurious signals are presumably less reproducible, we assume corresponding signals on the replicates to be independent, that is, $\rho_0 = 0$; whereas, since genuine signals usually are positively associated between replicates, we expect $\rho_1 > 0$, though ρ_1 is not required to be positive in our model. It also seems natural to assume that the underlying latent variables, reflecting 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, but not the means and variances of the latent variables. Thus, the parametric model generating our copula can be described as follows:

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

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

where $\mu_0 = 0$, $\mu_1 > 0$, $\sigma_0^2 = 1$, $\rho_0 = 0$, $0 < \rho_1 \leq 1$.

Let

$$(2.5b) \quad \begin{aligned} u_{i,1} &\equiv G(z_{i,1}) = \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,1} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,1}), \\ u_{i,2} &\equiv G(z_{i,2}) = \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,2} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,2}). \end{aligned}$$

Our actual observations are

$$(2.5c) \quad \begin{aligned} x_{i,1} &= F_1^{-1}(u_{i,1}), \\ x_{i,2} &= F_2^{-1}(u_{i,2}), \end{aligned}$$

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

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

(2.6a) $$+ \pi_1 h_1(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))]$$

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

is a copula density function with

$$h_0 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \text{and} \quad h_1 \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1^2 \\ \rho_1\sigma_1^2 & \sigma_1^2 \end{pmatrix}\right),$$

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) $$\Pr(K_i = 0 \mid (x_{i1}, x_{i2}); \theta) = \frac{\pi_0 h_0(G^{-1}(F_1(x_{i,1}), G^{-1}(F_2(x_{i,2}))))}{\sum_{k=0,1} \pi_k h_k(G^{-1}(F_1(x_{i,1}), G^{-1}(F_2(x_{i,2}))))}.$$

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. In this model, all the copulas in C_{mix} are assumed to have identical marginal distributions. In contrast, the copula in our model not only has mixed associations, but also allows different associations to occur with different marginal distributions (F_j^0 and F_j^1), thus can be viewed as a generalization of the case with the same marginal distribution. In addition, our modeling goal is to cluster the observations into groups with homogeneous associations, instead of data fitting. This difference in marginal distributions calls for nonstandard estimation, which 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 models 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 (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 on marginal distributions, which makes 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, Laird and Rubin (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 the supplementary materials [Li et al. (2011)]. 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 the supplementary materials [Li et al. (2011), Section 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. This

criterion can be used to assess the reproducibility of both individual signals and the overall 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 (2004); Storey (2002)]. Based on this mixture model, the local false discovery rate, which is the posterior probability of being in the null component $\text{Lfdr}(\cdot) = (1 - \pi) f_0(\cdot)/f(\cdot)$, was introduced to compute the individual significance level [Efron (2004)]. Sun and Cai (2007) show, again for the i.i.d. 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. However, the two classes have different interpretation and representation: The two classes represent irreproducible measurements and reproducible measurements in our model, 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

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

This quantity can be thought of 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,

$$(2.10)$$
$$\begin{aligned} \text{IDR}(\gamma) &= P(\textit{irreproducible} \mid i \in I_\gamma) \\ &= \frac{\pi_0 \int_{I_\gamma} dH_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}{\int_{I_\gamma} dH(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))}, \end{aligned}$$

where $I_\gamma = \{(x_{i,1}, x_{i,2}) : \text{idr}(x_{i,1}, x_{i,2}) < \gamma\}$, 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 $l = \max\{i : \frac{1}{i} \sum_{j=1}^{i} idr_j \leq \alpha\}$. 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.

3.1. *Illustration of correspondence curves.* To show the prototypical plots of more realistic cases, we use simulated data to compare and contrast the curves in presence and absence of the transition of association described in Section 2.1. (Figure 3). The case where no transition occurs is illustrated using two single-component bivariate Gaussian distributions with homogeneous association, $\rho = 0$ [Figure 3(a)] and $\rho = 0.8$ [Figure 3(b)], 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 3(c)] and $\rho_1 = 0.8$ [Figure 3(d)], respectively.

As in the idealized example (Figure 1), the characteristic transition of curves is observed when the transition of association is present [Figure 3(c), (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.

3.2. *Copula mixture model.* 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 assess the robustness of our method to the violation of one of its underlying model assumptions. In each simulation, we also compare the performance with two existing methods for combining significance scores across samples. However, as existing combination methods can be applied to only p-values, we use p-values as the significance scores in the comparison, though our method can be applied to arbitrary scores with continous marginal distributions. Here we consider the scenario when the p-values are not well calibrated but are reflective of the relative strength of evidence that the signals are real, and assess the accuracy of thresholds selected by all methods of comparison. These simulations also provide a helpful check on the convergence of our estimation procedure.
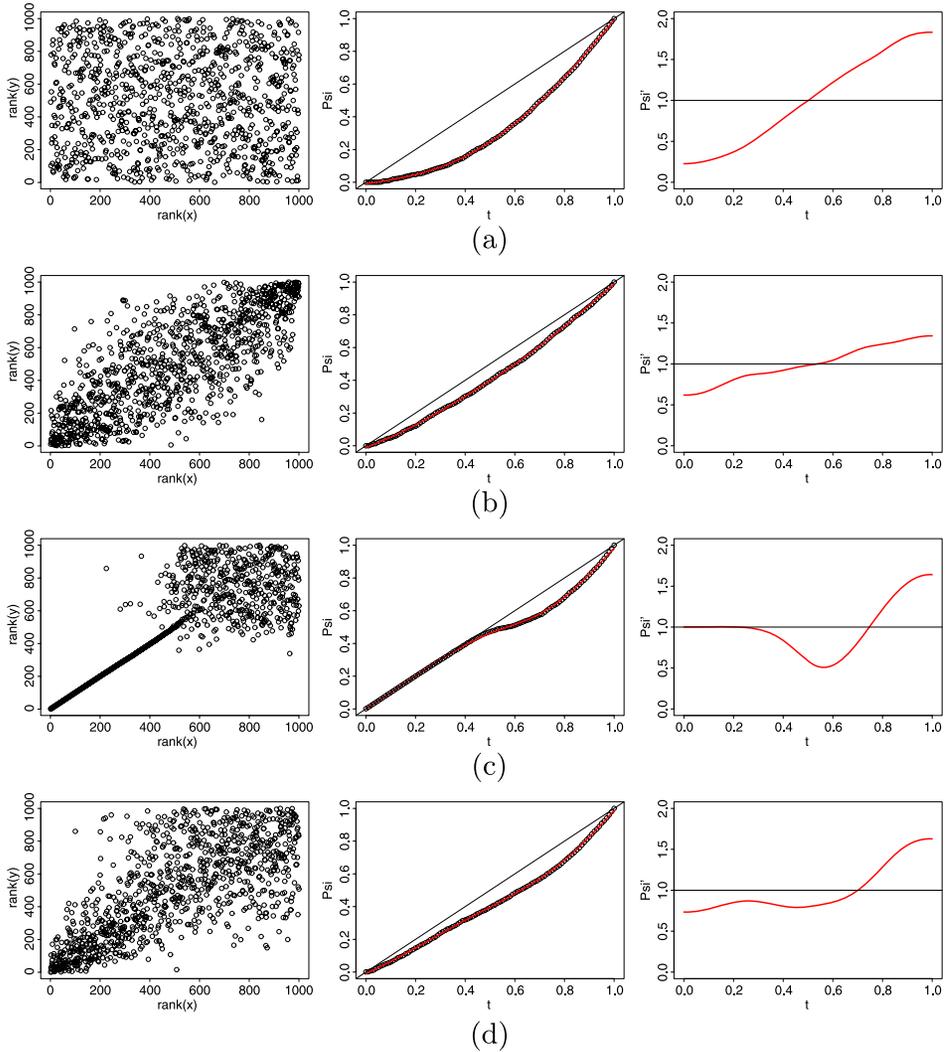
FIG. 3. *Behavior of correspondence curves when data consists of homogeneous and heterogeneous association. From left to right, the three columns are the scatterplot of ranks, the curve of Ψ and the curve of Ψ'. (a) Bivariate Gaussian distribution with $\rho = 0$. (b) Bivariate Gaussian distribution 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 the derivative on the spline that fits Ψ_n with $df = 6.4$.*

In each simulation study, we generate a sample of n pairs of signals on two replicates. Each pair of observed signals $(Z_{i1}, Z_{i,2})$ $(i = 1, \ldots, n)$ is a noisy realization of a latent signal Z_i, which is independently and identically generated from

the following normal mixture model:

$$K_i \sim \text{Bernoulli}(\pi_1),$$

(3.1)
$$Z_i \mid K_i = k \sim N(\mu_k, \tau_k^2), \qquad k = 0, \ldots, K - 1,$$

$$Z_{i,j} \mid K_i = k, Z_i = Z_i + \varepsilon_{ijk}, \qquad j = 1, 2,$$

$$\varepsilon_{ijk} \sim N(0, \omega_k^2).$$

As can easily be seen from the joint distribution of $(Z_{i,1}, Z_{i,2}) \mid K_i$,

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

where $\rho_k = \frac{\tau_k^2}{\tau_k^2 + \omega_k^2}$; this model is in fact a reparameterization of our model (2.5a) when $K = 2$: $(Z_{i,1}, Z_{i,2})$ directly corresponds to the latent Gaussian variables in (2.5a) when setting $\mu_0 = 0$, $\mu_1 > 0$, $\tau_k^2 + \omega_k^2 = 1$, $\rho_0 = 0$ and $k = 0, 1$. When $K > 2$, this model is convenient for simulating data from multiple components and investigating the robustness of our method to the violation of the assumption that the data consists of a reproducible and an irreproducible component. After $Z_{i,j}$ is simulated, we transform $Z_{i,j}$ to a Student's t distribution by a probability integral transformation $F_{t_5}^{-1}(G(Z_{i,j}))$, where F_{t_5} is the cdf of t distribution with $df = 5$ and G is as defined in (2.5b). We then obtain p-values from a one-sided z-test for $H_0 : \mu = 0$ vs $H_1 : \mu > 0$, and use them as the significance score $X_{i,j}$. This procedure is equivalent to applying a z-test to a t distribution, thus generating p-values that are not calibrated but are reflective of the relative strength of evidence that the signals are real. It is equivalent to letting $F_j^{-1}(\cdot) = 1 - \Phi(F_{t_5}^{-1}(\cdot))$ in (2.5c) for our model.

With p-values 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 (1925)] 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 signals 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 distribution $N(0, 1)$ under H_0, where Φ is the standard normal CDF. For each pair of signals, we compute Q_i (S_i, resp.) and its corresponding p-values p_i^Q (p_i^S, resp.), then estimate the corresponding false discovery rates (FDR) by computing q-values [Storey (2003)] based on p_i^Q (p_i^S, resp.), using R package "qvalue." FDR is estimated similarly for p-values on the individual replicates.

TABLE 1

Simulation parameters and parameter estimation in the simulation studies of 100 *data sets. Each data set consists of* 10,000 *pairs of observations. The simulation parameters are estimated from a ChIP-seq data set. In all simulations,* $\mu_0 = 0$, $\sigma_0^2 = 1$ *and* $\rho_0 = 0$. *In S1–S3,* $\pi_0 = 1 - \pi_1$. *S4 has a third component with* $\mu_2 = 0$, $\sigma_0^2 = 1$, $\rho_2 = 0.64$, $\pi_2 = 0.07$ *and* $\pi_0 = 1 - \pi_1 - \pi_2$. *The table shows the mean and the standard deviation of the estimated parameters over the* 100 *data sets using our model*

		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)
S4	True parameter	0.650	0.840	3.000	1.000
	Estimated values	0.669 (0.005)	0.850 (0.005)	3.021 (0.031)	1.058 (0.029)

For our method, we classify a call as correct (or incorrect), when a genuine (or spurious) signal is assigned an idr value smaller than an idr threshold. Correspondingly, for a call from individual replicates, Fisher's method or Stouffer's method, the same classification applies, when its corresponding q-value is smaller than the threshold. We compare the discriminative power of these methods by assessing the trade-off between the number of correct and incorrect calls made at various thresholds.

In an attempt to generate realistic simulations, we first estimated parameters from a ChIP-seq data set (described in Section 4 using the model in Section 2.2), then simulated the signals on a pair of replicates using the sampling model (3.1). We performed four simulations, S1, S2, S3 and S4, as follows, with simulation parameters in Table 1:

S1 This simulation was designed to demonstrate performance when the data are generated from the same copula mixture model we use for estimation. Data were simulated from the model (3.1) with $K = 2$, using the parameters estimated from the ChIP-seq data set considered below. The resulting data contained $\pi_1 = 65\%$ signals and $\pi_0 = 35\%$ noise.

S2 A simulation to assess performance of our method when the correlation between genuine signals is low. Data were simulated as in S1 ($K = 2$), except that $\rho_1 = 0.4$ and $\pi_1 = 0.3$.

S3 A simulation to assess performance of our method when only a small proportion of real but highly correlated signals are present. Data were simulated as in S1 ($K = 2$), except that $\pi_1 = 0.05$.

S4 Here simulation parameters were chosen to illustrate a scenario when repro-
ducible noise is present in addition to random noise and real signals. The goal
is to assess the sensitivity of our method to deviations from the assumption that
genuine signals are reproducible and noise is irreproducible. Data were simu-
lated from a three-component model (i.e., $K = 3$) using (3.1), where $\pi_2 = 7\%$
reproducible noise is added as the third component with $\rho_2 = 0.64$, $\mu_2 = 0$ and
$\sigma_2^2 = 1$, and the parameters for signals and random noise are as in S1, except
$\pi_0 = 28\%$ and $\mu_1 = 3$.

For each parameter set, we simulated 100 data sets, each of which consists of
two replicates with 10,000 signals on each replicate. In each simulation, we ran
the estimation procedure from 10 random initializations, and stopped the proce-
dure when the increment of log-likelihood is <0.01 in an iteration or the number
of iterations exceeds 100. All the simulations converge, when starting points are
close to the true parameters. The results that converge to the highest likelihood are
reported.

3.2.1. *Parameter estimation and calibration of IDR.*   In S1–S3, the parameters
estimated from our models are close to the true parameters (Table 1). 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.

The irreproducible discovery rate as a guide for the selection of the signals
needs to be well calibrated. To check the calibration of thresholds, we compare the
actual frequency of false calls, that is, empirical FDR, with the estimated IDR for
our method and with the q-values for other methods (Figure 4, left column).

As shown in Figure 4, the original significance scores and other combination
methods are overly conservative in their estimated FDR in all simulations, whereas
our method is reasonably well calibrated in S1, S2 and S3. When reproducible
noise is present (S4), our method slightly overestimates the proportion and the
correlation of the real signals, and underestimates the empirical FDR (Figure 4-
S4). This reflects that the data contains some artifacts that receive reproducible
high scores on their original measures and consequently receive relatively low idr
values. These artifacts are difficult to distinguish from genuine signals. We will
compare the discriminative power of all four methods in the next section.

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

In all simulations, our method consistently identifies more true signals than
the original significance score and the two p-value combination methods, at a
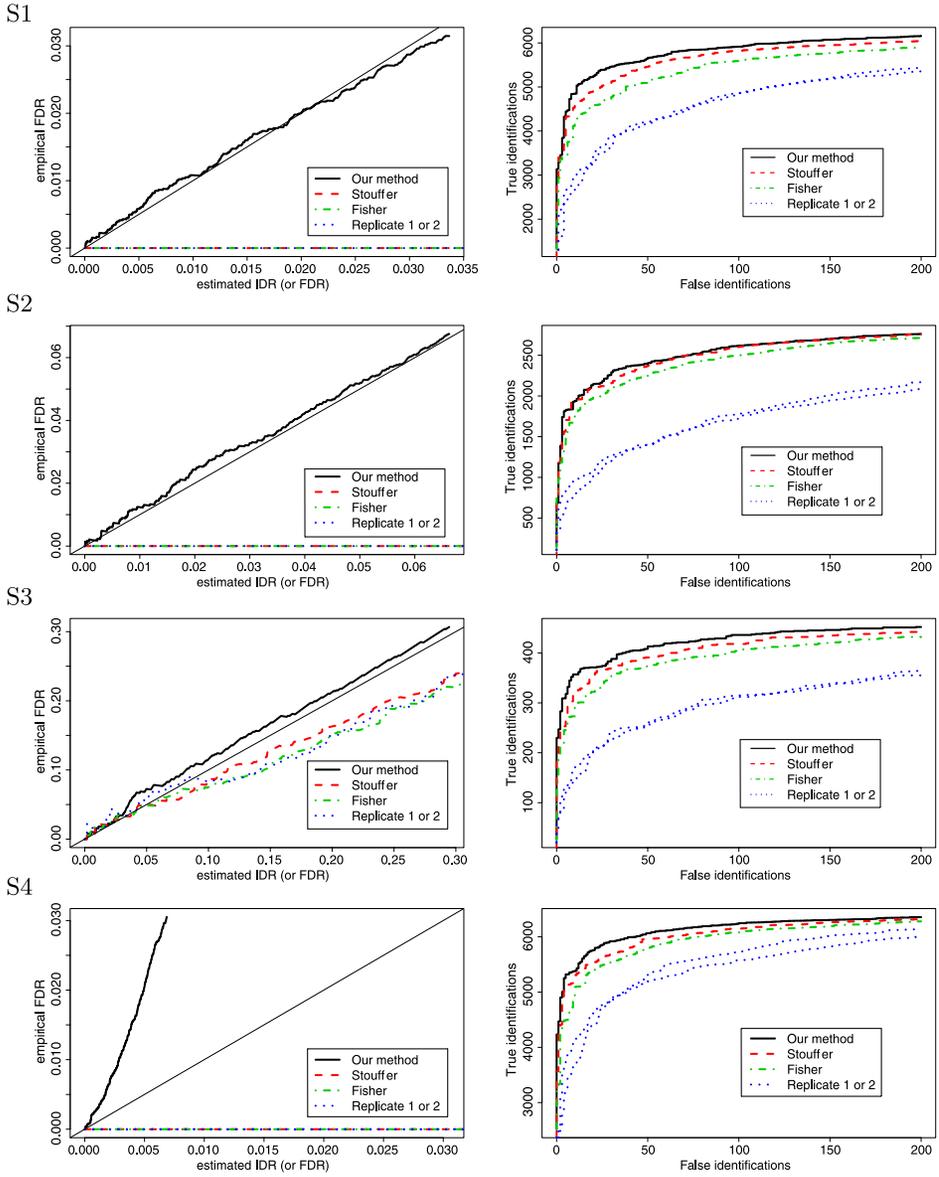
FIG. 4. *Calibration of IDR (Left) and comparison of discriminative power (Right) in simulation studies. Left: Estimated error rate (x-axis: IDR for our method and FDR for other methods) is compared with the actual frequency of false identifications (y-axis). Right: The number of correct and incorrect calls made at various thresholds in simulation studies. Incorrect calls: The number of spurious signals assigned idr values smaller than the thresholds (our method) or with q-values smaller than the cutoffs (other methods). Correct calls: The number of genuine signals assigned idr values smaller than the thresholds (our method) or with q-values smaller than the cutoffs (other methods).*

given number of false calls in the studied region. Even when reproducible arti-facts are present (S4) or only a small proportion of genuine signal is present (S3), our method still outperforms all methods compared here.

4. Applications on real data.

4.1. *Comparing the reproducibility of multiple peak callers for ChIP-seq exper-iments.*   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 reproducibility of multiple algorithms for iden-tifying protein-binding regions in ChIP-seq data (described below), selecting bind-ing regions using a uniform criterion for data from different sources (e.g., labs), and identifying experiment results in poor quality.

We now state the background of ChIP-Seq data in more detail and refer to Park (2009) for a recent review. A ChIP-seq experiment is a high-throughput assay to study protein binding sites on DNA sequences in a genome. In a typical ChIP-seq experiment, DNA regions that are specifically bound by the protein of interest are first enriched by immunoprecipitation, then the enriched DNA regions are se-quenced by high-throughput sequencing, which generates a genome-wide scan of tag counts that correspond to 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 local tag counts, based on either heuristics or some probabilistic models. The regions whose significance are above some prespecified threshold then are identified. To date, more than a dozen of the peak callers have been published. Some common measures of significance are fold enrichment, p-value or q-value [Storey (2003)].

Though these scores may reflect the relative strength of evidence for putative binding regions to be real, determination of a proper threshold is not straightfor-ward, especially for heuristic-based scores, where arbitrary judgment often has to be involved. In fact, this difficulty could also exist for probabilistic-based scores if the underlying probabilistic models are inadequate to capture the complexity of the data. Because tuning parameters for each data set are usually infeasible due to lack of ground truth, default thresholds are often used in practice, though they may not be the optimal choices for the data to be analyzed. Ideally, an objective performance assessment should reflect the behavior of peak callers instead of the effect of thresholds.

Here we use the binding regions identified at untuned thresholds in a CTCF ChIP-seq experiment (described below) to illustrate how our method is used for assessing and comparing the reproducibility of peak callers when tuning thresh-olds are unavailable, for setting a reproducibility-based threshold that is appliable to both heuristic and probabilistic-based significance scores, and for identifying results with low reproducibility. A detailed analysis on a comprehensive set of ENCODE data will appear elsewhere.

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 Bernstein Laboratory at the Broad Institute on human K526 cells. Peaks were identified in biology labs, using nine commonly used and publicly available peak callers, namely, Peakseq [Rozowsky et al. (2009)], MACS [Zhang et al. (2008)], SPP [Kharchenko, Tolstorukov and Park (2008)], Fseq [Boyle et al. (2008)], Hotspot [Thurman et al. (2011)], 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 (all peak callers except Hotspot) or more relaxed thresholds (Hotspot). Among them, Peakseq and SPP use q-value, MACS, Hotspot and SIS-SRS use p-value, and the rest use fold enrichment, as their significance measures. Only the outputs from peak callers were available for our analysis.

The peaks generated from different algorithms have substantially different peak widths. SPP and SISSRS generate peaks with fixed width of 100 bp and 40 bp, respectively; all other algorithms generate peaks with varying peak width (median $= 130$–760 bp). Because wider peaks are more likely to hit true binding sites by chance than shorter peaks, we normalized peak width by truncating the peaks wider than 40 bp down to intervals of 40 bp centered at the reported summits of peaks, so that reproducibility is compared on the same basis. The choice of 40 bp was made because the peak caller with the narrowest average peak width in our comparison reports peaks with a fixed width of 40 bp. Prior to applying our method, peaks on different replicates are paired as identifying the same binding region, if their coverage regions overlap (i.e., overlap ≥ 1 bp). Because peaks without matches do not have replicate measurements and are apparently irreproducible, here we elected to assess reproducibility of paired peaks in our analysis. Around 23–78% of peaks are retained for this analysis.

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 3, five peak callers (Peakseq, MACS, SPP, Fseq and Hotspot) show the characteristic transition from strong association to near independence [Figure 5(b)]. As described in Section 2.1, when heterogeneity of association is present, a high reproducibility translates to late occurrence 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 data set are Peakseq, MACS and SPP (Figure 5). For the other four peak callers (Erange, Cisgenome, Quest and SIS-SRS), the curves display a less clear transition and report substantially fewer (reproducible) peaks. This indicates that the default thresholds for these peak callers are likely to be too stringent to reach the breakdown of consistency, and that the reported peaks have relatively low reproducibility across replicates. This conclusion was confirmed later by biological verification (see Section 4.2.3).
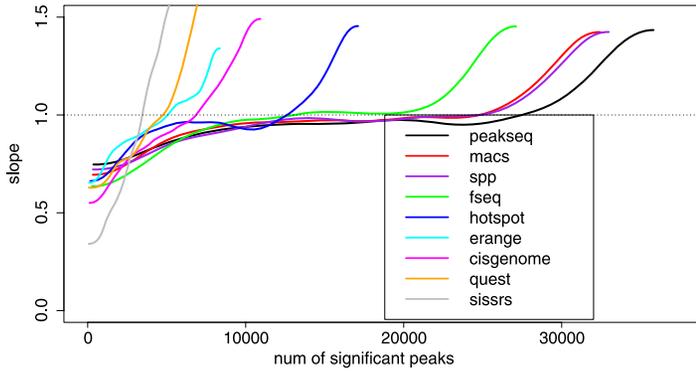
FIG. 5.   *The change of correspondence* (Ψ'_n) *along the decreasing order of significance, plotted for* 9 *peak callers on a CTCF ChIP-seq experiment from ENCODE. X-axis*: *The rank list of peaks identified on a replicate. Y-axis*: Ψ'_n.

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 one-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 pattern of transition in the correspondence profiles (Figure 5).

To select binding sites, we rank putative peaks by the values of local idr and compute the irreproducible discovery rate (IDR) for peaks selected at various local idr cutoffs using (2.10), as described in Section 2.3. We illustrate the IDR as a function of the numbers of top peaks (ranked by local idr) for all peak callers in Figure 6. For a given IDR level, one can determine the number of peaks to be called from this plot, regardless of what type of scores are used to measure

Parameters estimated from the copula mixture model and the one-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

the significance of peaks. For example, at 5% IDR, the top 27,500 peaks with the smallest local idr can be called using MACS. Using the same reproducibility criterion, peaks can be selected for other peak callers similarly.

We also compare the overall reproducibility of different peak callers using Figure 6. For example, while Peakseq, MACS and SPP on average have about 3% irreproducible peaks when selecting the top 25,000 peaks, most of the other peak callers have already reached a much higher IDR before identifying the top 10,000 peaks. According to the number of peaks identified before reaching 5% IDR, the three most reproducible peak callers on this data set are Peakseq, MACS and SPP,
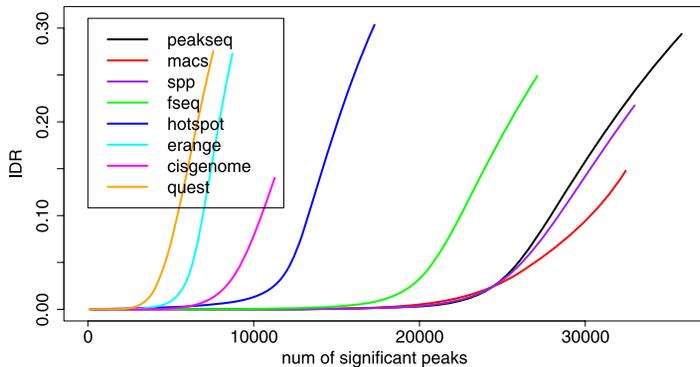

FIG. 6. *Irreproducible discovery rate (IDR) at different numbers of selected peaks, plotted at various idr cutoffs for eight peak callers on a CTCF Chip-seq experiment from ENCODE. Peaks are selected using local idr. X-axis: The rank list of peaks, ranked by local idr, Y-axis: Irreproducible discovery rate (IDR). SISSRS is not shown because its results are highly inconsistent and all peaks are grouped into a low correlation group.*

then followed by Fseq, then others. This result is consistent with the graphical comparison based on the correspondence profile (Figure 5).

4.2.3. *Evaluating the biological relevance of the reproducibility assessment.* To evaluate the biological relevance of our reproducibility assessment, we check the accuracy of peak identifications using external biological information. Because a complete list of true binding regions is not known for the examined data set, the accuracy of peak identifications is assessed using high-confidence binding motifs computationally predicted using sequence information [Kheradpour et al. (2007)], which is a commonly used device in this setting [e.g., Zhang et al. (2008); Kharchenko, Tolstorukov and Park (2008), among many others]. Though high-confidence motifs are not required to be bound and true binding sites are not required to exhibit a motif signature, the high-confidence motif instances are assumed, standard in this context, to contain a representative subset of true binding regions and are expected to have a relatively high occurrence in high-scored ChIP-seq peaks [Kharchenko, Tolstorukov and Park (2008)]. We selected ChIP-seq peaks reported by each peak caller at various IDR thresholds, and examined the number of high-confidence motifs (FDR ≤ 0.1 at the PWM threshold of p-value $= \frac{1}{4^{10}}$) that coincide with the reported ChIP-seq peaks (defined as overlap ≥ 1 bp) (Figure 7).

For the peak callers whose reported peaks fit the two-component model (i.e., Peakseq, MACS, SPP, Fseq and Hotspot), we marked the number of ChIP-seq peaks selected at IDR = 5%. For these algorithms, the motif occurrences first increase with the increase of reported ChIP-seq peaks, then plateau before reaching the default thresholds (Figure 7). The mark of 5% IDR approximately corresponds
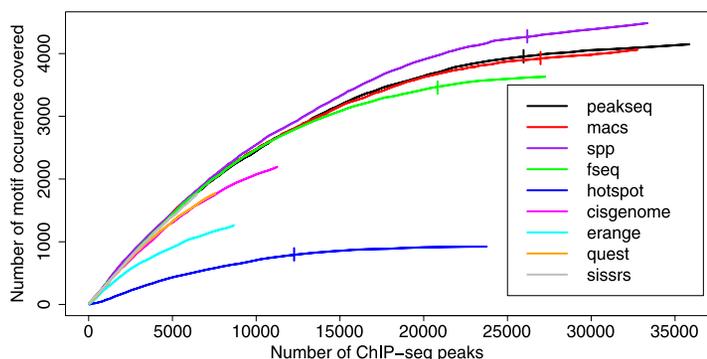

FIG. 7. *The coverage of high-confidence CTCF motif at different numbers of selected ChIP-seq peaks, plotted at various idr cutoffs for nine peak callers on a CTCF Chip-seq experiment from ENCODE. The bars on the curves of Peakseq, MACS, SPP, Fseq and Hotspot show the number of peaks selected at IDR = 0.05. No selection is made for the rest of the peak callers because model selection favors the one-component model for peaks identified by these callers.*

to the occurrences of the plateau, with few additional motif occurrences if more ChIP-seq peaks are called.

On the other hand, for the peak callers (Erange, Cisgenome, Quest and SIS-SRS) whose peaks fit the one-component model, the motif occurrence still shows an increasing trend at default thresholds. This confirms the observation from correspondence curves (Figure 5) that the default thresholds for these peak callers are likely to be overly stringent for this data set.

Overall, the results of this analysis agree with the assessment from our reproducibility comparison: the three peak calling results with the highest reproducibility (SPP, Peakseq and MACS) in Figure 6 show the highest rates of motif occurrence among all algorithms of comparison; the ones that are reported to be less reproducible do show lower rates of motif occurrence. 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 results in high-throughput experiments and setting selection thresholds using a reproducibility criterion. Using simulated and real data, we have illustrated the potential of our method for providing reproducibility assessment that is not confounded with prespecified threshold choices, determining biologically relevant thresholds, improving the accuracy of signal identification, and identifying suboptimal results.

As no assumption is made on the scale of the scores, the proposed method is applicable for any scoring system that produces continuous ranking to reflect the relative ordering of the signals. It provides a principled way to select signals that are scored on heuristic measures, and complements the thresholds determined on individual replicates. Moreover, because consistency between replicates is an internal standard that is independent of the scoring schemes and comparable across data sets, the proposed reproducibility criterion is suited for setting uniform standards for selecting signals for data from multiple sources, such as consortium studies. Because our measure of consistency is not confounded by platform-dependent thresholds, inter-platform consistency can be assessed easily.

Of course, reproducibility is only a necessary but not a sufficient condition to accuracy. If replicates are generated in the presence of a systematic bias that introduces false association, the threshold derived from this procedure may underestimate the empirical false discovery rate. Though the thresholds determined by our method show reasonable biological relevance in the data examined here and many other ENCODE ChIP-seq experiments (to appear in another manuscript), we emphasize that some cares are necessary to ensure that the replicates maintain the level of independence that they should.

We also note that the reproducibility of outputs from a data-analytical method (e.g., a peak caller) on replicate samples reflects the combined properties of the method and the samples. As the behavior of the data-analytical method may vary across different samples, the reproducibility assessment in our example should be

interpreted as being specific to the studied data set, instead of a general conclusion. A detailed comparison of the performance of peak callers has been evaluated on a comprehensive set of data and will appear elsewhere.

The algorithm to implement the estimation strategy outlined in Section 2.2.3 is provided as supplemental material to this article. An R package is downloadable at the following website: http://cran.r-project.org/web/packages/idr/index.html.

SUPPLEMENTARY MATERIAL

Supplementary materials for Measuring reproducibility of high-throughput experiments (DOI: 10.1214/11-AOAS466SUPP; .pdf). This supplement consists of four parts. Part 1 describes the algorithm for estimating parameters in our copula mixture model. Part 2 provides a theoretical justification for the efficiency of our estimator for the proposed copula mixture model when n is large. Part 3 derives the properties of the correspondence curves in Section 2.1.1. Part 4 provides an extension of our model to the case with multiple replicates.

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BLEST, D. C. (2000). Rank correlation—an alternative measure. *Aust. N. Z. J. Stat.* **42** 101–111. MR1747465

BOULESTEIX, A. L. and SLAWSKI, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics* **10** 556–568.

BOYLE, A. P., GUINNEY, J., CRAWFORD, G. E. and FUREY, T. S. (2008). F-Seq: A feature density estimator for high-throughput sequence tags. *Bioinformatics* **24** 2537–2538.

DA COSTA, J. P. and SOARES, C. (2005). A weighted rank measure of correlation. *Aust. N. Z. J. Stat.* **47** 515–529. MR2235420

DEHEUVELS, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.* (5) **65** 274–292. MR0573609

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

EFRON, B. (2004). Local false discovery rate. Technical report, Dept. Statistics, Stanford Univ.

ENCODE PROJECT CONSORTIUM (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306** 636–640.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Oliver & Boyd, Edinburgh.

FISHER, N. I. and SWITZER, P. (1985). Chi-plots for assessing dependence. *Biometrika* **72** 253–265. MR0801767

FISHER, N. I. and SWITZER, P. (2001). Graphical assessment of dependence: Is a picture worth 100 tests? *Amer. Statist.* **55** 233–239. MR1963399

GENEST, C. and BOIES, J.-C. (2003). Detecting dependence with Kendall plots. *Amer. Statist.* **57** 275–284. MR2016261

GENEST, C., GHOUDI, K. and RIVEST, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82** 543–552. MR1366280

GENEST, C. and PLANTE, J.-F. (2003). On Blest's measure of rank correlation. *Canad. J. Statist.* **31** 35–52. MR1985503

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

HU, L. (2006). Dependence patterns across financial markets: A mixed copula approach. *Applied Financial Economics* **16** 717–729.

JI, H., JIANG, H., MA, W., JOHNSON, D. S., MYERS, R. M. and WONG, W. H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology* **26** 1293–1300.

JOE, H. (1997). *Multivariate Models and Dependence Concepts. Monogr. Statist. Appl. Probab.* **73**. Chapman & Hall, London. MR1462613

JOTHI, R., CUDDAPAH, S., BARSKI, A., CUI, K. and ZHAO, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res.* **36** 5221–5231.

KALLENBERG, W. C. M. and LEDWINA, T. (1999). Data-driven rank tests for independence. *J. Amer. Statist. Assoc.* **94** 285–301. MR1689233

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

KHERADPOUR, P., STARK, A., ROY, S. and KELLIS, M. (2007). Reliable prediction of regulator targets using 12 drosophila genomes. *Genome Res.* **17** 1919–1931.

KUO, W., LIU, F., TRIMARCHI, J., PUNZO, C., LOMBARDI, M., SARANG, J., WHIPPLE, M. E. et al. (2006). A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature Biotechnology* **24** 832–840.

LEHMANN, E. L. (2006). *Nonparametrics*: *Statistical Methods Based on Ranks*, 2nd ed. Springer, New York.

LI, Q., BROWN, J. B., HUANG, H. and BICKEL, P. J. (2011). Supplement to "Measuring reproducibility of high-throughput experiments." DOI:10.1214/11-AOAS466SUPP.

MAQC CONSORTIUM (2006). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24** 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** 318–324.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. and WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5** 621–628.

NELSON, R. B. (1999). *An Introduction to Copulas*, 2nd ed. Springer, New York.

OAKES, D. (1994). Multivariate survival distributions. *J. Nonparametr. Stat.* **3** 343–354. MR1291555

PARK, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10** 669–680.

ROZOWSKY, J., EUSKIRCHEN, G., AUERBACH, R. K., ZHANG, Z. D., GIBSON, T., BJORNSON, R., CARRIERO, N., SNYDER, M. and GERSTEIN, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology* **27** 66–75.

SKLAR, M. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8** 229–231. MR0125600

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the *q*-value. *Ann. Statist.* **31** 2013–2035. MR2036398

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS, J. (1949). *The American Soldier: Vol. 1. Adjustment During Army Life.* Princeton Univ. Press, Princeton, NJ.

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

THURMAN, R., HAWRYLYCZ, M., KUEHN, S., HAUGEN, E. and STAMATOYANNOPOULOS, S. (2011). Hotspot: A scan statistic for identifying enriched regions of short-read sequence tags. Unpublished manuscript, Univ. Washington.

VALOUEV, A., JOHNSON, D. S., SUNDQUIST, A., MEDINA, C., ANTON, E., BATZOGLOU, S., MYERS, R. M. and SIDOW, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nature Methods* **5** 829–834.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSS-BAUM, C., MYERS, R. M., BROWN, M., LI, W. and LIU, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology* **9** R137.

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