Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Two-sample test with g-modeling and its applications

Jingyi Zhai | Hui Jiang

Department of Biostatistics, University of Michigan, MI, USA

Correspondence
Hui Jiang, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109. Email: jianghui@umich.edu

Summary

Many real data analyses involve two-sample comparisons in location or in distribution. Most existing methods focus on problems where observations are independently and identically distributed in each group. However, in some applications the observed data are not identically distributed but associated with some unobserved parameters which are identically distributed. To address this challenge, we propose a novel two-sample testing procedure as a combination of the g-modeling density estimation introduced by Efron and the two-sample Kolmogorov-Smirnov test. We also propose efficient bootstrap algorithms to estimate the statistical significance for such tests. We demonstrate the utility of the proposed approach with two biostatistical applications: the analysis of surgical nodes data with binomial models and differential expression analysis of single-cell RNA sequencing (scRNA-seq) data with zero-inflated Poisson model.

KEYWORDS:
Two-sample test, g-modeling; Bootstrap, Differential expression analysis, Single-cell RNA-seq, Zero-inflated Poisson.

1 | INTRODUCTION

Two-sample comparison occurs frequently in statistical analysis, for example, the testing of the drug effect between the control and the treatment groups in clinical trials. Relatively simple comparisons are often conducted to detect the difference in location between two groups where the parametric two-sample t-test or the non-parametric Wilcoxon rank-sum test is widely used. Comparing two samples in distribution is more challenging than comparing them in location. For more complicated and noisy data, the two-sample Kolmogorov-Smirnov (K-S) test is often used to detect the difference between two unknown distributions by comparing the empirical distributions of two groups of observations.[1] All the above tests and most other widely used statistical tests assume that the observations in each group are independently and identically distributed (i.i.d.). Unfortunately, such assumption may be violated in complex real-world problems and the problem becomes harder when the observations are no longer identically distributed. For instance, we may have two groups of independent samples where each observation follows the same type of distribution but with different underlying parameters such as different means. If we assume that these unknown parameters follow certain distributions, the objective of the two-sample comparison becomes the comparison of the distribution of the underlying parameters in the two groups. Specifically, consider the situation where there are two groups of observations $X_1, \ldots, X_{N_X}$ and $Y_1, \ldots, Y_{N_Y}$, where

$$X_i \overset{ind}{\sim} p_i(X_i|\Theta_i), \quad Y_k \overset{ind}{\sim} p_k(Y_k|\Lambda_k), \quad i = 1, \ldots, N_X, \quad k = 1, \ldots, N_Y, \\ \text{and} \qquad \Theta_i \overset{ind}{\sim} G, \quad \Lambda_k \overset{ind}{\sim} H, \tag{1}$$

where Θ_i's and Λ_k's are two groups of unknown quantities drawn from two unknown distributions G and H, respectively, and p_i's and p_k's are some known parametric densities (or probability mass functions for discrete data). Our interest is in testing whether G equals H. Under this model, X_i's (and Y_k's) are independently but non-identically distributed, e.g., when X_i is binomial with parameters n_i and Θ_i, where n_i's are fixed but different quantities for different i's. In such case, the standard two-sample K-S test is no longer applicable.

Since we only observe X_i's and Y_k's but not Θ_i's and Λ_k's, we cannot compare the two unknown distributions G and H directly. The related one-sample problem, i.e., the problem of estimating G from X_i's (and similarly to estimate H from Y_k's), a.k.a. empirical Bayes deconvolution,[3] has been well-studied with many methods developed,[4–11] but largely they suffer from intensive computation and slow convergence.[12] Recently, a g-modeling approach was proposed by Efron for the efficient estimation of G (or H) in such case.[2]

To address the above two-sample comparison problem, we propose a novel two-sample testing procedure combining the g-modelling method for density estimation and the two-sample Kolmogorov-Smirnov (K-S) test statistic to detect for differences in distribution between two samples. We also develop efficient bootstrap algorithms to estimate the statistical significance for such tests. Our approach can be applied on a wide range of data types. We apply our approach on simulated data from a surgical nodes dataset under two different scenarios. In this application, the numbers of malignant satellite from patients are assumed to follow binomial distributions. Our test is shown to have high power to detect small difference between the two groups. Moreover, we also applied our test to the differential expression (DE) analysis on a real scRNA-seq dataset. Different from the original purpose of using g-modeling to model the test statistics across all the genes to controlling for multiple testing and false discovery rate,[2] here we model the read counts from each gene across all the samples to detect DE genes individually. Comparing with other existing DE methods, our test can detect more DE genes and has higher accuracy.

The rest of the paper is organized as follows: In Section 2, we introduce the notations and briefly review the g-modelling method for density estimation. We then introduce our proposed approach for two-sample tests in distribution. Section 3 presents two applications: the analysis of surgical nodes data with binomial models and DE analysis of scRNA-seq data with zero-inflated Poisson model. Section 4 concludes the paper with a discussion.

2 | METHODS

Since we build our approach for two-sample comparison based on the g-modeling method, we first review it briefly in the section below.

2.1 | One-sample density estimation with g-modeling

Starting with the one-sample density estimation problem based on the observations $X_1, \dots, X_N$, we follow the same setting as the g-modelling method,[2] where the sample space of Θ is discretized as $\boldsymbol{\tau} = (\theta_1, \dots, \theta_m)$ for computational convenience. The g-modelling framework further assumes that Θ follows a semi-parametric exponential family distribution as follows:

$$Pr(\Theta = \theta_j) = g_j(\boldsymbol{\alpha}) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha} - \phi(\boldsymbol{\alpha})\}, \quad j = 1, \dots, m,$$

where $\boldsymbol{\alpha}$ is a p-dimensional vector of parameters, $\boldsymbol{Q}$ is a fixed and known $m \times p$ matrix taken as the design matrix from natural spline basis,[2] $\boldsymbol{Q}_j$ is the j-th row of $\boldsymbol{Q}$ (as a p-dimensional column vector), and the normalization term $\phi(\boldsymbol{\alpha})$ is $\phi(\boldsymbol{\alpha}) = \log \sum_{j=1}^m \exp(\boldsymbol{Q}_j^T \boldsymbol{\alpha})$. Conditional on Θ_i, the observed X_i follows a known parametric distribution as $X_i \overset{ind}{\sim} p_i(X_i|\Theta_i)$, for $i = 1, \dots, N$, and we define $p_{ij} = p_i(X_i = x_i|\Theta_i = \theta_j)$. Then the marginal probability of X_i and log-likelihood of the observed data can be computed as $Pr(X_i = x_i) = f_i(\boldsymbol{\alpha}) = \sum_{j=1}^m p_{ij} g_j(\boldsymbol{\alpha})$, and $l_i(\boldsymbol{\alpha}) = \log f_i(\boldsymbol{\alpha})$, respectively. Here we assume discrete X_i (or discretized X_i if it was continuous). In order to improve the accuracy for estimation, the log-likelihood is regularized with a ℓ_2 penalty term. Hence, the objective function for maximum likelihood estimation is $m(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - s(\boldsymbol{\alpha})$, where $s(\boldsymbol{\alpha}) = c_0 \|\boldsymbol{\alpha}\|$ with c_0 being a tuning parameter which we take as 1. Now $\boldsymbol{\alpha}$ can be estimated by maximizing the above penalized log-likelihood. We denote the maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}}$ and obtain the estimated $g_j(\boldsymbol{\alpha})$ by $g_j(\hat{\boldsymbol{\alpha}})$ for $j = 1, \dots, m$.

In practice, we find that discretized Θ may introduce numerical problems, especially when we are dealing with scRNA-seq data where the read counts are sparse and have a large range. For instance, it could happen that $p_{ij} = 0$ for all $j = 1, \dots, m$, which leads infinite log-likelihood. Therefore, we make an adjustment in the calculation of p_{ij} to avoid such problem. For each

x_i, we estimate its MLE $\hat{\theta}_i^*$, and calculate p_{ij} for the θ_j that is the closest to $\hat{\theta}_i^*$ as $p_{ij} = p_i(X_i = x_i|\Theta_i = \hat{\theta}_i^*)$. For all other j, we still calculate p_{ij} as $p_{ij} = p_i(X_i = x_i|\Theta_i = \theta_j)$.

2.2 | Two-sample comparison based on g-modeling

Here we propose a two-sample test built on the above one-sample density estimation procedure using g-modelling. With the notations in (1), assume Θ_i and Λ_k have the same discretized sample space $\tau = (\theta_1, \dots, \theta_m)$, $g(\boldsymbol{\alpha}_X)$ and $h(\boldsymbol{\alpha}_Y)$ are the semi-parametric distributions for G and H respectively, and have the following form

$$\begin{aligned} Pr(\Theta_i = \theta_j) &= g_j(\boldsymbol{\alpha}_X) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_X - \phi(\boldsymbol{\alpha}_X)\}, \text{ for } j = 1, \dots, m, \\ Pr(\Lambda_k = \theta_j) &= h_j(\boldsymbol{\alpha}_Y) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_Y - \phi(\boldsymbol{\alpha}_Y)\}, \text{ for } j = 1, \dots, m, \end{aligned} \tag{2}$$

where $\boldsymbol{Q}$ is a fixed and known $m \times p$ structure matrix for both g and h, and $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_Y$ are parameters that can be estimated with g-modeling using the data from the two groups, respectively. To test for the difference between the two distributions G and H against the null hypothesis $H_0 : G = H$, we use the two-sample K-S test statistic which can be calculated as

$$T = \max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, \quad j = 1, \dots, m, \tag{3}$$

where $\hat{G}_j$ and $\hat{H}_j$ are the values of estimated cumulative distribution functions (CDFs) of G and H evaluated at θ_j, respectively.

Different from the generalized linear mixed model introduced originally for two-sample comparison when changes in location between two distributions are of concern,[2] our two-sample test can detect changes not only in location but also in distribution.

2.3 | Simple bootstrap procedure for p-value estimation

To estimate the statistical significance for the test statistic T defined in (3), we can use a simple parametric bootstrap procedure to directly simulate the null distribution of T. Under the null hypothesis, G and H are identical. Therefore, we first pool the two groups together and employ the g-modeling approach to obtain a pooled density estimate $\hat{g}(\hat{\boldsymbol{\alpha}}_p)$. Then, for the b-th bootstrap iteration, $b = 1, \dots, B$, we take the following steps:

1. Sample $\Theta_i^{(b)}, i = 1, \dots, N_X$, and $\Lambda_k^{(b)}, k = 1, \dots, N_Y$, with respect to $\hat{g}(\hat{\boldsymbol{\alpha}}_p)$.
2. Sample $X_i^{(b)}$ from $p_i(X_i|\Theta_i^{(b)}), i = 1, \dots, N_X$, and $Y_k^{(b)}$ from $p_k(Y_k|\Lambda_k^{(b)}), k = 1, \dots, N_Y$.
3. Estimate $\hat{G}^{(b)}$ from $X_1^{(b)}, \dots, X_{N_X}^{(b)}$ and $\hat{H}^{(b)}$ from $Y_1^{(b)}, \dots, Y_{N_Y}^{(b)}$ using g-modeling.
4. Calculate $T^{(b)} = \max_j |\hat{G}_j^{(b)} - \hat{H}_j^{(b)}|, j = 1, \dots, m$.

Finally, we estimate the p-value as $\hat{p} = (\sum_{b=1}^{B} 1_{T^{(b)} \geq T} + 1)/(B+1)$, where we add one to both the numerator and the denominator to avoid a p-value of zero.

Since MLE problems need to be solved for each bootstrap iteration, the above simple bootstrap procedure may be computationally intensive, especially when there are a large number of tests to be performed. For instance, in DE analysis one often needs to test for thousands of genes. In order to reduce the computational load, we employ an early stopping rule in our experiments.[13] Specifically, after the b-th bootstrap iteration, we calculate $\hat{p}_b^* = \sum_{l=1}^{b} 1_{T^{(l)} \geq T}/b$. If $\hat{p}_b^* > (a/b + c)/(1 + c)$ where a and c are some constants, then we stop the bootstrap procedure and output $\hat{p}^s = \hat{p}_b^*$. Otherwise, the bootstrap procedure will continue until $b = B$ and outputs $\hat{p}^s = \hat{p}_B^*$, where $\hat{p}^s$ is the final p-value estimate for our two-sample test. Following recommendation,[13] we take $c = (1 + \delta) \times p_0/(1 - p_0)$, where p_0 is the p-value cutoff, and a and δ are parameters of choice. In our application, the p-value cutoff is chosen as $p_0 = 0.01$, and we set $a = 4$, $\delta = 0.4$. Thus, $c = 0.0141$.

2.4 | Accelerated bootstrap procedure based on asymptotic distribution of the test statistic

From our experiments, we find that the simple bootstrap procedure described in Section 2.3 provides accurate p-value estimates, but at the price of intensive computation, due to the need to estimate $\hat{G}^{(b)}$ and $\hat{H}^{(b)}$ using g-modeling in each bootstrap iteration. In this section, we propose an accelerated bootstrap procedure based on approximating the null distribution of the test statistic using large sample theory.

Suppose X and Y are sampled from exponential family distributions such as normal, Poisson or binomial distributions. With g-modeling, we assume that G and H as semi-parametric exponential distributions as defined in (2). After obtaining the MLE of $\boldsymbol{\alpha}$ denoted as $\hat{\boldsymbol{\alpha}}$, we denote the estimated PDF and CDF of Θ as $g(\hat{\boldsymbol{\alpha}})$ and $G(\hat{\boldsymbol{\alpha}})$. Since both $G(\hat{\boldsymbol{\alpha}})$ and $g(\hat{\boldsymbol{\alpha}})$ are evaluated on the grid $\tau = (\theta_1, \dots, \theta_m)$, w.l.o.g., assuming it is an equally-spaced grid for simplicity, we have $G(\hat{\boldsymbol{\alpha}}) = \boldsymbol{A}g(\hat{\boldsymbol{\alpha}})$, where

$$\boldsymbol{A} = \begin{bmatrix} a & 0 & \dots & 0 \\ a & a & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \dots & a \end{bmatrix},$$

where a is the grid size.

The asymptotic distribution of $G(\hat{\boldsymbol{\alpha}})$ is[2]:

$$G(\hat{\boldsymbol{\alpha}}) - G(\boldsymbol{\alpha}_0) \doteq N[\text{Bias}\{G(\boldsymbol{\alpha}_0)\}, \text{Cov}\{G(\boldsymbol{\alpha}_0)\}],$$

where $\boldsymbol{\alpha}_0$ is the true value of $\boldsymbol{\alpha}$ and

$$\text{Bias}\{G(\boldsymbol{\alpha}_0)\} = \boldsymbol{A}D(\boldsymbol{\alpha}_0)\boldsymbol{Q}\text{Bias}(\boldsymbol{\alpha}_0), \quad \text{Cov}\{G(\boldsymbol{\alpha}_0)\} = \boldsymbol{A}D(\boldsymbol{\alpha}_0)\boldsymbol{Q}\text{Cov}(\boldsymbol{\alpha}_0)\boldsymbol{Q}^T D(\boldsymbol{\alpha}_0)\boldsymbol{A}^T,$$

where

$$D(\boldsymbol{\alpha}_0) = \text{Diag}\{g(\boldsymbol{\alpha}_0)\} - g(\boldsymbol{\alpha}_0)g(\boldsymbol{\alpha}_0)^T, \quad \text{Bias}(\boldsymbol{\alpha}_0) = -\{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1}\dot{s}(\boldsymbol{\alpha}_0),$$
$$\text{Cov}(\boldsymbol{\alpha}_0) = \{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1}\mathcal{I}(\boldsymbol{\alpha}_0)\{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1},$$

and

$$s(\boldsymbol{\alpha}_0) = c_0||\boldsymbol{\alpha}_0||, \quad \dot{s}(\boldsymbol{\alpha}_0) = c_0\frac{\boldsymbol{\alpha}_0}{||\boldsymbol{\alpha}_0||}, \quad \ddot{s}(\boldsymbol{\alpha}_0) = c_0\frac{c_0}{||\boldsymbol{\alpha}_0||}(\boldsymbol{I}_p - \frac{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T}{||\boldsymbol{\alpha}_0||^2}).$$

Here $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and $\text{Diag}\{g(\boldsymbol{\alpha}_0)\}$ is a diagonal matrix with $g(\boldsymbol{\alpha}_0)$ as the diagonal components. $\mathcal{I}(\boldsymbol{\alpha}_0)$ is the corresponding Fisher information matrix calculated as

$$\mathcal{I}(\boldsymbol{\alpha}) = \boldsymbol{Q}^T[W_i(\boldsymbol{\alpha})W_i(\boldsymbol{\alpha})^T + W_i(\boldsymbol{\alpha})g(\boldsymbol{\alpha})^T + g(\boldsymbol{\alpha})W_i(\boldsymbol{\alpha})^T - \text{Diag}\{W_i(\boldsymbol{\alpha})\}]\boldsymbol{Q},$$

where $W_i(\boldsymbol{\alpha})$ is an m-vector and its j-th element is defined as $w_{ij}(\boldsymbol{\alpha}) = g_j(\boldsymbol{\alpha})\{p_{ij}/f_j(\boldsymbol{\alpha}) - 1\}$, and $\text{Diag}\{W_i(\boldsymbol{\alpha})\}$ is an $m \times m$ diagonal matrix with $W_i(\boldsymbol{\alpha})$ on the diagonal. In practice, the true value $\boldsymbol{\alpha}_0$ is unknown, so we replace $\boldsymbol{\alpha}_0$ with $\hat{\boldsymbol{\alpha}}$ in above formulas to estimate the bias and covariance of G.

Now we move on to consider our two-sample test. As the null hypothesis defined where $G = H$, the true values $\boldsymbol{\alpha}_{X0} = \boldsymbol{\alpha}_{Y0}$. Hence, we can pool the two groups X and Y together for estimating the parameter vector similarly as the simple bootstrap procedure and obtain the estimate $\hat{\boldsymbol{\alpha}}_p$. Then we use $\hat{\boldsymbol{\alpha}}_p$ in the calculation of the biases and the covariance matrices. Since the two groups are independent, the asymptotic null distribution of $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ is

$$G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \doteq N[\text{Bias}\{G(\hat{\boldsymbol{\alpha}}_p)\} - \text{Bias}\{H(\hat{\boldsymbol{\alpha}}_p)\}, \text{Cov}\{G(\hat{\boldsymbol{\alpha}}_p)\} + \text{Cov}\{H(\hat{\boldsymbol{\alpha}}_p)\}]. \tag{4}$$

Our two-sample K-S test statistics T defined in (3) is the maximum absolute difference between $G(\hat{\boldsymbol{\alpha}}_p)$ and $H(\hat{\boldsymbol{\alpha}}_p)$. Although it is difficult to derive the null distribution of T analytically, we can use parametric bootstrap to simulate its null distribution based on (4) and then estimate the p-value of the test. By doing so, since we only need to sample from a multivariate normal distribution and therefore have avoided the estimation of $\hat{G}^{(b)}$ and $\hat{H}^{(b)}$ using g-modeling in each bootstrap iteration as in the simple bootstrap procedure, the computational burden is greatly reduced. See Section S1.1 in the supplementary materials for more details.

The accelerated bootstrap procedure based on the asymptotic null distribution derived in this section is computationally more efficient than the simple bootstrap procedure described in Section 2.3, as we directly obtain bootstrap samples from the multivariate normal distribution. However, it only provides an approximated p-value estimate which requires relatively large sample sizes in both groups for the approximation to be accurate, especially in settings where the convergence of the asymptotic distribution may be slow.

3 | APPLICATIONS

3.1 | Malignant nodes analysis on intestinal surgery

We apply our method on a satellite nodes dataset of intestinal surgery in Gholami and others (2002).[14] There are 844 cancer patients who have removed satellite nodes for later testing. Each patient has a pair of $(n_i, X_i), i = 1, \dots, 844$, where n_i is the number of removed satellites and X_i is the number of malignant satellites found. Binomial distribution is assumed as $X_i \sim Bin(n_i, \Theta_i)$, where Θ_i denotes the probability of any one satellite being malignant for the patient i.

To extend the real dataset into a two-group comparison setting, we simulate data based on these 844 pairs in three cases. We add a binary group indicator C ($C = 0$ or $C = 1$) and randomly split data into two groups with equal size. Now we have

$$\Theta_i | C_i = 0 \sim G, \qquad \Theta_i | C_i = 1 \sim H.$$

G is estimated with all pairs with $C = 0$ and H is estimated with all pairs with $C = 1$.

In the first case – null hypothesis, we directly apply g-modeling to estimate G and H, and then perform our two-sample test with 99 bootstrap iterations. We repeat the random assignment procedure for $1,000$ times and obtain the histogram and Q-Q plot of empirical p-values (see Figure 1a and Figure 1b). Both the histogram and the Q-Q plot shows a uniform distribution the p-values which indicated insignificant difference in the distribution of the malignant probability of satellites between the two randomly assigned groups, which is expected.

In the second case – alternative hypothesis 1, we keep $\hat{G}$ as the final estimated density for the group with $C = 0$, and for the group with $C = 1$, we sample $\hat{\Theta}_i$ from $\hat{H}$ and implement the transformation $\Theta_i^* = (1 - w)\hat{\Theta}_i + wI(\hat{\Theta}_i > 0)$ where $0 < w < 1$ is a tuning parameter. Θ_i^* is more distinct from $\hat{\Theta}_i$ with larger w. Here Θ_i^* is always greater than Θ_i.

In the third case – alternative hypothesis 2, we also maintain $\hat{G}$ as the final estimated density for the group with $C = 0$, and for the group with $C = 1$, we sample $\hat{\Theta}_i$ from $\hat{H}$ and implement the transformation $\Theta_i^* = (1 - w)\hat{\Theta}_i + w\{0.5I(0 < \hat{\Theta}_i < 1) + I(\hat{\Theta}_i = 1)\}$ where $0 < w < 1$ is a tuning parameter. Θ_i^* is more distinct from $\hat{\Theta}_i$ with larger w. Θ_i^* can be different from Θ_i in two directions – either larger than or smaller than Θ_i.

In both alternative hypotheses, X_i^* is generated from $Bin(n_i, \Theta_i^*)$. Then we obtain the final estimated density for the group $C = 1$ from the generated pairs (n_i, X_i^*). Through 99 bootstrap iterations, we compute the empirical p-values. The data generation procedure is repeated 100 times with the same w, which ranges from 0 to 0.05 for the power plot in the alternative hypothesis 1, while $w \in (0, 0.1)$ for the power plot in the alternative hypothesis 2. From Figure 2a and Figure 2b, we can see that the power increases as w increases in both cases. However, the alternative case 1 has higher power than the alternative case 2 with the same w. In the alternative case 1, the change can only occur in one direction. However the transformation in the alternative case 2 involves two directions of change which reduces the final combined transformation. Therefore, with the same w, the transformation in the alternative case 1 is more dramatic than that in the alternative case 2.

3.2 | Differential expression analysis of single-cell RNA-seq data

Another application of our proposed two-sample test is the differential expression (DE) analysis of RNA sequencing (RNA-Seq) or the more recent single-cell RNA sequencing (scRNA-seq) data. Such data is often sparse, i.e. have lots of zero counts,[15–17] and the gene distribution of scRNA-seq data is complex since there is substantial heterogeneity among different cell samples.[18,19] Most importantly, since each measurement (read count for a given gene) from a sample depends on some sample-specific factors (e.g. library size) of that sample, the measurements across samples are not identically distributed.

Since the true expression level of a gene can vary across samples (e.g., cells in scRNA-seq), we model it as the underlying unknown parameters of interest (i.e., Θ_i and Λ_k) with g-modeling. The observed read counts (i.e., X_i and Y_k) are generated from these underlying parameters via a parametric distribution (e.g., zero-inflated Poisson) with the library size as a known covariate. Many existing DE methods assume that the gene expression level follows some parametric distribution (e.g., normal or gamma distributions) and are interested in detecting its locational changes between conditions. Our approach, however, use the semi-parametric g-modeling framework and can detect distributional changes between conditions.

In the literature, a number of methods have been introduced to detect DE genes from scRNA-seq data. Model-based analysis of single-cell transcriptomics (MAST) and single-cell differential expression (SCDE) fit two-stage models to handle inflated zero counts.[20,21] Nonparametric methods such as SigEMD, EMDomics and D3E address the multimodality issue in scRNA-seq data[22–24]. However, as pointed out by Jaakkola and others (2016) and Wang and others(2019),[25,26] none of these methods is able to handle inflated zero counts and multimodality issues simultaneously. We compare our method on a scRNA-seq dataset with seven other existing methods designed for differential expression analysis: (i) Model-based analysis of single-cell transcriptomics (MAST) models scRNA-seq data with a mixture of two components. One component describes the unobserved or dropout

measurements, and the other component explains the observed gene expression in cells.[20] A two-part generalized linear model is used to fit the data. MAST only considers the differences in location between the two groups. (ii) Single-cell differential expression (SCDE) is a three-step Bayesian approach, including data filtering, finding an error model and differential expression test.[21] (iii) Single-cell Differential Distributions (scDD) is a mixture modelling method based on a Bayesian framework to detect genes with expression changes in distributions between conditions.[27] MAST, SCDE and scDD are all designed specifically for scRNA-seq data. Four other methods designed for differential expression analysis in microarray or bulk RNA-seq data are also included in our comparison. (iv) Differential expression analysis for sequence count data (DESeq) is based on a negative binomial model with mean and variance linked through local regression.[28] (v) Linear models for microarray and RNA-Seq data (Limma-Voom) fit a gene-wise linear model on the gene expression values and applied modified t-statistical to test for differential expressed genes.[29] (vi) Reproducibility-optimized test statistic (ROTS) uses a modified t-statistic by maximizing the reproducibility of top-ranked features across group-preserving bootstrap samples.[30] (vii) A negative binomial generalized log-linear model is fitted with likelihood ratio test on read counts by edgeR to detect DE genes.[31]

Accounting for the excess number of zero counts in scRNA-seq data, we modify our test statisitcs and the associated two bootstrap procedures. Specifically, given a gene, we assume that the reads count (X) for that gene from each cell follow a Zero-Inflated Poisson distribution $ZIP(\lambda, \pi)$ where λ is the usual Poisson rate parameter and π is the excessive probability of zero counts with the following probability mass function

$$p(X = k|\pi, \lambda) = \begin{cases} \pi + (1-\pi)e^{\lambda}, & k = 0, \\ (1-\pi)\dfrac{\lambda^k e^{\lambda}}{k!}, & k > 0. \end{cases} \tag{5}$$

However, the Poisson expression rates and are different for different cells. These Poisson expression rates cannot be observed but assumed to follow a certain distribution which can be estimated from the reads counts with g-modeling density estimation. We denote $G(\boldsymbol{\alpha}_X)$ and $H(\boldsymbol{\alpha}_Y)$ as the distributions of Poisson expression rates in two groups. We define π_X and π_Y as the excessive probability of zeros in two groups. Then, the problem of DE analysis is to detect the difference between $G(\boldsymbol{\alpha}_X)$ and $H(\boldsymbol{\alpha}_Y)$ as well as the difference between π_X and π_Y. In such case, since π_X and π_Y also need to be estimated together with $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_Y$, we use a modified K-S test statistic defined as

$$T = \max\{\max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, |\hat{\pi}_X - \hat{\pi}_Y|\}. \tag{6}$$

Note that here we apply the g-modeling method for a purpose different from that in Efron's original paper. Instead of treating the test statistics across all the genes as observations,[2] here we model the read counts from a given gene across all the cells as observations.

For the modified test statistic T defined in (6), a similar simple parametric bootstrap procedure with slight modifications can be used. We first estimate $\hat{\pi}_p$ together with $\hat{\boldsymbol{\alpha}}_p$ by pooling two groups together. Then, in each bootstrap iteration, we simulate data with the zero-inflated Poisson distribution according to (5) and estimate $(\hat{G}^{(b)}, \hat{\pi}_X^{(b)})$ and $(\hat{H}^{(b)}, \hat{\pi}_Y^{(b)})$ using MLE from the simulated data, respectively, and calculate $T^{(b)}$ as $T^{(b)} = \max(\max_j |\hat{G}_j^{(b)} - \hat{H}_j^{(b)}|, |\hat{\pi}_X^{(b)} - \hat{\pi}_Y^{(b)}|)$.

The accelerated bootstrap procedure can also be adjusted accordingly. Specifically, we now have two parts in the modified test statistics, so we define an augmented vector of parameters as $\boldsymbol{\alpha}' = \begin{pmatrix} \pi \\ \boldsymbol{\alpha} \end{pmatrix}$. Given $\Theta_i = \theta_j$ and π, X_i follows a zero-inflated Poisson distribution $ZIP(\theta_j, \pi)$. We can obtain the joint asymptotic distribution of $\hat{\pi}_X - \hat{\pi}_Y$ and $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ in a similar manner as the general case in Section 2.4. Here we also pool the two groups of sample together to estimate the augmented vector of parameters as $\hat{\boldsymbol{\alpha}}'_p = \begin{pmatrix} \hat{\pi}_p \\ \hat{\boldsymbol{\alpha}}_p \end{pmatrix}$ for the asymptotic null distribution. Defining two transformation functions $C(\hat{\boldsymbol{\alpha}}') = \begin{bmatrix} \hat{\pi} \\ G(\hat{\boldsymbol{\alpha}}) \end{bmatrix}$ and $K(\boldsymbol{\alpha}') = \begin{bmatrix} \pi \\ g(\boldsymbol{\alpha}) \end{bmatrix}$, we can then obtain the joint asymptotic multivariate normal distribution of $\hat{\pi}_X - \hat{\pi}_Y$ and $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ as

$$\begin{bmatrix} \hat{\pi}_X - \hat{\pi}_Y \\ G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \end{bmatrix} \doteq N[\text{Bias}\{C(\hat{\boldsymbol{\alpha}}'_p)\} - \text{Bias}\{C(\hat{\boldsymbol{\alpha}}'_p)\}, \text{Cov}\{C(\hat{\boldsymbol{\alpha}}'_p)\} + \text{Cov}\{C(\hat{\boldsymbol{\alpha}}'_p)\}],$$

where

$$\text{Bias}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Bias}\{K(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\dot{\boldsymbol{K}}^T\text{Bias}(\hat{\boldsymbol{\alpha}}'),$$
$$\text{Cov}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Cov}\{K(\hat{\boldsymbol{\alpha}}')\}\boldsymbol{B}^T = \boldsymbol{B}\dot{\boldsymbol{K}}^T\text{Cov}(\hat{\boldsymbol{\alpha}}')\dot{\boldsymbol{K}}\boldsymbol{B}^T,$$

and

$$\dot{\boldsymbol{K}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \dot{g} & \\ 0 & & & \end{bmatrix}, \quad \dot{g} = \boldsymbol{Q}^T D, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \boldsymbol{A} & \\ 0 & & & \end{bmatrix}.$$

Here $\dot{\boldsymbol{K}}$ is the matrix of the first derivatives of K with respect to $\boldsymbol{\alpha}'$ when $\boldsymbol{\alpha}'$ takes the value $\hat{\boldsymbol{\alpha}}'_p$. $\dot{g}$ is the matrix of the first derivatives of g with respect to $\boldsymbol{\alpha}$ when we use $\hat{\boldsymbol{\alpha}}$ to approximate $\boldsymbol{\alpha}$. The bias and covariance of $\hat{\boldsymbol{\alpha}}'$ can be derived similarly as in the simple case. See Section S1.2 in the supplementary materials for more details.

To combine the above asymptotic null distribution with bootstrap for p-value estimation, we can generate bootstrap samples for the difference in π and the difference in cumulative density function jointly from using the above asymptotic multivariate normal distribution.

In practical settings such as the zero-inflated Poisson case, the data is skewed with an excess number of zero counts, and outliers are more likely to occur. We find that the asymptotic null distribution-based accelerated bootstrap procedure may fail occasionally when there are extreme outliers and consequently the estimated covariance matrix becomes singular. We use numerical remedies to stabilize the covariance matrix (e.g., adding a constant diagonal matrix to it) to handle such issues.

3.2.1 | Simulations with the zero-inflated Poisson model

We first evaluate our proposed approach with simulated data from the ZIP model. In particular, we simulate two samples as follows

$$\begin{gathered} X_i \sim ZIP(d_{X_i}\Theta_i, \pi_X), \quad Y_k \sim ZIP(d_{Y_k}\Lambda_k, \pi_Y), \quad i,k = 1, \dots N, \\ d_{X_i} \sim Unif(0.5, 1), \quad d_{Y_k} \sim Unif(0.5, 1), \quad \pi_X = \pi_Y = 0.5, \end{gathered} \tag{7}$$

where d_{X_i} and d_{Y_k} are known constants modeling the sequencing depths (or library sizes), Θ_i and Λ_k are unknown parameters modeling gene expression levels, and π_X and π_Y are unknown parameters modeling the excessive probabilities of zeros. X_i's and Y_k's are the simulated reads counts for a gene from two groups of cells.

Under the null hypothesis, we simulate both Θ_i and Λ_k from chi-square distribution with ten degrees of freedom, that is, $\Theta_i \sim \chi^2_{10}$ and $\Lambda_k \sim \chi^2_{10}$. Under the alternative hypothesis, we simulate Θ_i and Λ_j from two chi-square distributions with different degrees of freedom, respectively, $\Theta_i \sim \chi^2_{10}$ and $\Lambda_k \sim \chi^2_{10+\Delta}$, where we vary the effect size Δ from 1 to 5. We use the modified K-S statistic (6) to capture the difference between the distributions of Θ_i and Λ_k as well as the difference between π_X and π_Y.

We run 1000 simulations and use $B = 99$ bootstrap samples for each simulation to save computational cost. Figures 3a and 3b show the histogram and Q-Q plot of the p-values estimated from data simulated under the null hypothesis, which is uniformly distributed on $(0, 1)$ as expected. Figure 3c shows the estimated distributions of Θ and Λ against the true distribution χ^2_{10}. We see that there are some small but noticeable biases due to the semi-parametric g-modeling and penalized MLE at the tail and peak points, which is expected.[2] The estimated π_X and π_Y also seem reasonable as their histograms are centered around the true value 0.5 as Figures 3d and 3e show.

Figure 4 shows the statistical power of the simple and accelerated bootstrap procedures with varying differences in degrees of freedom between two chi-square distributions ($\Delta = 0$ simulated under the null hypothesis, and $\Delta = 1, \dots, 5$ simulated under the alternative hypothesis), and with varying sample sizes ($N = 100, 300, 500, 2000$) in each group. We can see that the power of both bootstrap procedures increases with larger differences between the two distributions. When $\Delta = 0$, both procedures have controlled type-I error rate at the predefined significance level $\alpha = 0.05$. As Δ increases, the accelerated bootstrap procedure has substantially lower power than the simple bootstrap procedure when the sample size is 100, but the power is similar for the two methods when the sample size is 300 or larger.

Additional simulations based on normal, binomial and Poisson distributions are provided in Section S2 in the supplementary materials. Furthermore, to compare the speed of the two bootstrap approaches, Figure 5 and Table S1 in the supplementary materials show the computational time of the two bootstrap procedures with varying sample sizes for normal-based model. We can see that the time increases linearly with sample size for both procedures. However, the accelerated bootstrap procedure is about 30 to 50 times faster than the simple bootstrap procedure. Hence, the accelerated bootstrap procedure is more suitable for datasets with more observations (e.g., $N > 300$ in each group) for being faster and reasonably powerful, while the simple bootstrap procedure is more suitable for datasets with smaller sample sizes for being more powerful and reasonably fast.

3.2.2 | Real data analysis

We run our proposed two-sample test on a real scRNA-seq dataset on human embryonic cells in early development.[32] We compare 81 cells in embryonic day 3 (E3) to 190 cells embryonic day 4 (E4), similarly as performed in.[33] There are dramatic changes between these two days, so this subset is suitable for running differential expression analysis. We select 2000 genes with the highest mean read counts across all cell lines and pick the genes with higher standard deviation when there are ties. We use the total read counts for each cell line as the constants d_{X_i} and d_{Y_k} adjusted in the zero-inflated Poisson model (7). Due to the relatively small sample size of this dataset, i.e., less than 200 cells in each group, the accelerated bootstrap procedure based on asymptotic distributions (noted as ASY) will have lower power in detecting DE genes. Therefore, we focus on the simple bootstrap procedure with our two-sample test based on the modified K-S statistic (6) (noted as KS) in this experiment. We also apply MAST, ROTS, DESeq, Limma-Voom, SCDE, scDD and edgeR for comparison, based on the code in[25]. Since we test on these 2000 genes simultaneously, we use Benjamini-Hochberg(BH) method to adjust the p-values and control for multiple testing by calculating the false discovery rate (FDR) for each gene.[34] Since most expressed genes have relatively small FDR, we use 999 bootstrap samples for our test with early stopping rule and consider genes with FDR < 0.01 as differentially expressed (DE). From Table 1, out test with the simple bootstrap procedure (KS) detect the largest number of DE genes among the 2000 selected genes across all methods. An example where a DE gene can only detected by our KS method is provided in Section S3 in the supplementary materials. The average expression of this gene is similar in the two groups but the shape of gene distribution differs. The Venn diagrams in Figure 6 show that our KS method has significantly overlapping DE genes with MAST, DESeq, SCDE, scDD and edgeR, while the overlapping DE gene list between our test and Limma-Voom or ROTS is not significant as both overlapping p-values are greater than 0.05. Limma-Voom and ROTS were not specifically designed for scRNA-seq data which might the reason for the poor agreement between our method and these two methods. ASY method also has similar agreement with other existing methods (See Figure S2 in the supplementary materials for details). Moreover, the two p-value estimation procedures for our test are highly consistently in the detection of DE genes.

3.2.3 | Validation with real data-based simulation

Since we do not know the ground truth in the real dataset, to further assess the comparison between our method and the other methods, we simulate data based on the real dataset. First, we restrict the data to a subset of 100 genes selected with the highest mean and standard deviation of read counts across all cell lines. Second, based on the results from our KS method, we select those genes with FDR < 0.01 as true DE genes and the remaining ones as true null genes. For the true null genes, we pool the two groups together to estimate the distribution of gene expression and the probability of excessive zero counts using g-modelling. For the true DE genes, we estimate the distribution of the gene expression and the probability of excessive zero counts separately for the two groups using g-modelling. Third, based on the estimated distributions of gene expression and the estimated probability of excessive zero counts, we simulate the raw reads counts from our zero-inflated Poisson model. Finally, we apply our KS method, ASY method and the other methods to our simulated data to identify DE genes. To account for the randomness of data generation, we repeat the above simulation five times.

With our KS method, 86 DE genes and 14 null genes are detected from the real data, and therefore they are used as true DE genes and true null genes in our simulations. On average, our KS method detects 84 true DE genes and mis-identifies 3 genes. However our ASY method shows poorer performance with 70 true DE genes found and mis-identifies 16 genes. However, there are no null genes mis-identified as DE genes by ASY method. Due to the small sample size of this experiment, the ASY method is more conservative than the KS method, and therefore the ASY method detected fewer DE genes. Among other methods, scDD detects 81 true DE genes as the most but mis-identifies 5 genes. Limma-Voom detects 76 DE genes among which 6 are misidentified. Similarly as Limma-Voom, there is 79 DE genes found by edgeR with 7 of them mis-identified. MAST does not detect any true null genes as DE genes, but it only detects 70 DE genes. ROTS has the poorest performance with both low power and high FDR. SCDE has similar performance as DESeq, where their power is similar to MAST but FDR is higher. Table 2 shows the observed FDR and AUC of all the methods in all five simulations, with the average ROC curves shown in Figure 7. We can see that our KS method, ASY method and MAST have controlled FDRs at level 0.01, while ROTS, DESeq, Limma-Voom, SCDE, scDD and edgeR have inflated FDRs. In terms of AUC, our KS method and ASY methods rank the top two, followed by MAST and scDD, and ROTS ranks the last. When ignoring the choice of p-value cutoff, the rankings of the p-values from our KS and ASY methods were very similar and therefore the AUC of the ASY method is similar with the KS method. Although scDD has high power in detecting DE genes, its relatively high FDR may compromise its performance of real data application. Overall, both our KS method and ASY method outperform other methods in this comparison.

4 | DISCUSSION

In many statistical problems, we observe two unknown distributions indirectly and aim to investigate the difference between them.[35] The unknown distribution can be estimated through deconvolution, in accordance with existing methods.[12] However, these methods are only designed for one-sample estimation. Thus, we combine the existing g-modelling method with two-sample K-S test statistic for two-sample density comparison. In terms of p-value estimation, we propose two versions of bootstrap procedures to cover the wide range of sample sizes and balance the needs for accuracy and speed. For small sample size, the simple bootstrap procedure has higher power than the accelerated bootstrap procedure for detecting the difference in distributions. For large sample size, the accelerated bootstrap procedure based on the asymptotic null distribution provides similar statistical power while being much more computationally efficient than the simple bootstrap procedure.

Our approach can be applied to a wide range of areas, as our approach is capable of handling various types of data including count or continuous outcomes. In the analysis of surgical nodes data with binomial models, our proposed test has controlled type I error and sufficient power in several cases of differences in distribution. In terms of scRNA-seq data application, the existing parametric methods all assumed that the unknown gene distributions follow a particular parametric family, and only detect changes in one (or a few) parameters (usually just the location parameter) of that family, which is restrictive. On the contrary, our proposed test assume that the underlying parameters (i.e., true expression levels of a gene in different samples) follow unknown distributions, and g-modeling allows us to model these unknown distributions and detect changes in distribution from one condition to the other. Compared with other existing methods for differential expression analysis on scRNA-seq data, our approach can detect more DE genes and has well-controlled false discovery rate.

There are several directions for potential future research. The current choice of grid for discretizing the sample space of θ is subjective, and data adaptive approaches could be considered to obtain more efficient density estimation. The computation efficiency might be further improved by the deriving the null distribution the our test statistics directly to avoid the use of any bootstrap procedure. The accelerated bootstrap procedure might be calibrated to boost its power for data with small sample size. The two-group comparison can be further extended to multi-group comparison with a k-sample test statistics. The adjustment for additional covariates could be included in the model to control for confounding effects and to improve the the statistical power. We can also extend our two-sample test to generalized linear mixed models to account for correlated or clustered observations.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data and computer codes that support the findings in this paper are available at at https://github.com/kkttzjy/gmodeltest. These data were derived from the following resources available in the public domain: https://cran.rstudio.com/web/packages/deconvolveR/index.html and https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3929/.

References

1. Pratt JW, Gibbons JD. *Kolmogorov-Smirnov Two-Sample Tests*: 318-344; New York, NY: Springer New York . 1981
2. Efron B. Empirical Bayes deconvolution estimates. *Biometrika* 2016; 103(1): 1-20. doi: 10.1093/biomet/asv068
3. Efron B, Hastie T. *Computer age statistical inference : algorithms, evidence, and data science*. New York, NY, USA : Cambridge University Press, 2016 Cambridge University Press . 2016.
4. Laird N. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association* 1978; 73(364): 805–811.
5. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* 1983; 78(381): 47-55. doi: 10.1080/01621459.1983.10477920

6. Zhang C. Empirical Bayes and compound estimation of normal means. *Statistica Sinica* 1997; 7(1): 181–193.

7. Jiang W, Zhang CH. General Maximum Likelihood Empirical Bayes Estimation of Normal Means. *The Annals of Statistics* 2009; 37(4): 1647–1684.

8. Robbins H. An Empirical Bayes Approach to Statistics. In: University of California Press; 1956; Berkeley, Calif.: 157–163.

9. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics MonographsCambridge University Press . 2010

10. Brown LD, Greenshtein E, Ritov Y. The Poisson Compound Decision Problem Revisited. *Journal of the American Statistical Association* 2013; 108(502): 741–749.

11. Efron B. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association* 2011; 106(496): 1602-1614. doi: 10.1198/jasa.2011.tm11181

12. Efron B. Two Modeling Strategies for Empirical Bayes Estimation. *Statist. Sci.* 2014; 29(2): 285–301. doi: 10.1214/13-STS455

13. Jiang H, Salzman J. Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika* 2012; 99(4): 973-980. doi: 10.1093/biomet/ass051

14. Gholami S, Janson L, Worhunsky DJ, et al. Number of Lymph Nodes Removed and Survival after Gastric Cancer Resection: An Analysis from the US Gastric Cancer Collaborative. *Journal of the American College of Surgeons* 2015; 221(2): 291—299. doi: 10.1016/j.jamcollsurg.2015.04.024

15. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology* 2016; 17(1): 63. doi: 10.1186/s13059-016-0927-y

16. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* 2015; 58(4): 610-620. doi: 10.1016/j.molcel.2015.04.005

17. Alam M, Al Mahi N, Begum M. Zero-Inflated Models for RNA-Seq Count Data. *Journal of Biomedical Analytics* 2018; 1: 55-70. doi: 10.30577/jba.2018.v1n2.23

18. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 2015; 16: 133. doi: 10.1038/nrg3833

19. Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013; 498: 236. doi: 10.1038/nature12172

20. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* 2015; 16(1): 278. doi: 10.1186/s13059-015-0844-5

21. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nature Methods* 2014; 11: 740. doi: 10.1038/nmeth.2967

22. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods* 2018; 145: 25-32. doi: https://doi.org/10.1016/j.ymeth.2018.04.017

23. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* 2015; 32(4): 533-541. doi: 10.1093/bioinformatics/btv634

24. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 2016; 17(1): 110. doi: 10.1186/s12859-016-0944-6

25. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics* 2016; 18(5): 735-743. doi: 10.1093/bib/bbw057

26. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* 2019; 20(1): 40. doi: 10.1186/s12859-019-2599-6

27. Korthauer KD, Chu LF, Newton MA, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* 2016; 17(1): 222. doi: 10.1186/s13059-016-1077-y

28. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology* 2010; 11(10): R106. doi: 10.1186/gb-2010-11-10-r106

29. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* 2004; 3(1): 3-25. doi: 10.2202/1544-6115.1027

30. Seyednasrollah F, Rantanen K, Jaakkola P, Elo LL. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Research* 2015; 44(1): e1-e1. doi: 10.1093/nar/gkv806

31. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 2012; 40(10): 4288-4297. doi: 10.1093/nar/gks042

32. Petropoulos S, Edsgärd D, Reinius B, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* 2016; 165(4): 1012-1026. doi: 10.1016/j.cell.2016.03.023

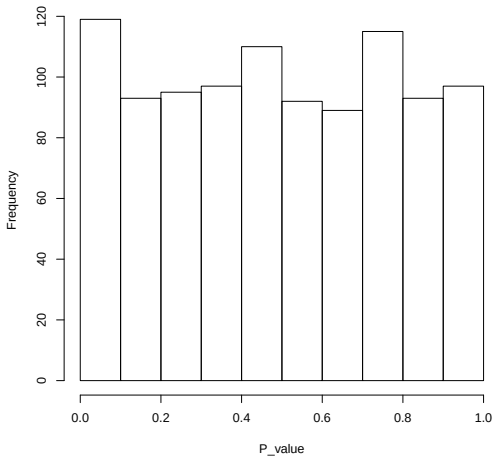
33. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 2018; 34(18): 3223-3224. doi: 10.1093/bioinformatics/bty332

34. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; 57(1): 289–300.

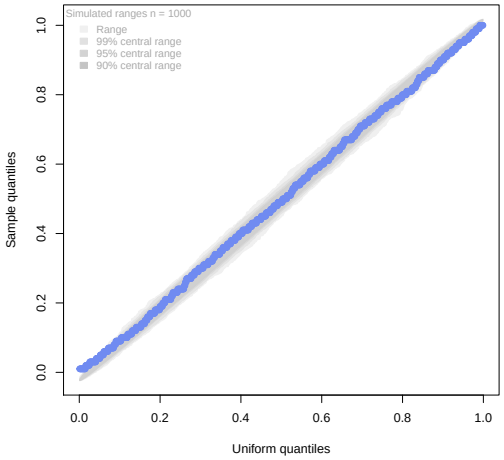
35. Madrid-Padilla OH, Polson NG, Scott J. A deconvolution path for mixtures. *Electron. J. Statist.* 2018; 12(1): 1717-1751. doi: 10.1214/18-EJS1430

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

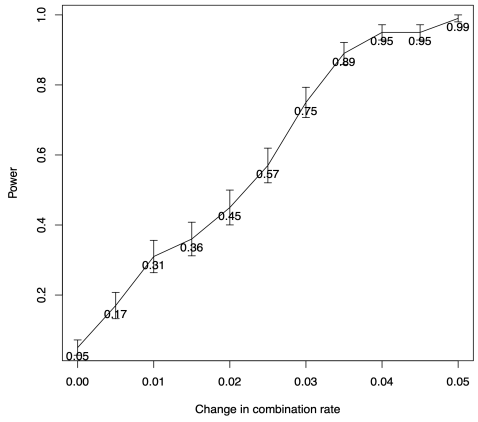


(a) Histogram of p-values

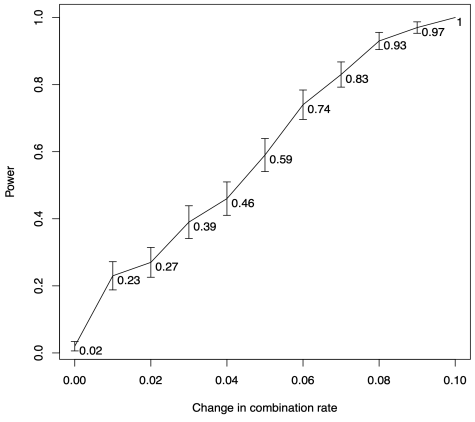


(b) Q-Q plot of p-values

FIGURE 1 (a) Histogram for estimated *p*-values from 1000 simulations under the null hypothesis and (b) Q-Q plot for estimated p-values against $U(0, 1)$.



(a) Alternative case 1

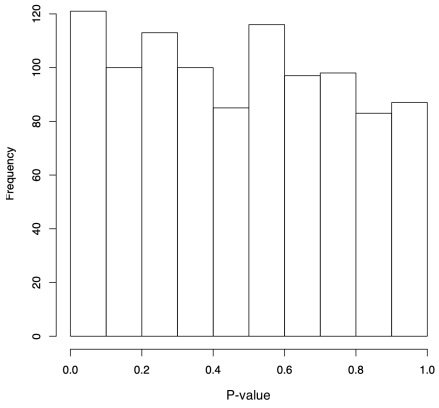


(b) Alternative case 2

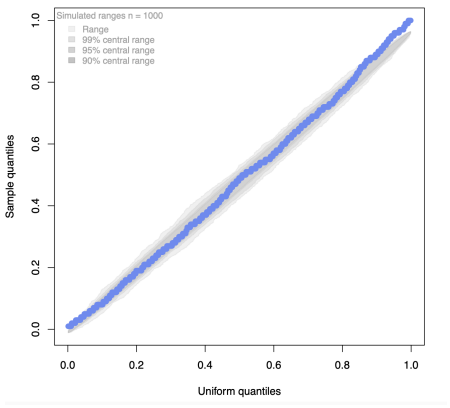
FIGURE 2 Power plots for alternative cases (a) 1 and (b) 2. For each line, the power changes with the difference in π varying from 0.1 to 1. The vertical bar represents the mean $\pm$ sd for each power value.

TABLE 1 Number of DE genes found among 2000 selected genes by proposed methods and other methods.

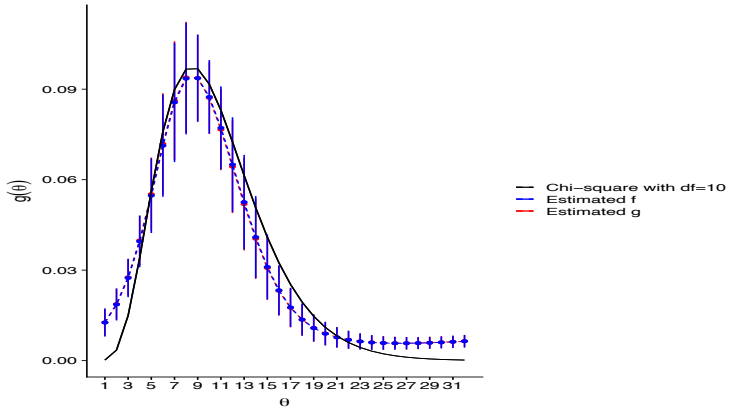
Method	KS	ASY	MAST	ROTS	DESeq	Limma-Voom	SCDE	scDD	edgeR
DE genes	1748	1294	1570	1647	1398	1546	1492	1642	1612



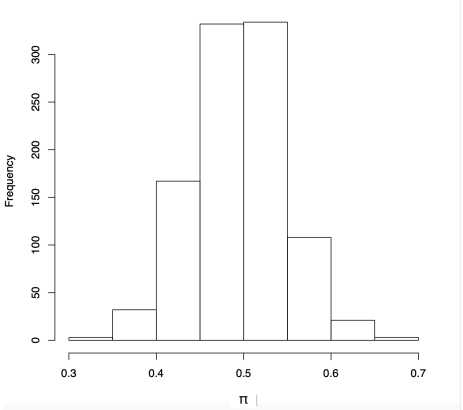
(a) Histogram of p-values



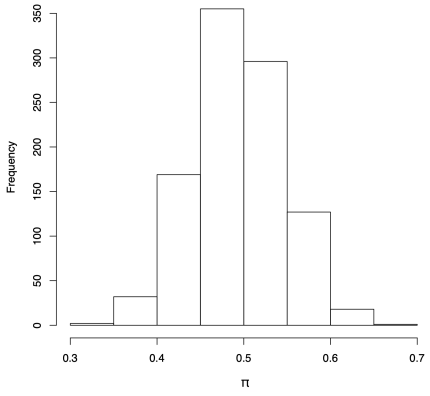
(b) Q-Q plot of p-values



(c) Density comparison

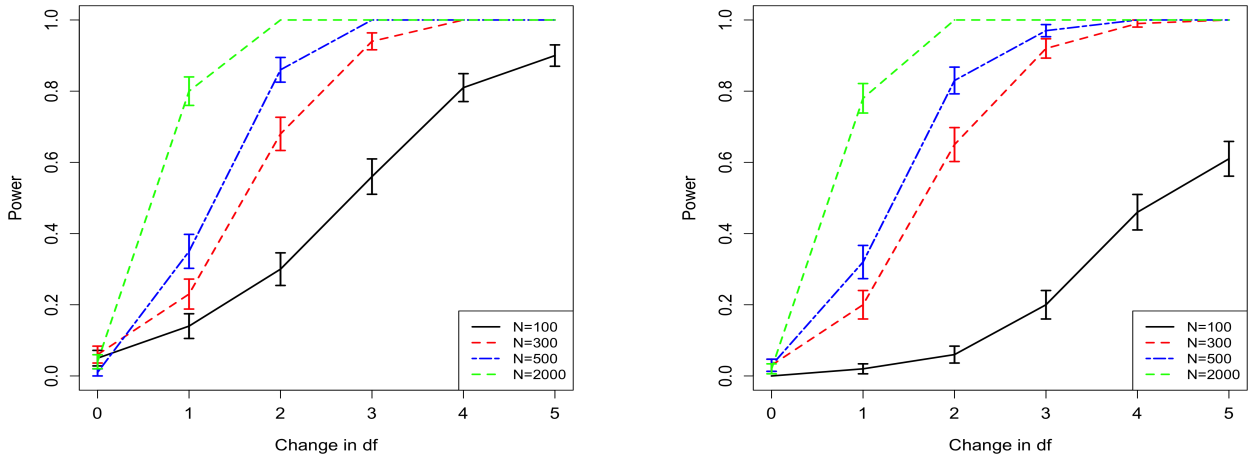


(d) Histogram of π_X



(e) Histogram of π_Y

FIGURE 3 (a) Histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density χ^2_{10}, and histograms for estimated (d) π_X and (e) π_Y.



(a) Simple bootstrap

(b) Accelerated bootstrap

FIGURE 4 Power plots of proposed (a) simple and (b) accelerated bootstrap procedures with varying sample sizes. For each line, the power changes with the difference in degrees of freedom varying from 0 to 5. The vertical bars represent the mean ± sd for each power value.

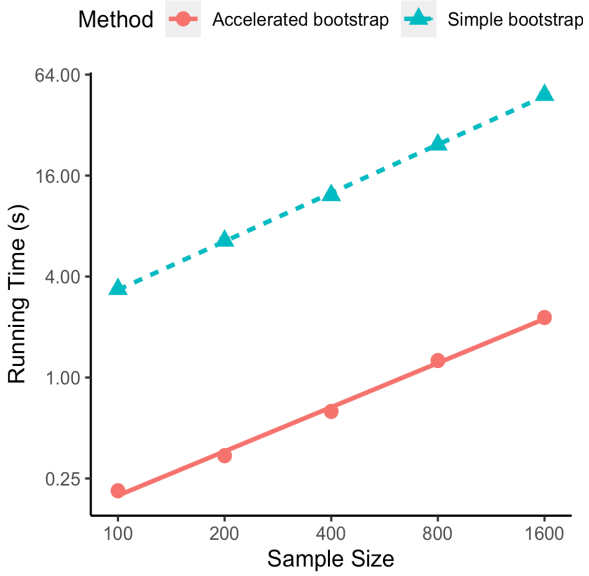


FIGURE 5 Log-log plot of Running time (in seconds) for proposed simple and accelerated bootstrap procedures against sample size in each group under null hypothesis for normal-based model.

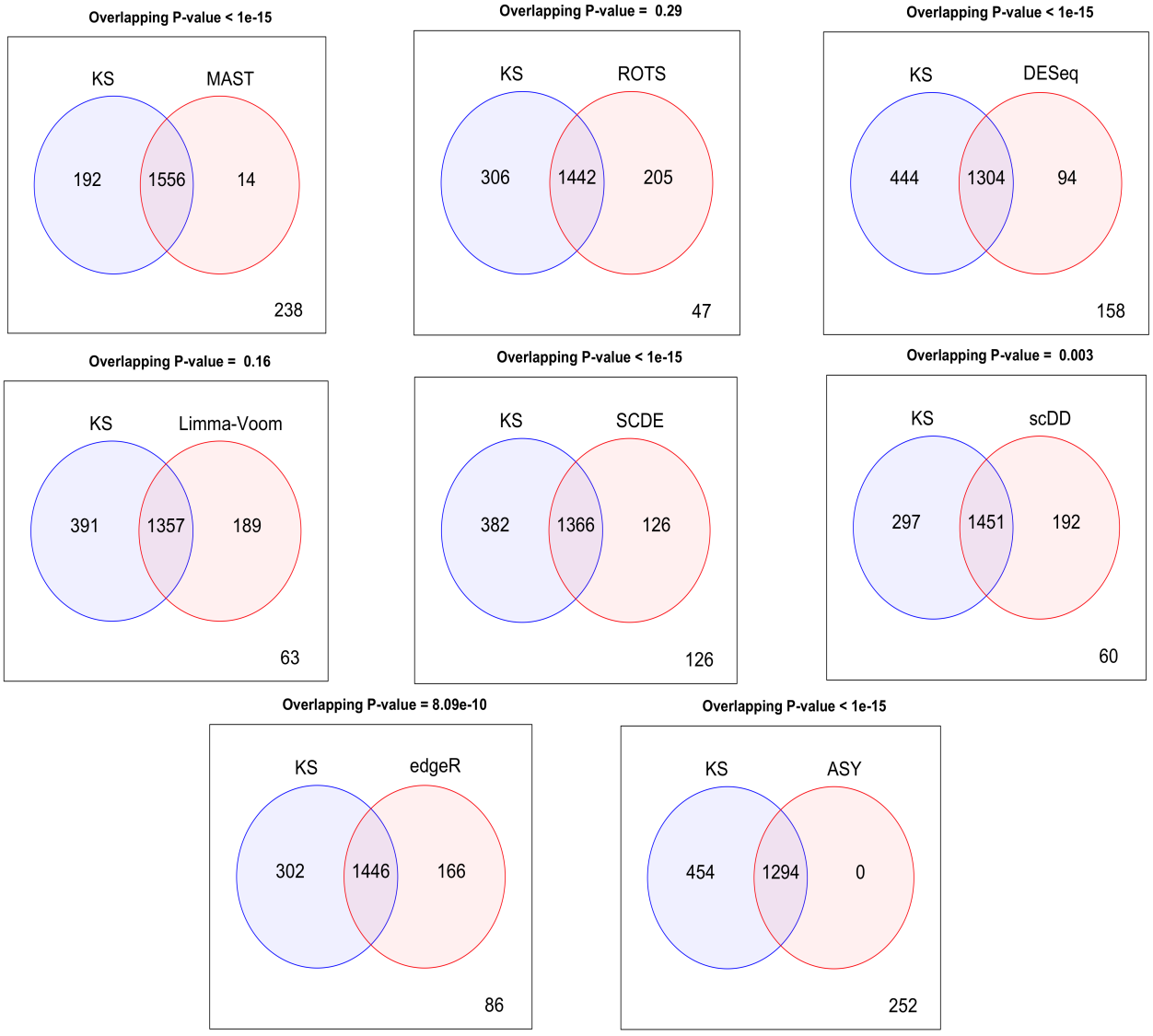


FIGURE 6 Venn diagrams for numbers of DE genes found by proposed method (KS) versus other methods with p-values from hypergeometric tests.

TABLE 2 Observed FDRs and AUCs of proposed methods (KS and ASY) and other methods in five simulations.

Simulation		1	2	3	4	5	Mean (SD)
Observed FDR	KS	0.00	0.07	0.00	0.00	0.00	0.01 (0.03)
	ASY	0.00	0.00	0.00	0.00	0.00	0.00 (0.00)
	MAST	0.00	0.00	0.00	0.00	0.00	0.00 (0.00)
	ROTS	0.07	0.00	0.00	0.64	0.79	0.30 (0.38)
	DESeq	0.07	0.29	0.36	0.07	0.14	0.19 (0.13)
	Limma-Voom	0.29	0.36	0.79	0.36	0.36	0.43 (0.20)
	SCDE	0.00	0.36	0.43	0.21	0.07	0.21 (0.18)
	scDD	0.29	0.14	0.36	0.14	0.29	0.24 (0.10)
	edgeR	0.43	0.57	0.79	0.43	0.50	0.54 (0.15)
AUC	KS	1.00	1.00	1.00	1.00	1.00	1.00 (0.00)
	ASY	0.98	1.00	0.98	0.99	0.99	0.99 (0.01)
	MAST	0.98	0.97	0.98	0.94	0.94	0.96 (0.02)
	ROTS	0.81	0.80	0.71	0.42	0.48	0.64 (0.18)
	DESeq	0.84	0.83	0.80	0.86	0.85	0.84 (0.02)
	Limma-Voom	0.84	0.81	0.74	0.85	0.82	0.81 (0.04)
	SCDE	0.96	0.81	0.77	0.80	0.91	0.85 (0.08)
	scDD	0.95	0.99	0.93	0.97	0.95	0.96 (0.02)
	edgeR	0.85	0.82	0.76	0.82	0.82	0.81 (0.03)

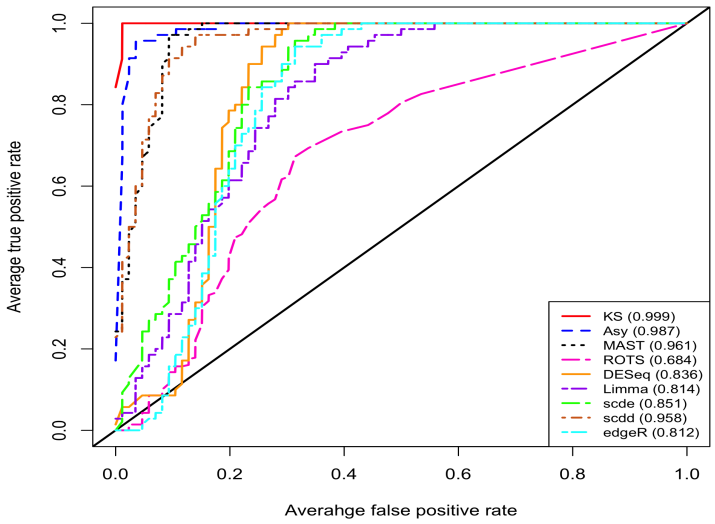
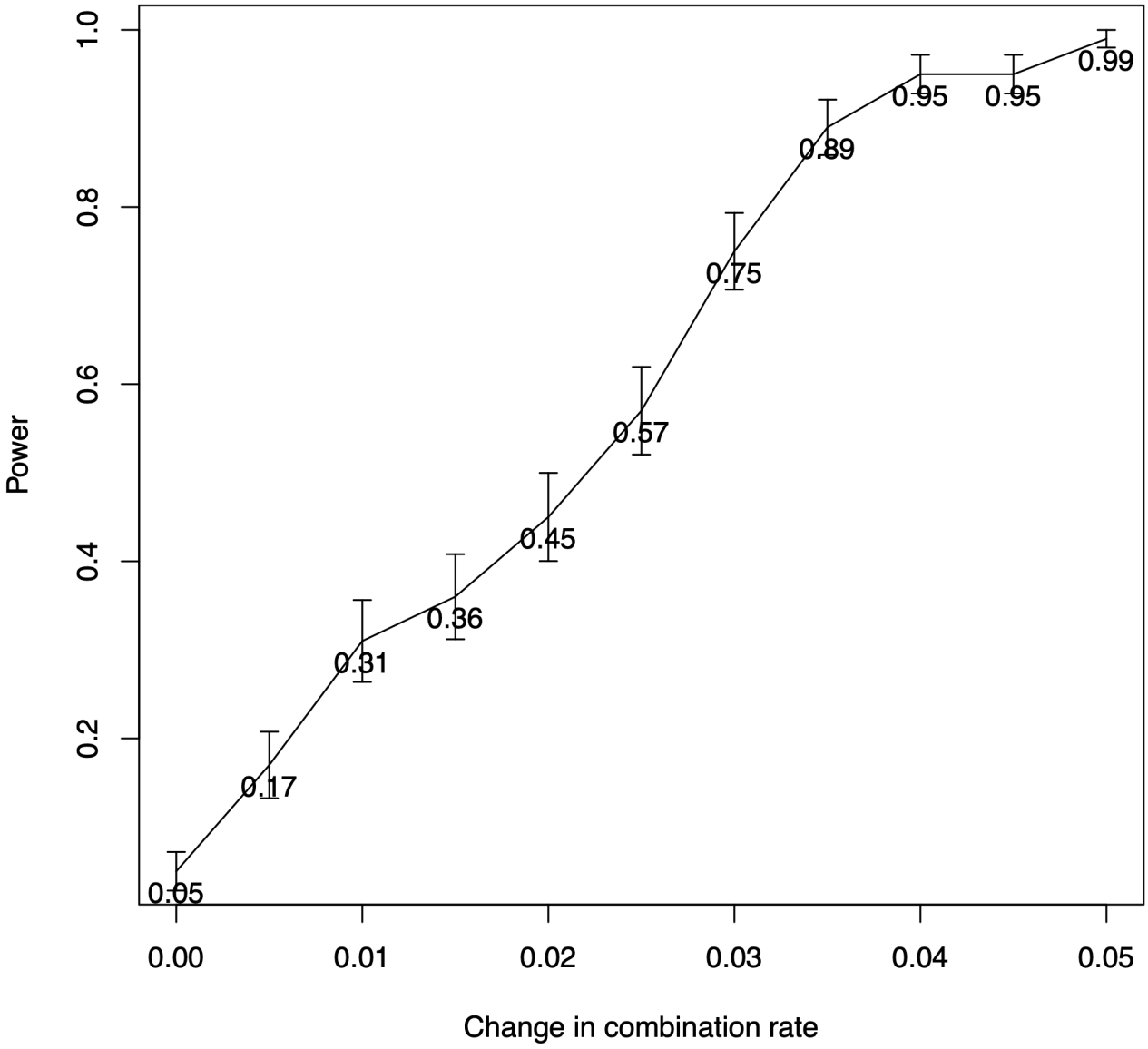
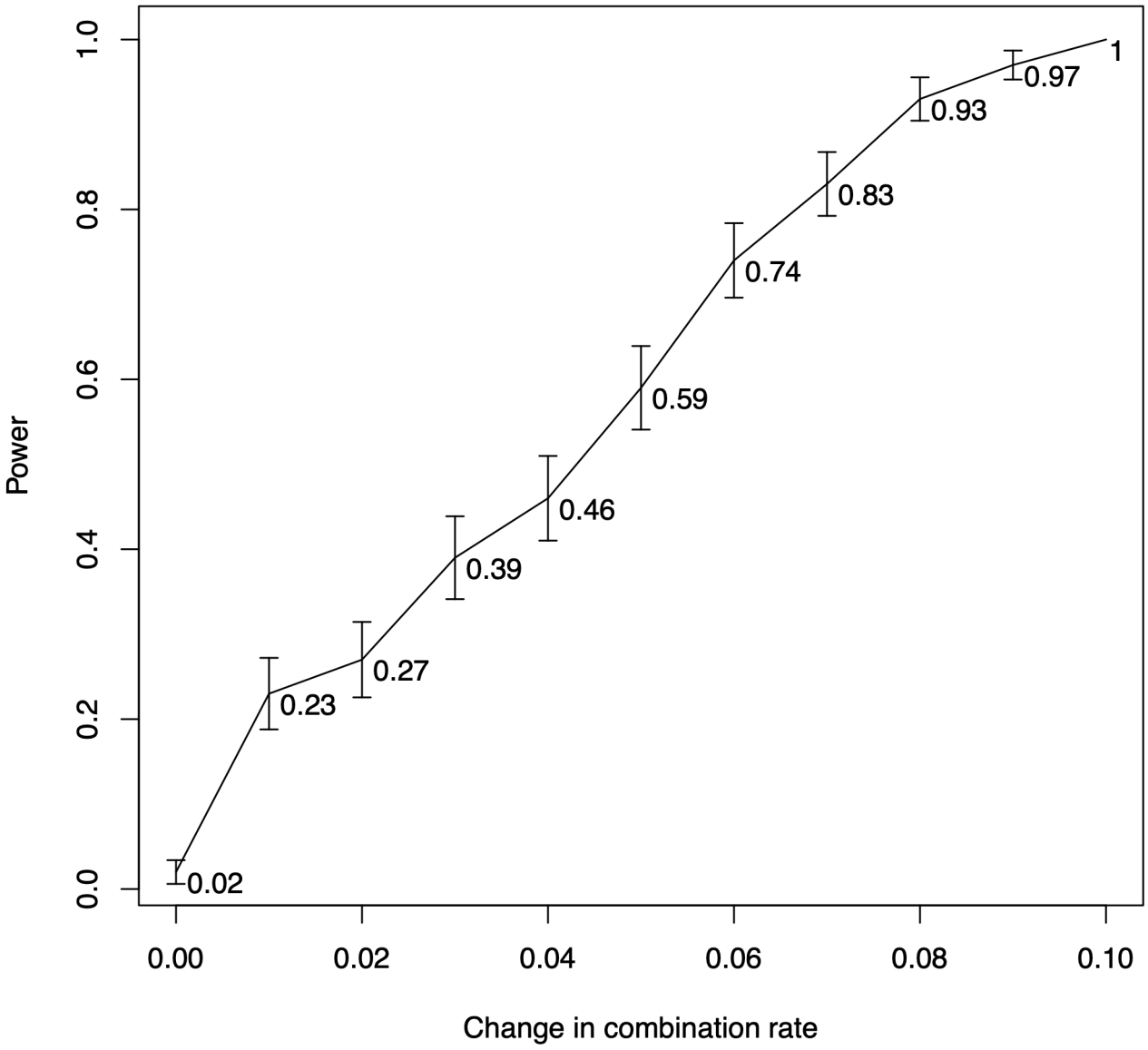
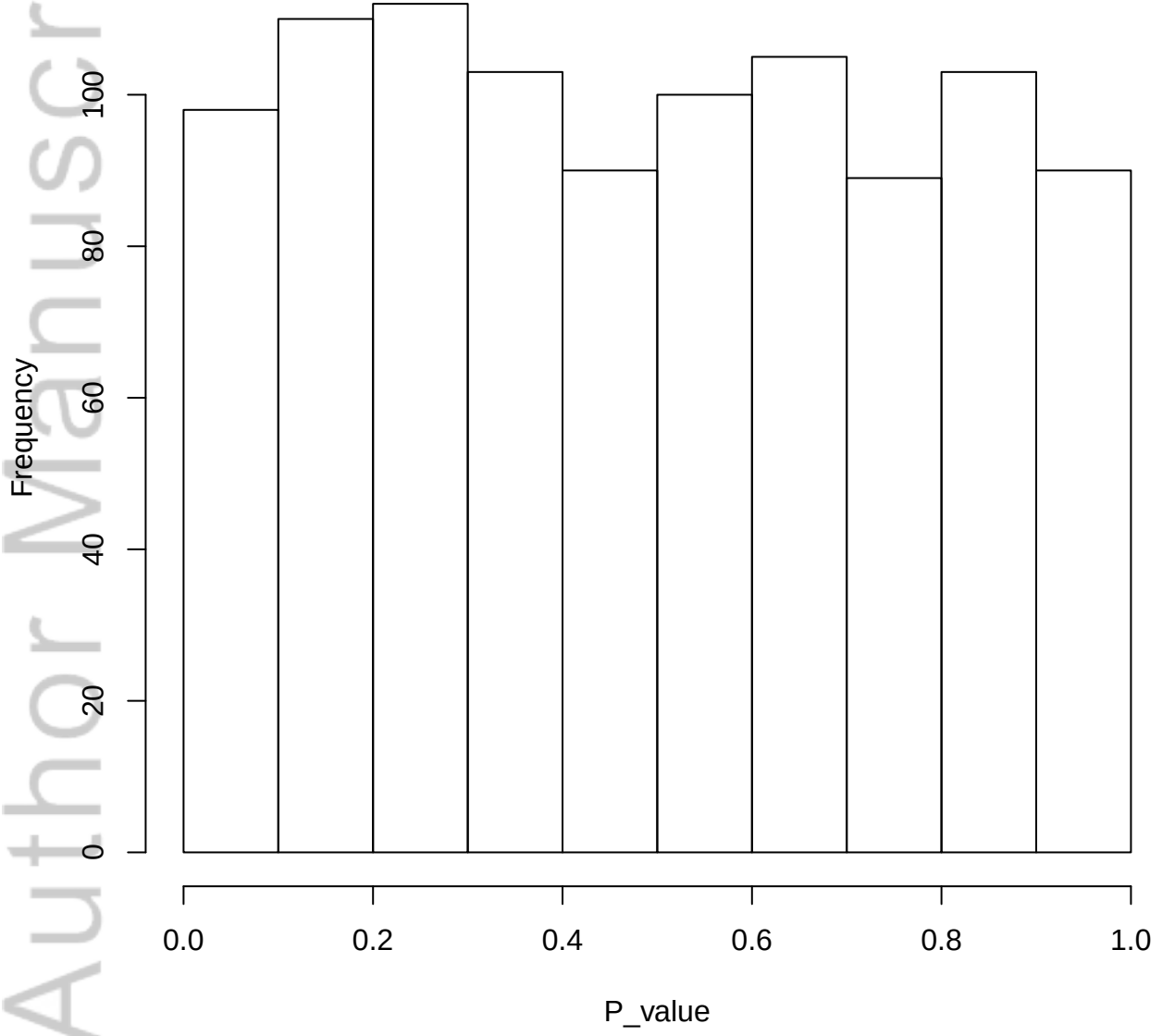


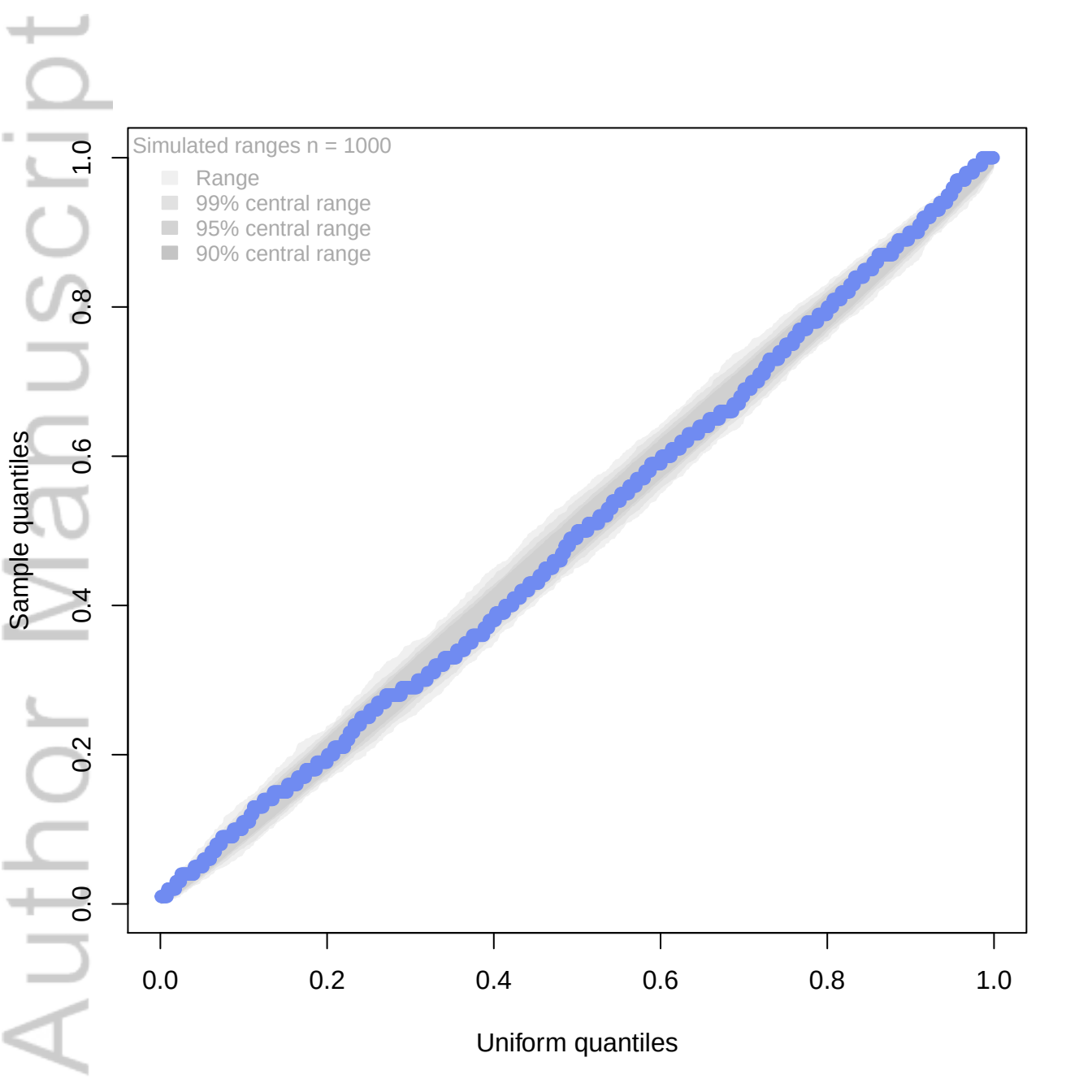
FIGURE 7 Mean ROC curves across five simulations for eight methods.

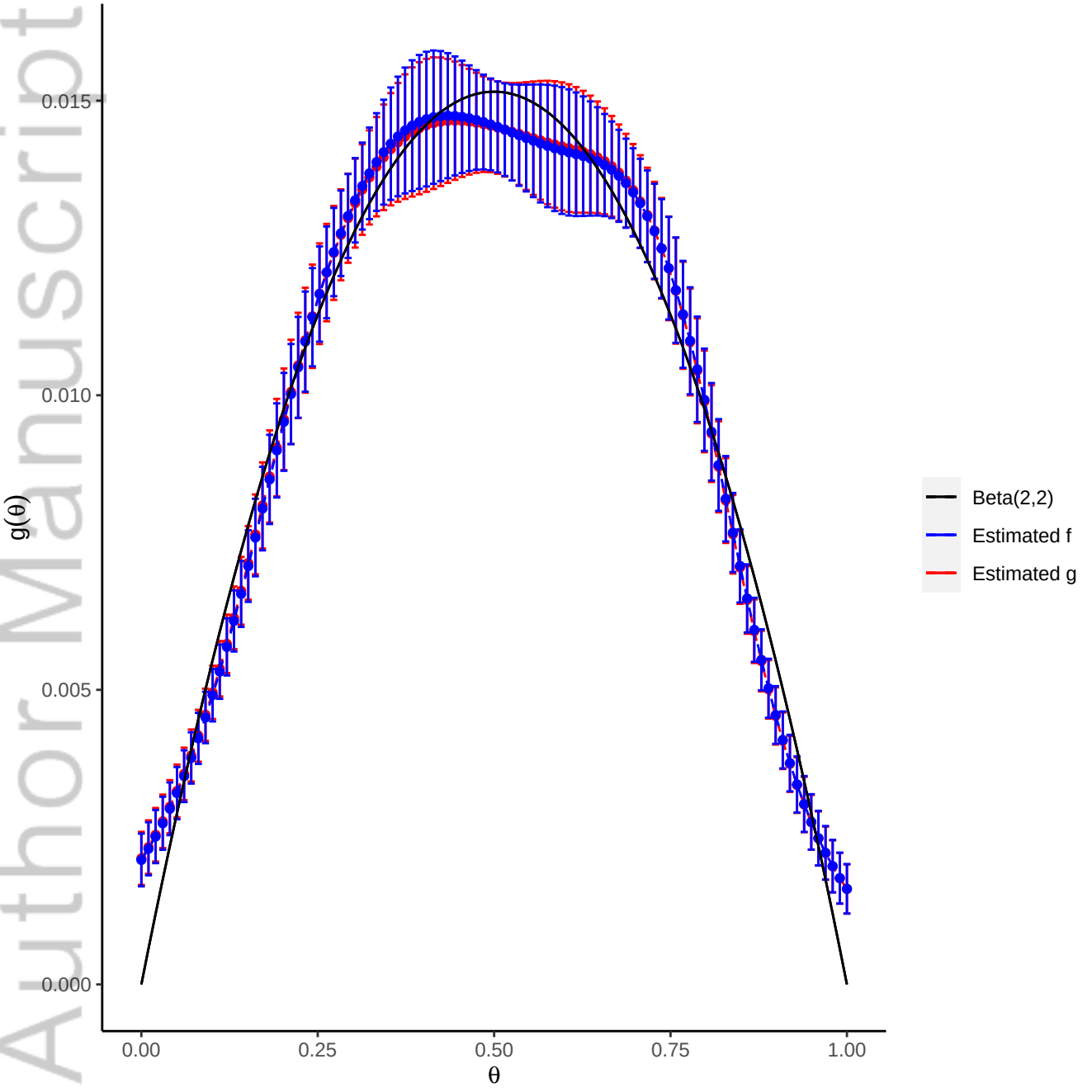
How to cite this article:

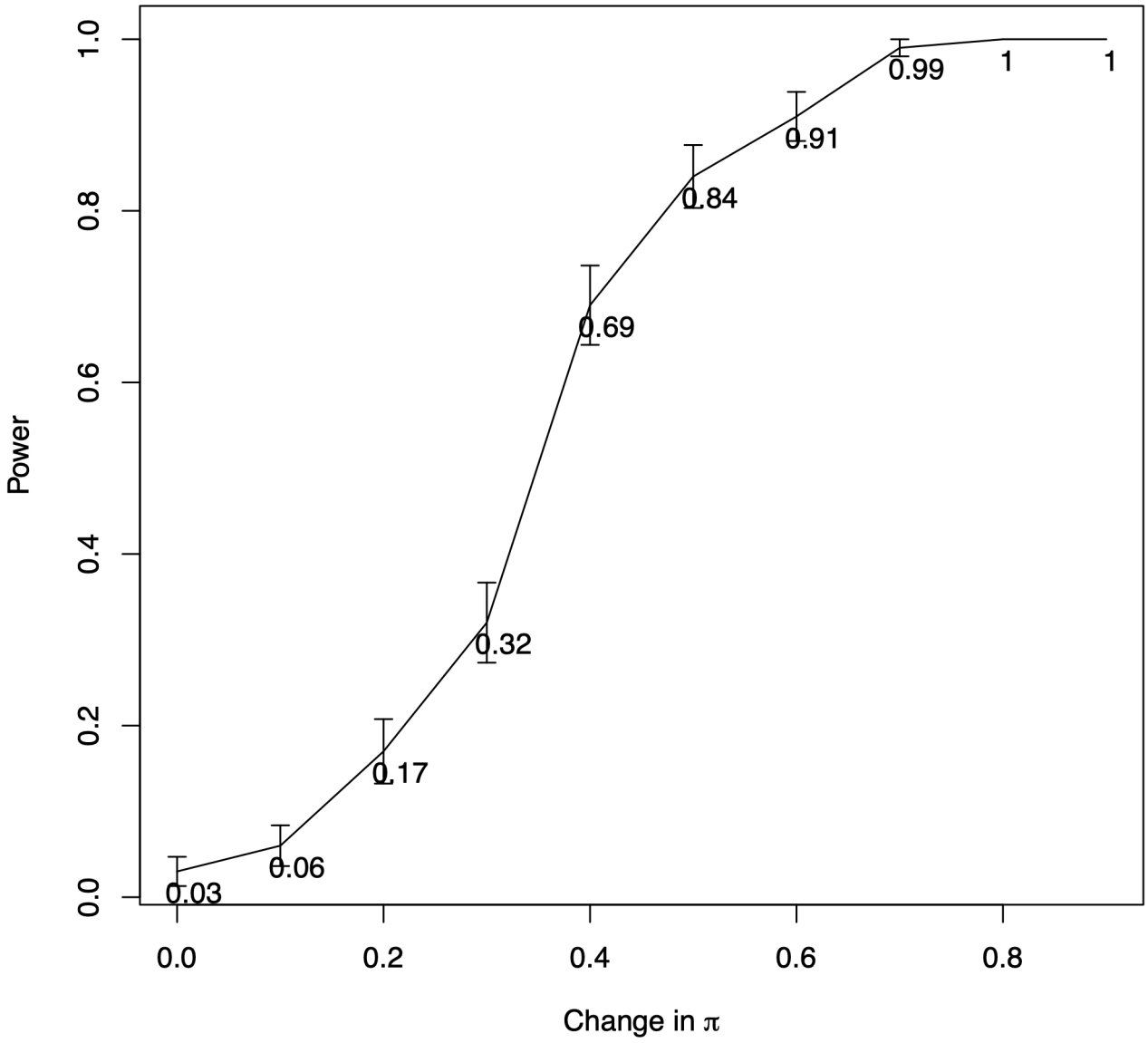


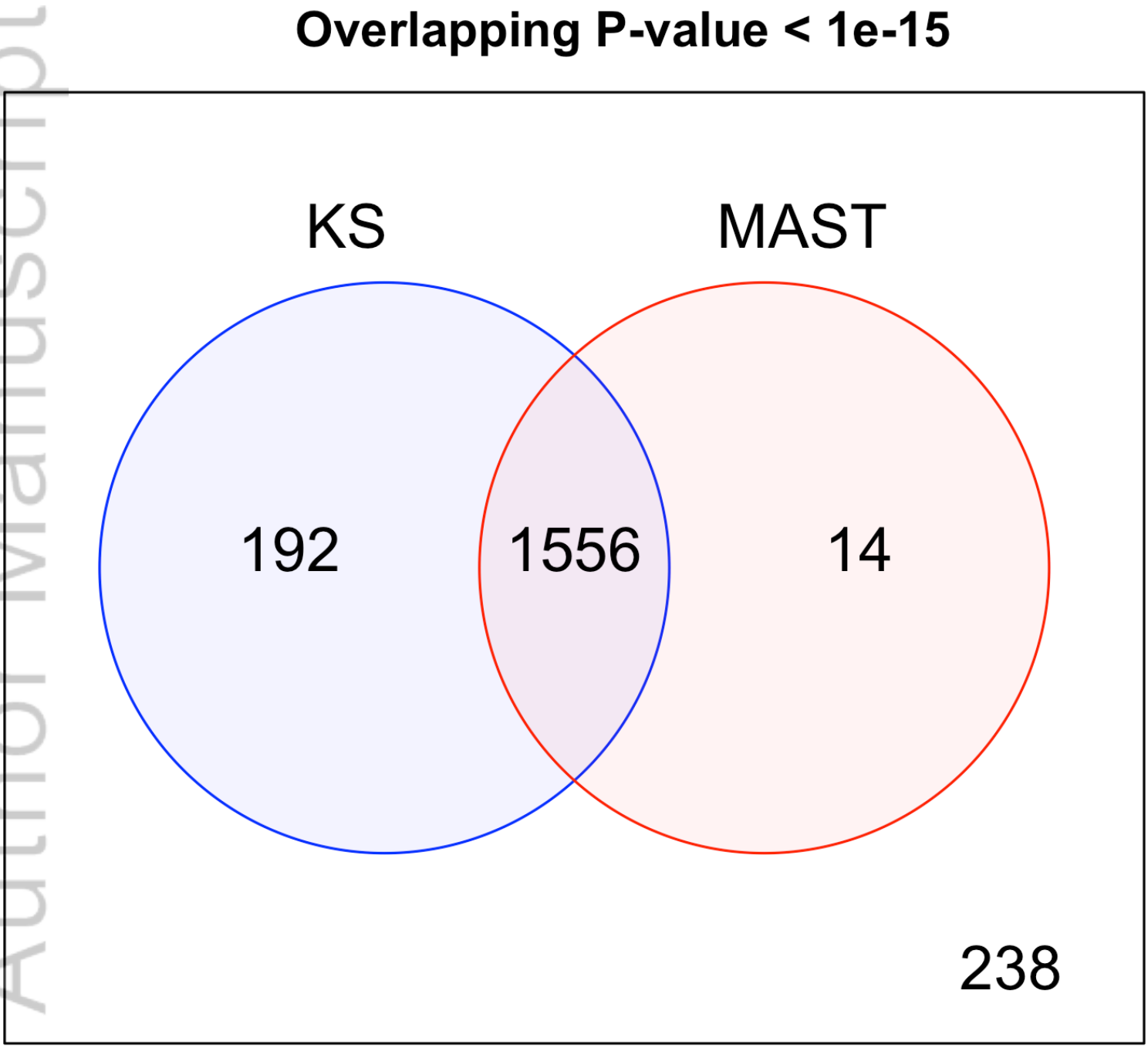




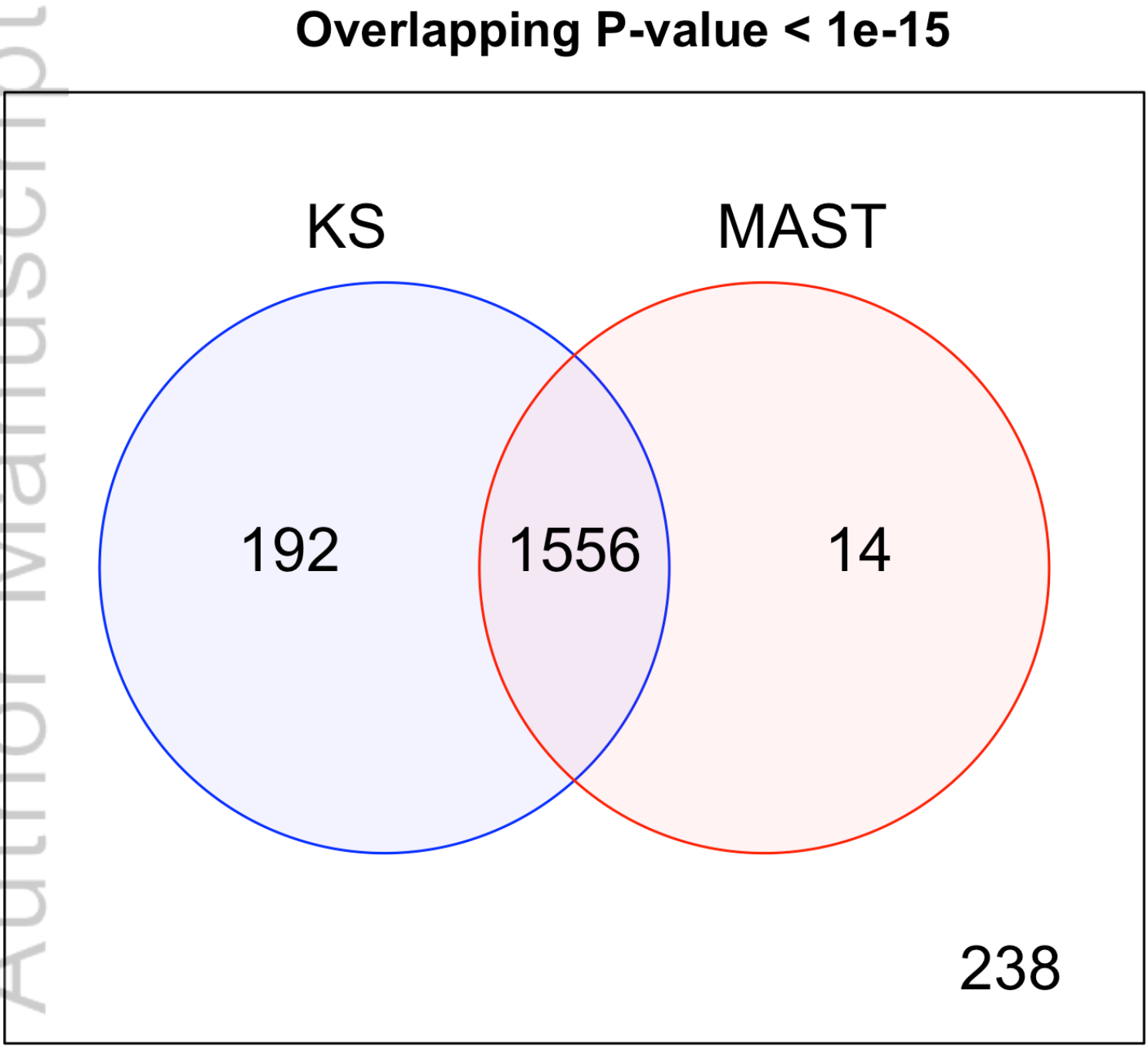




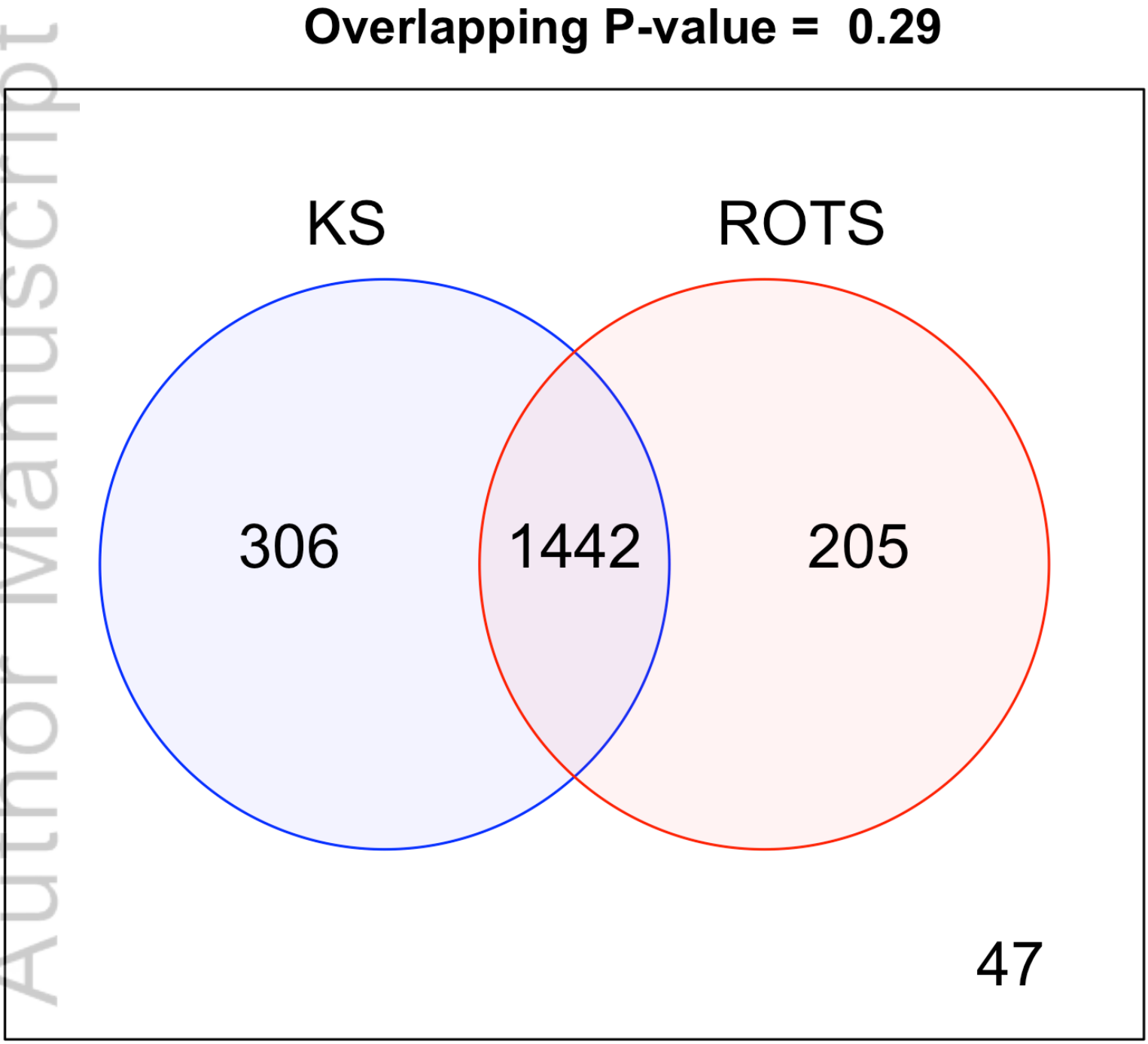




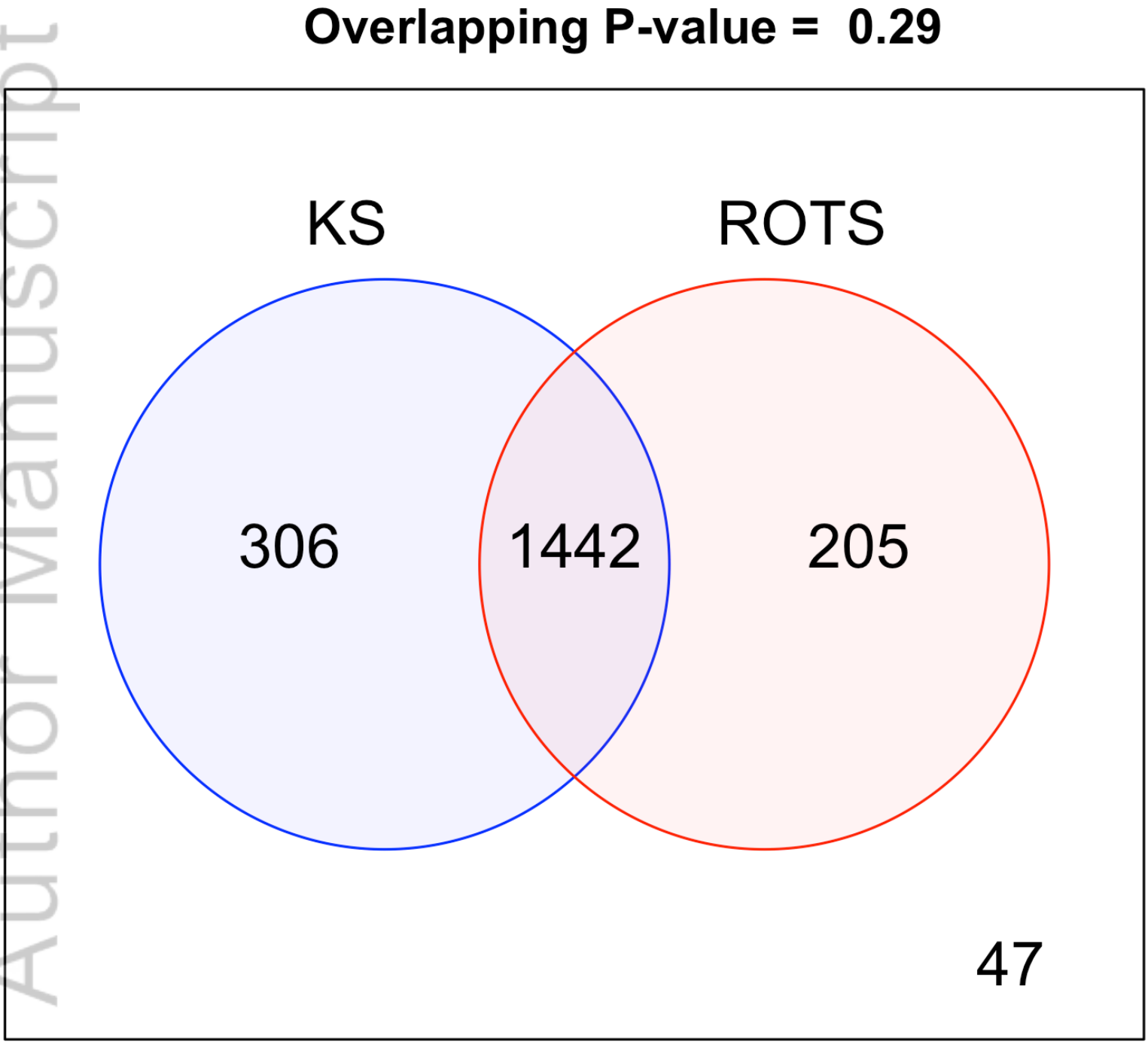
C1.png



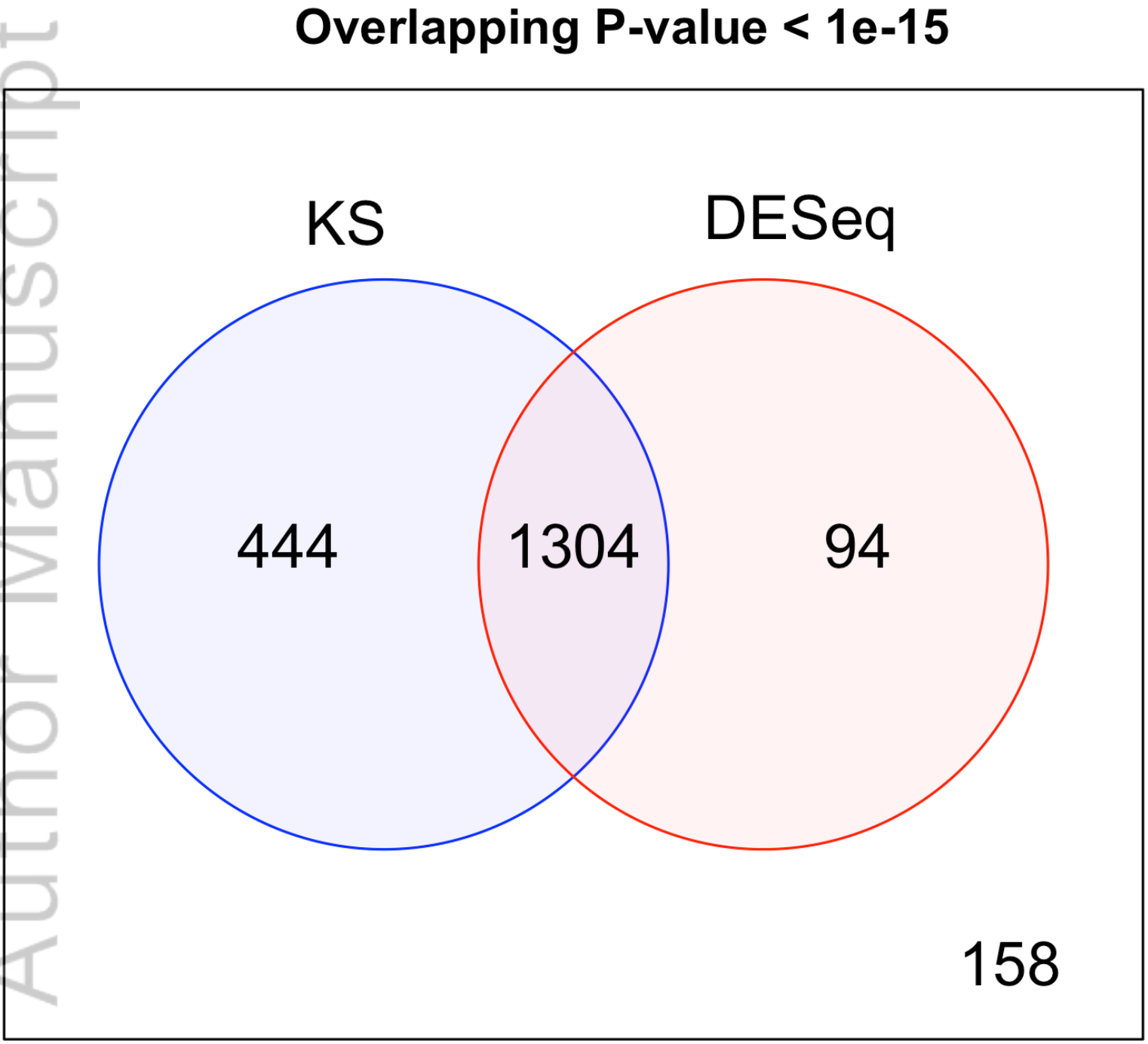
C1.png



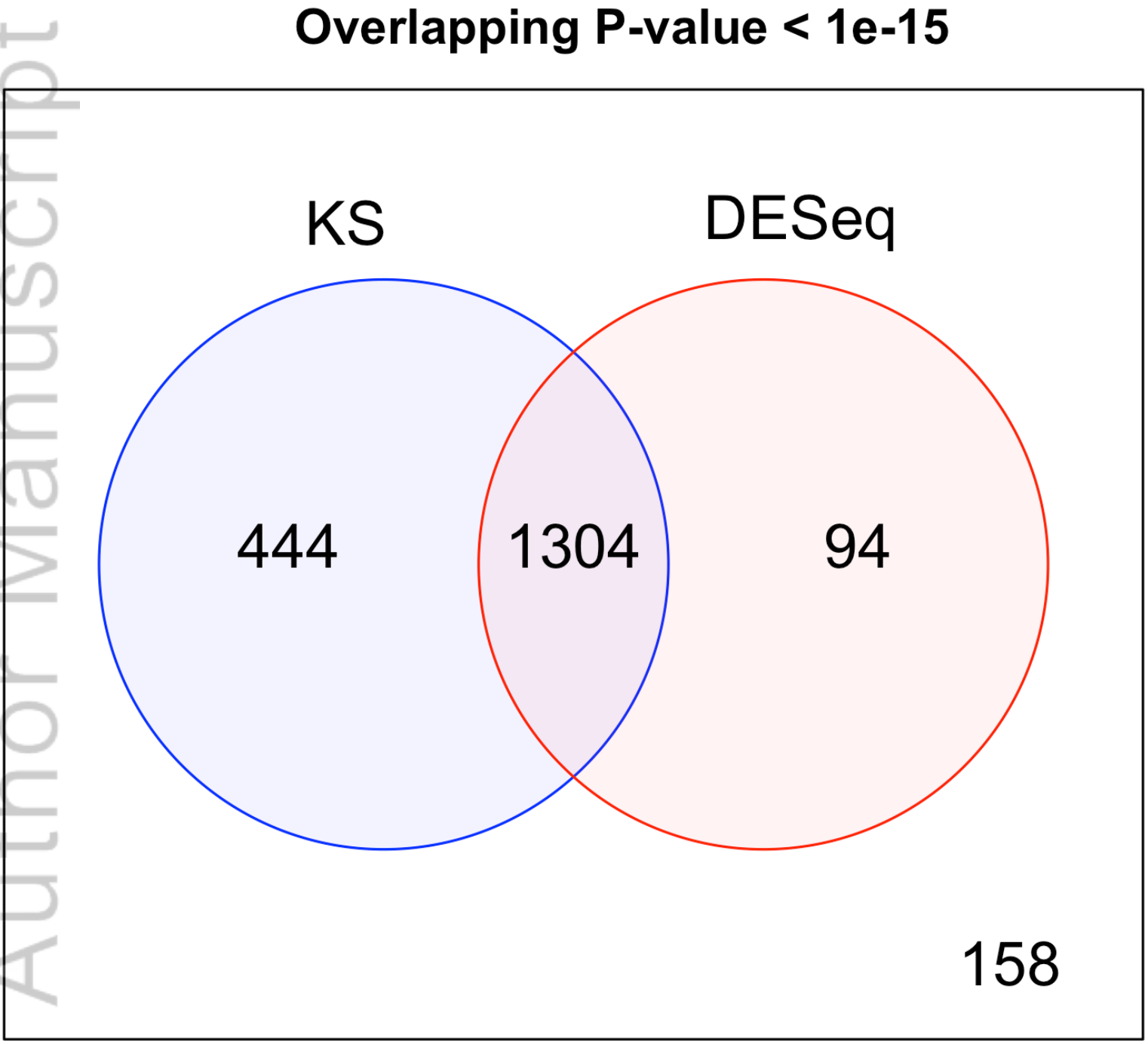
C2.png



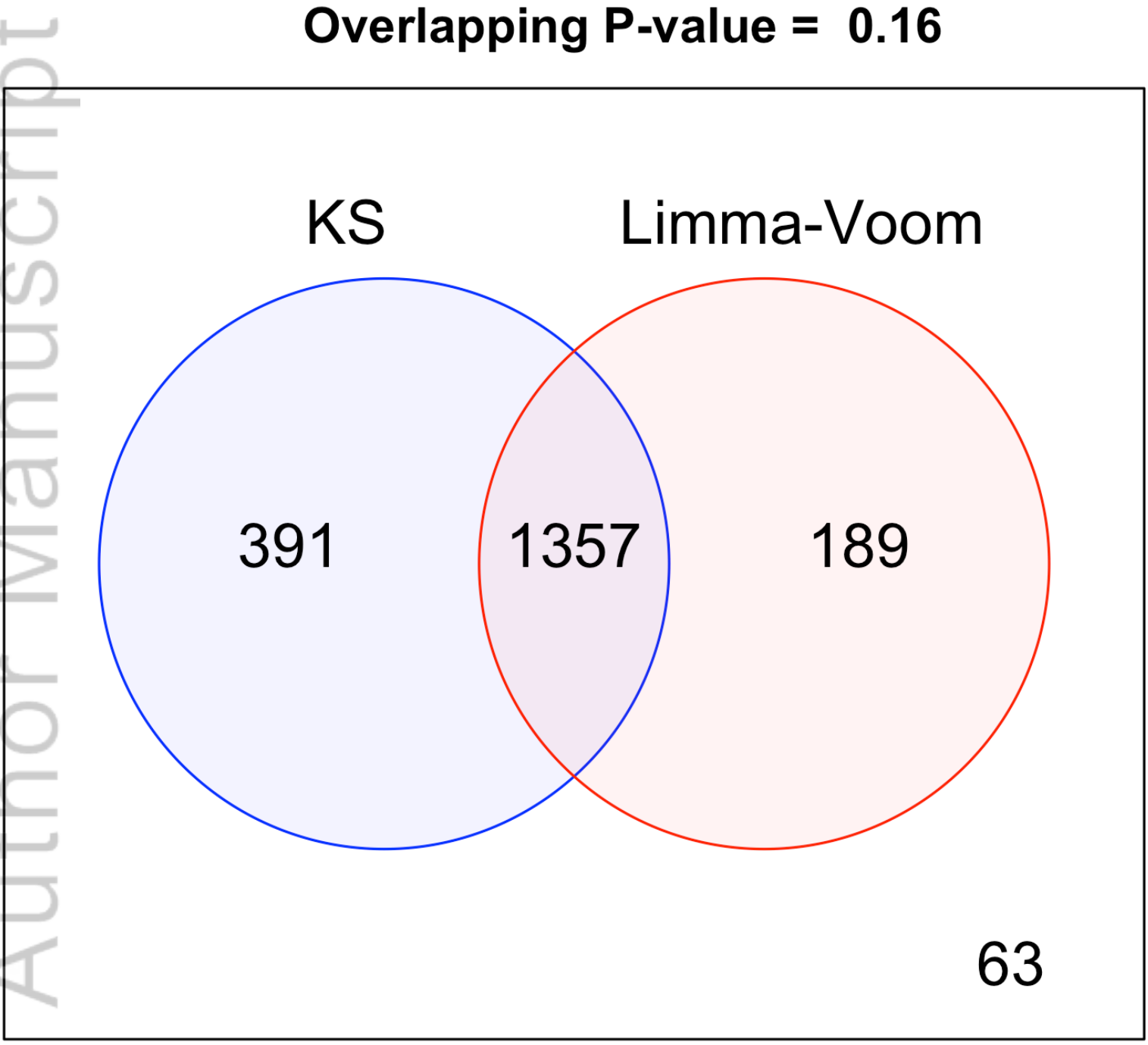
C2.png



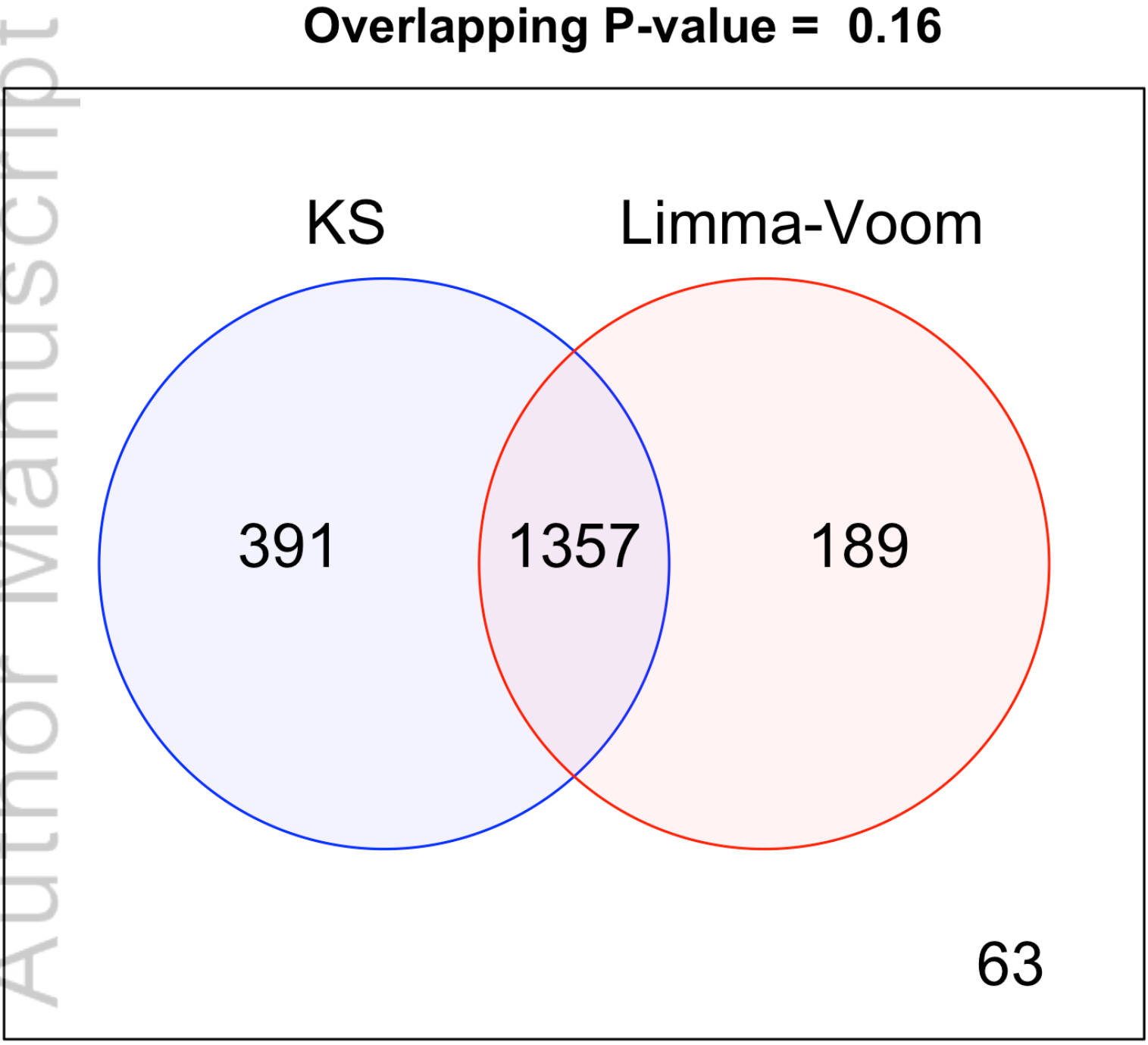
C3.png



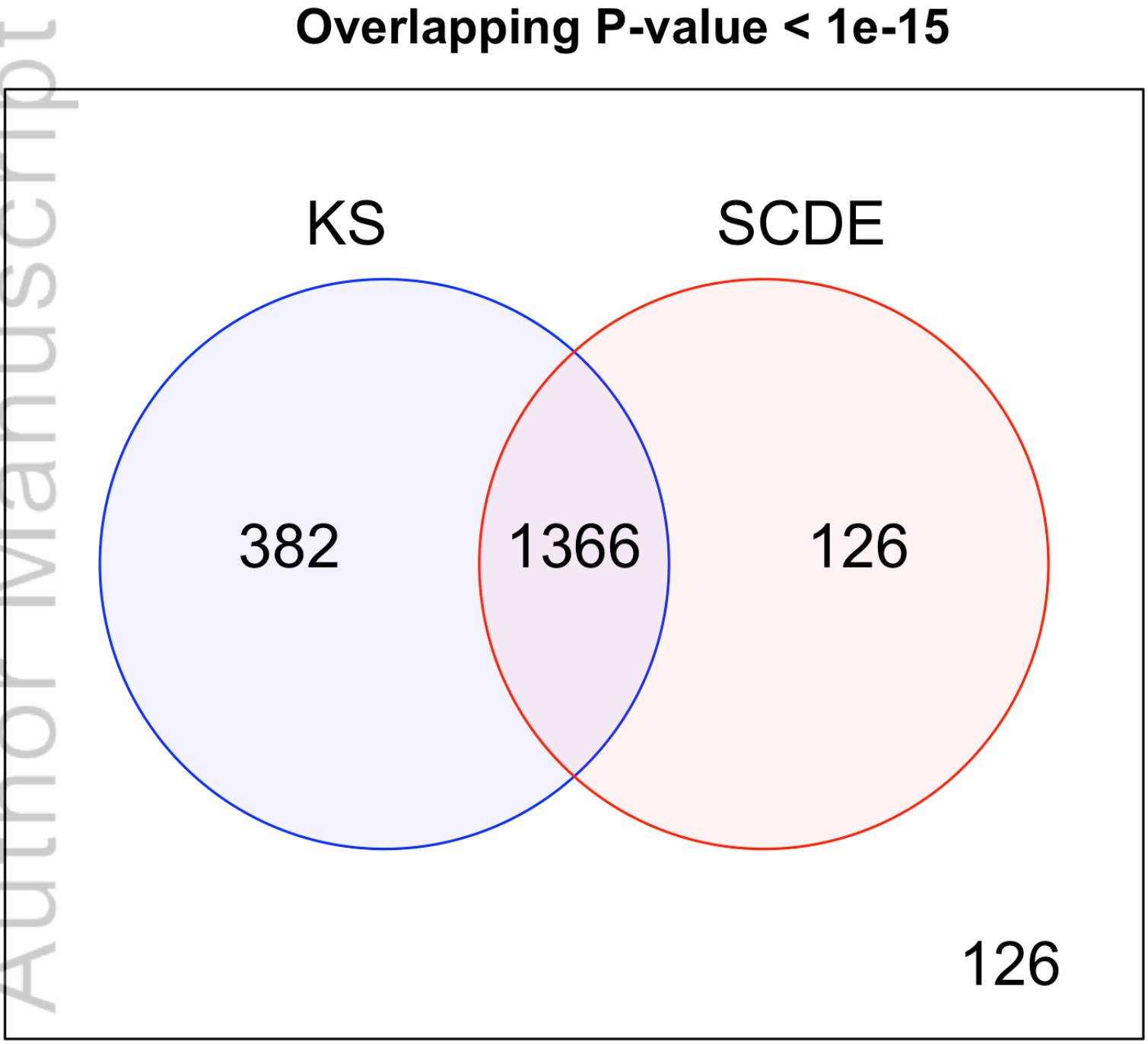
C3.png



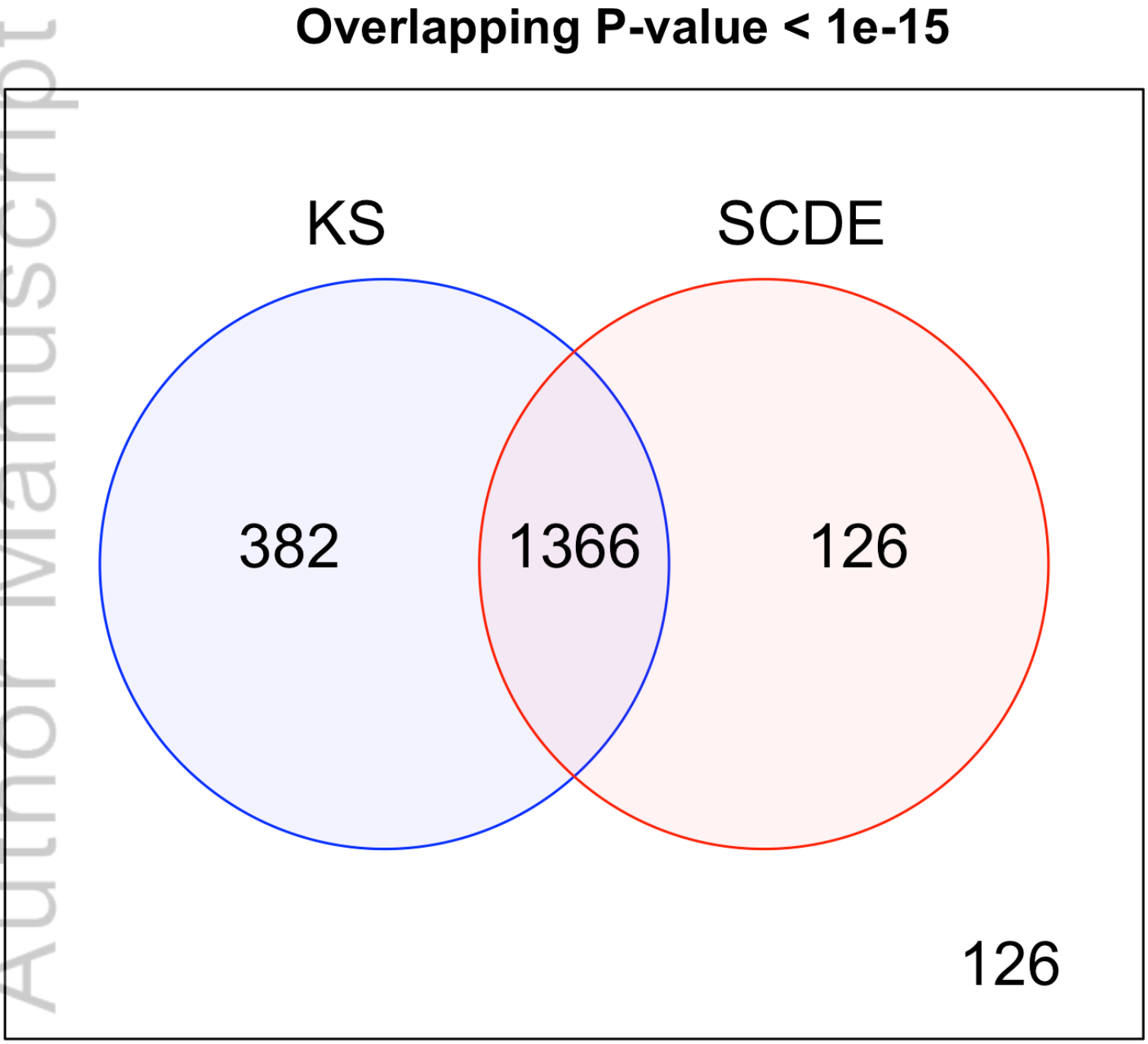
C4.png



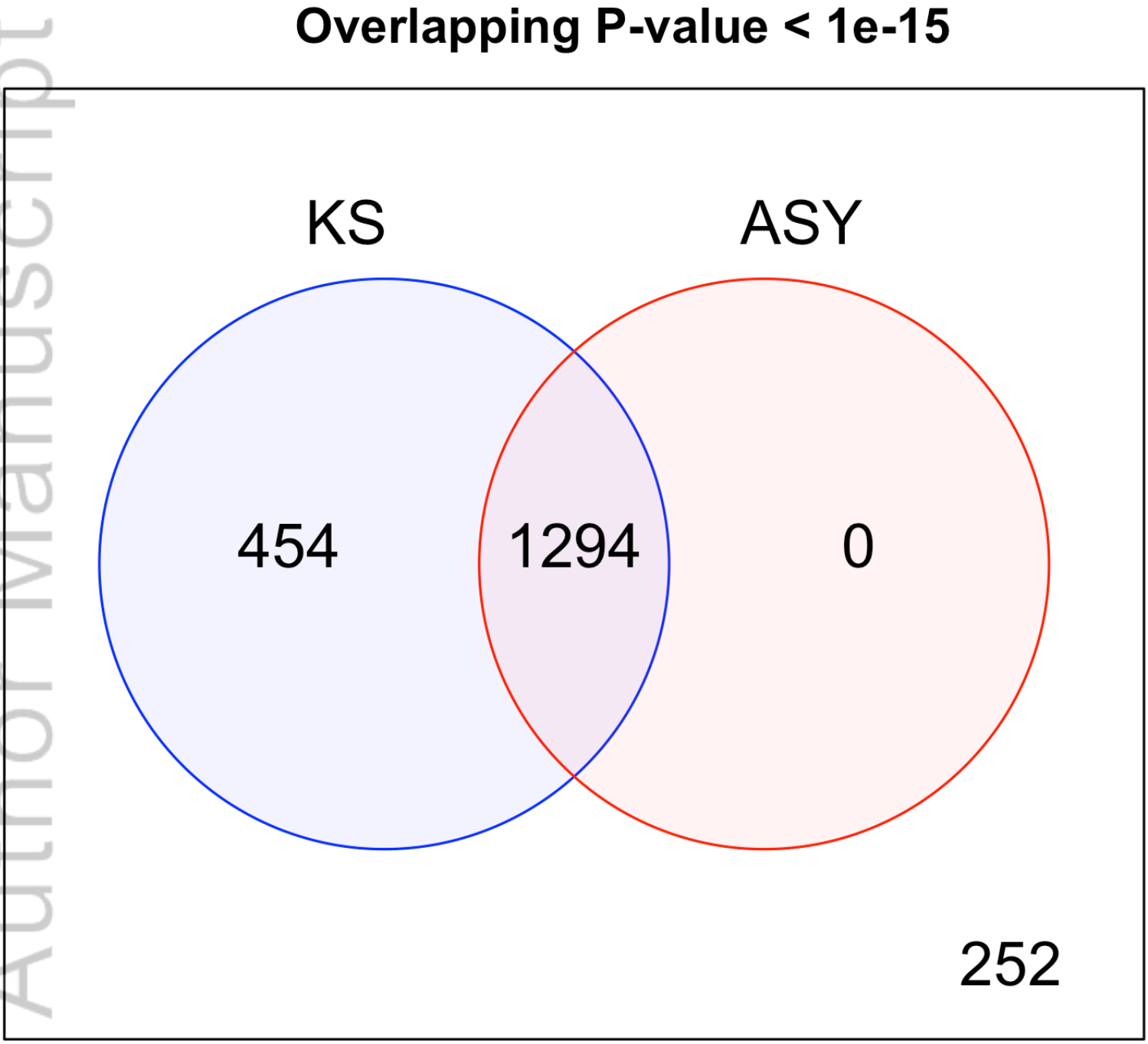
C4.png



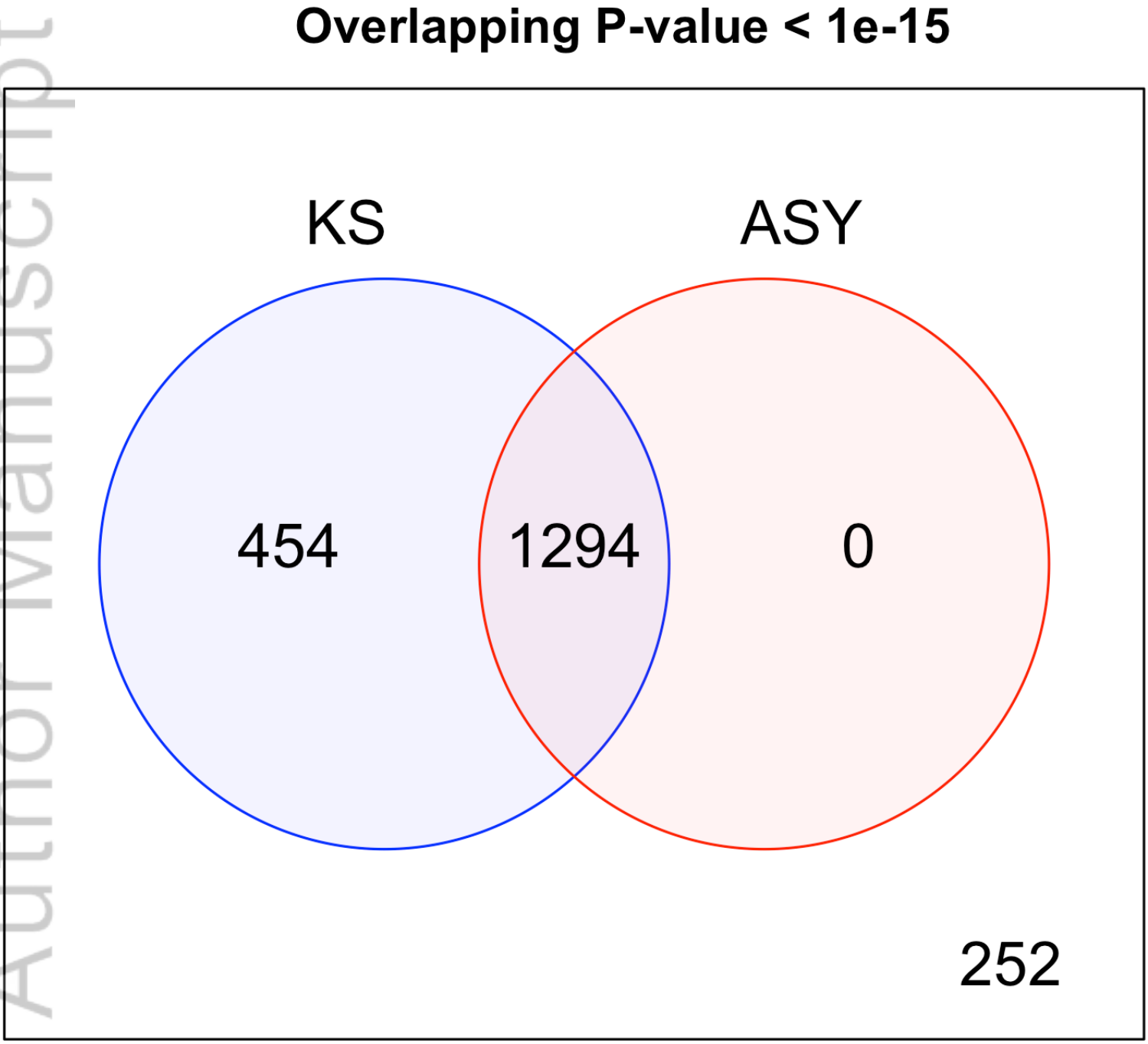
C5.png



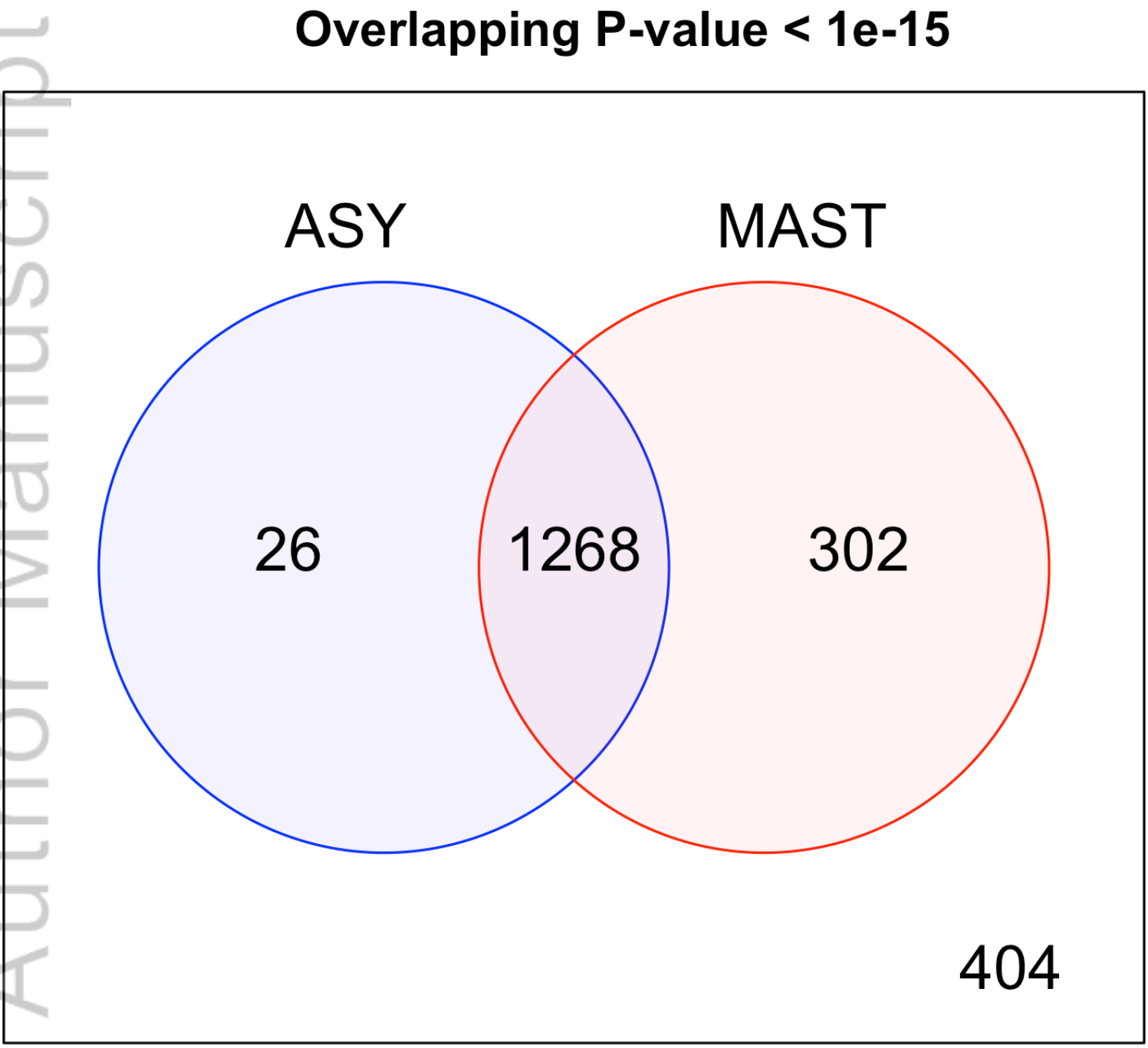
C5.png



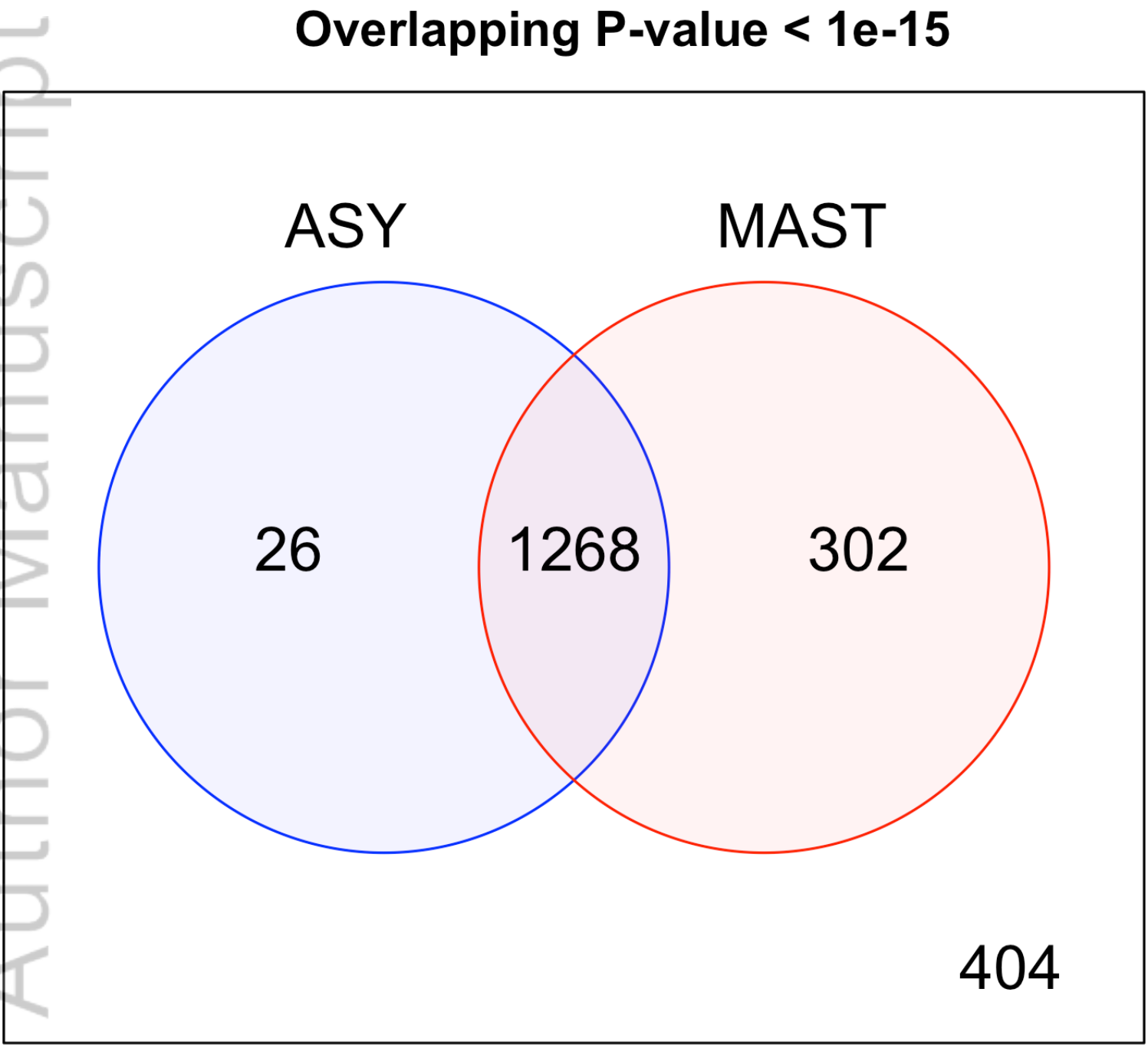
C6.png



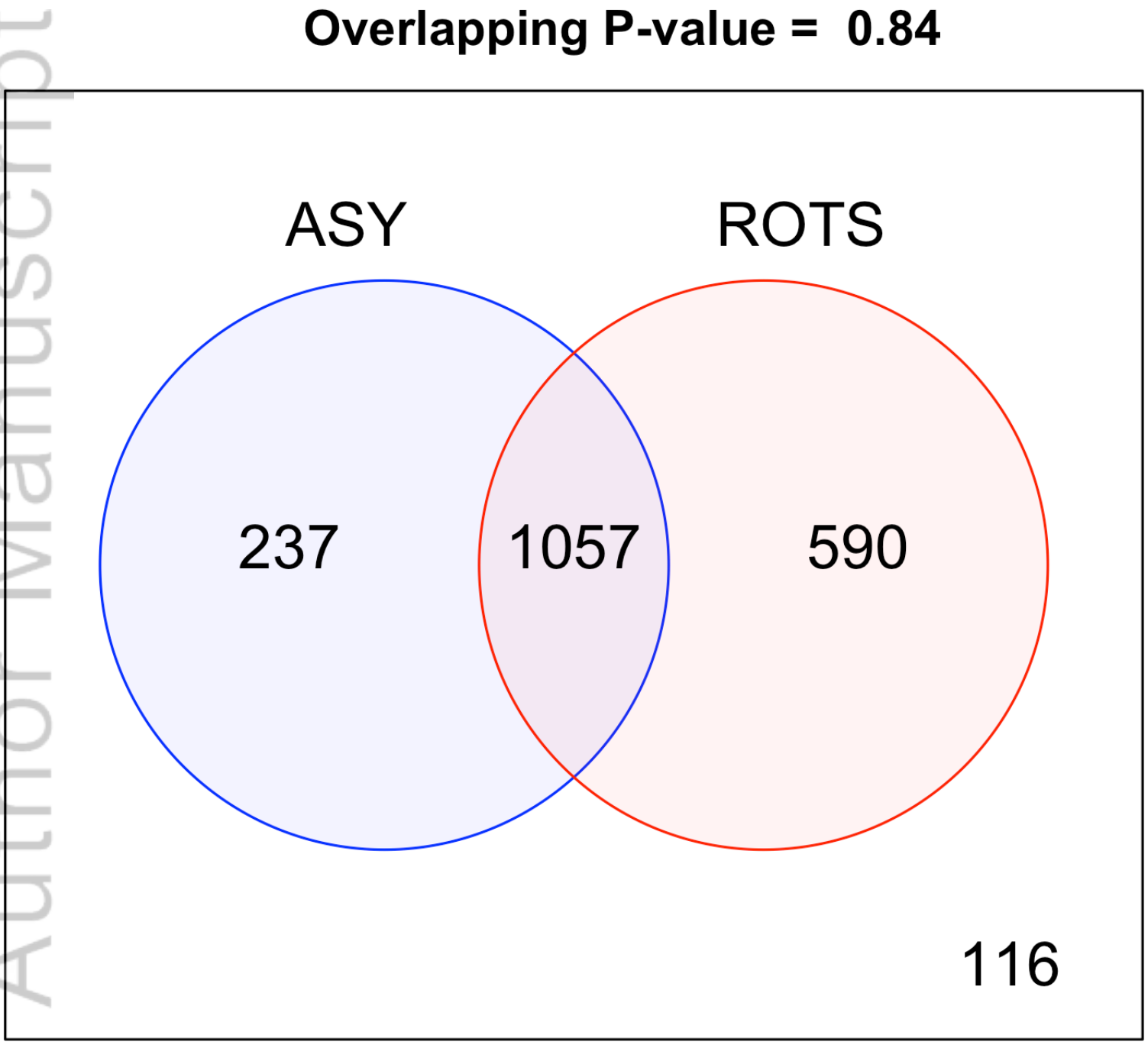
C6.png



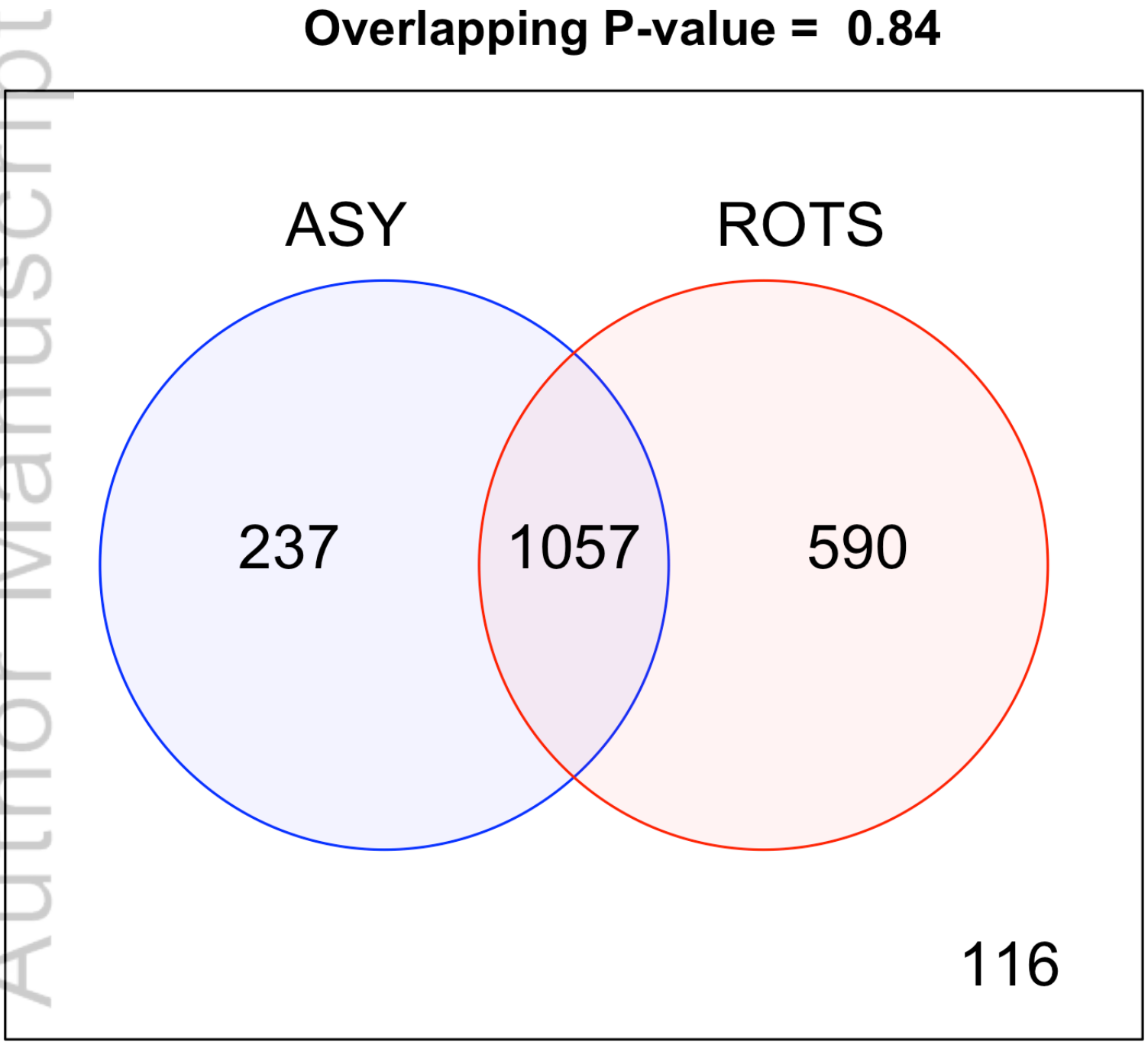
C7.png



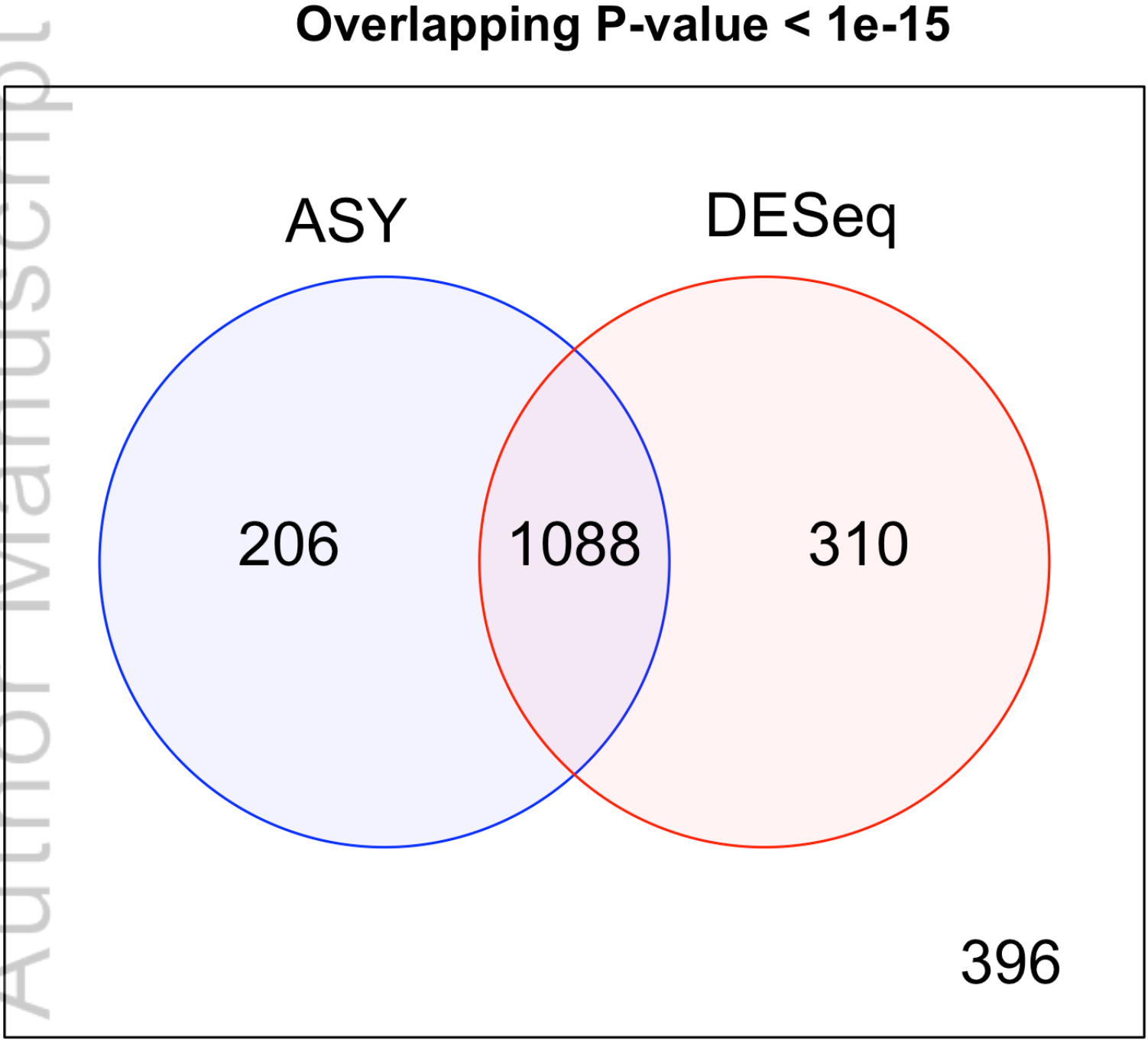
C7.png



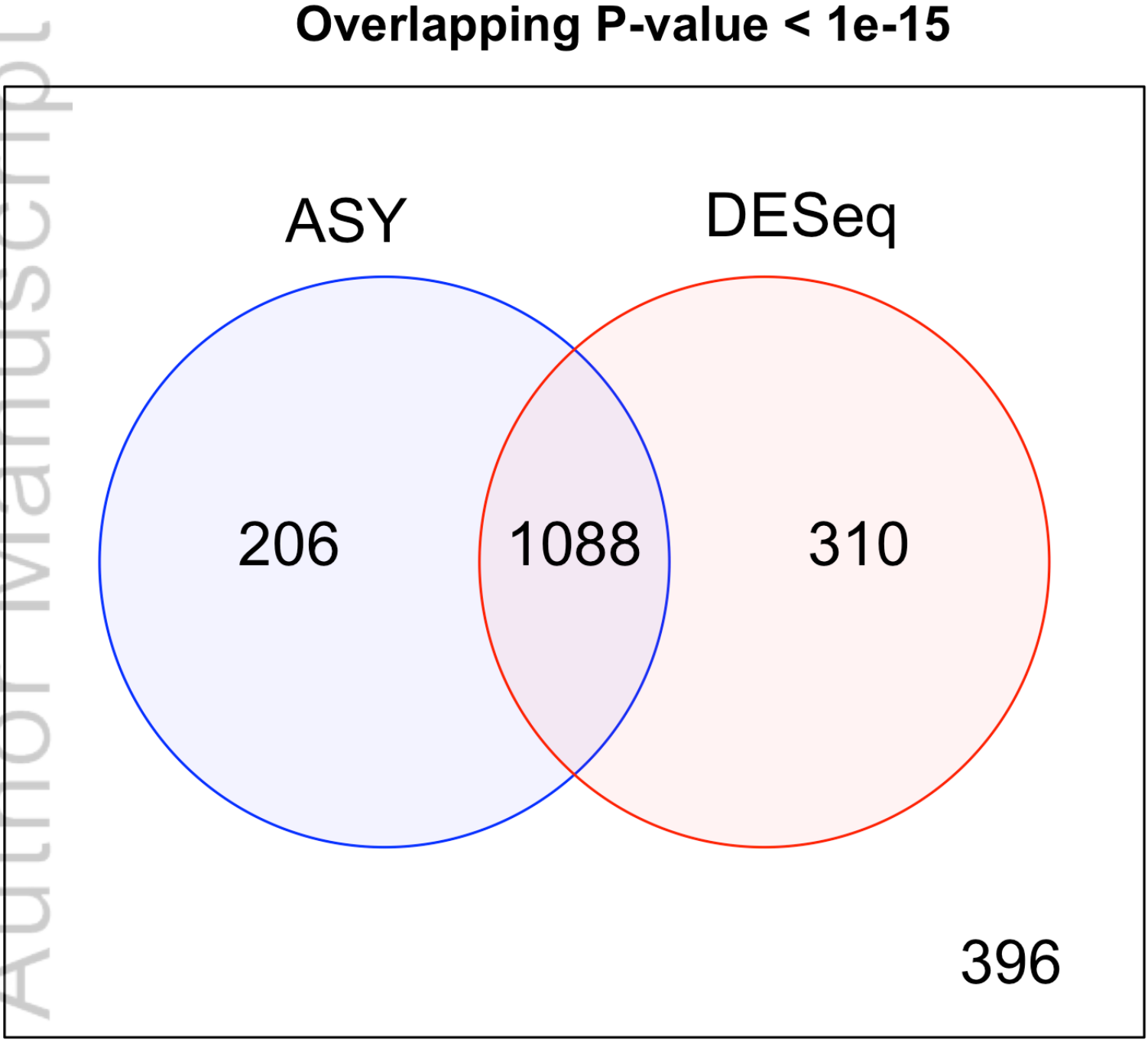
C8.png



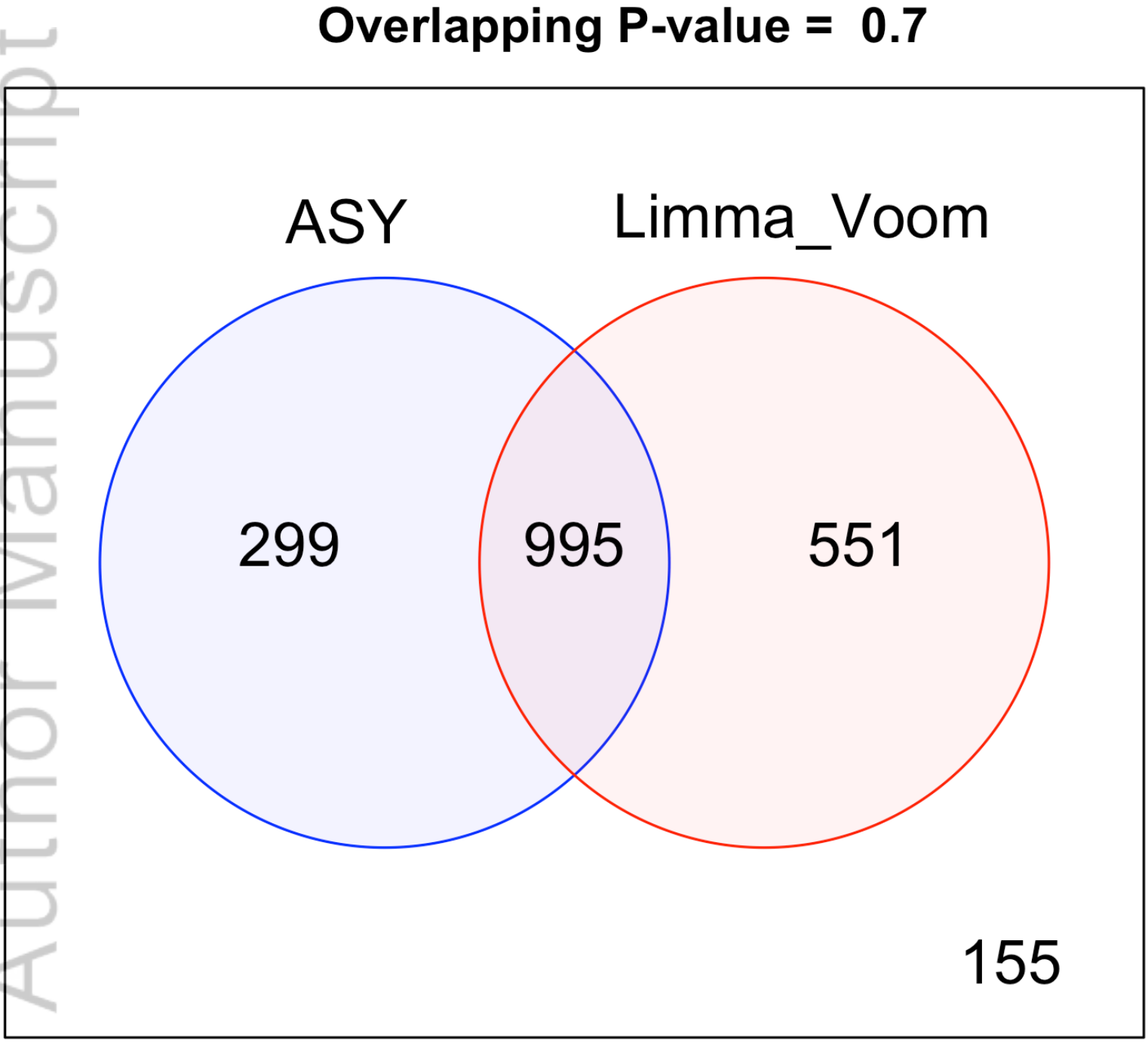
C8.png



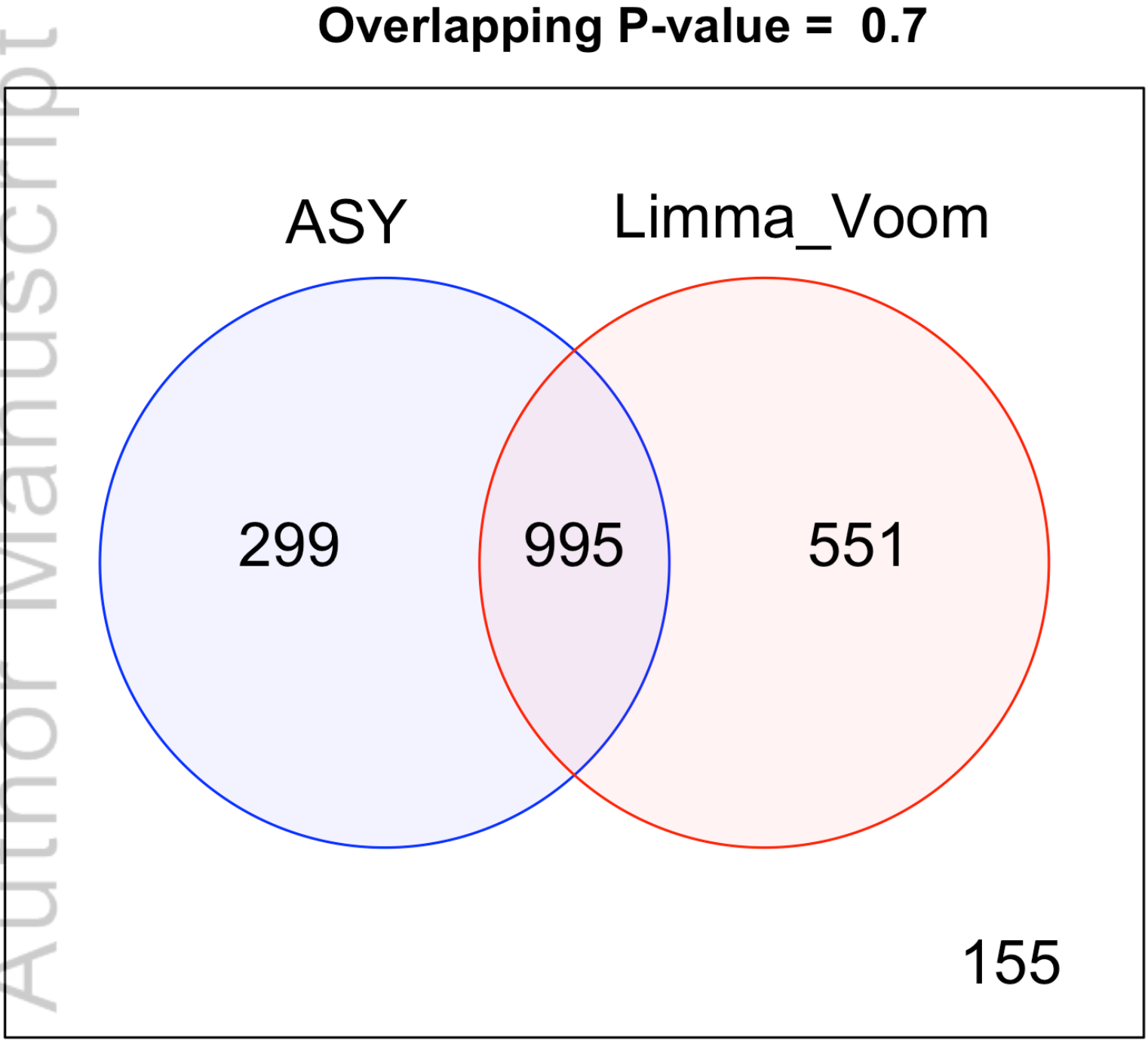
C9.png



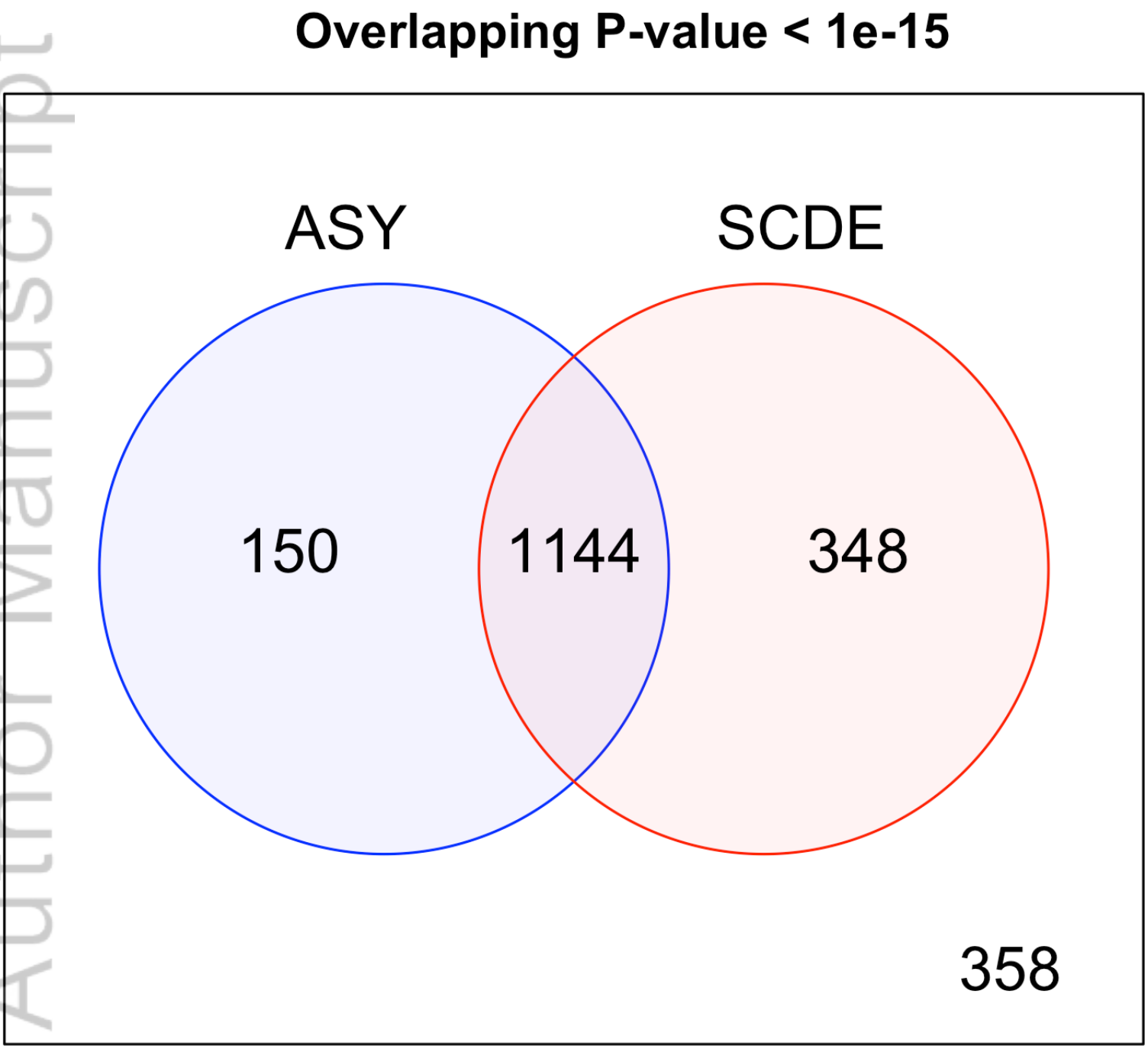
C9.png



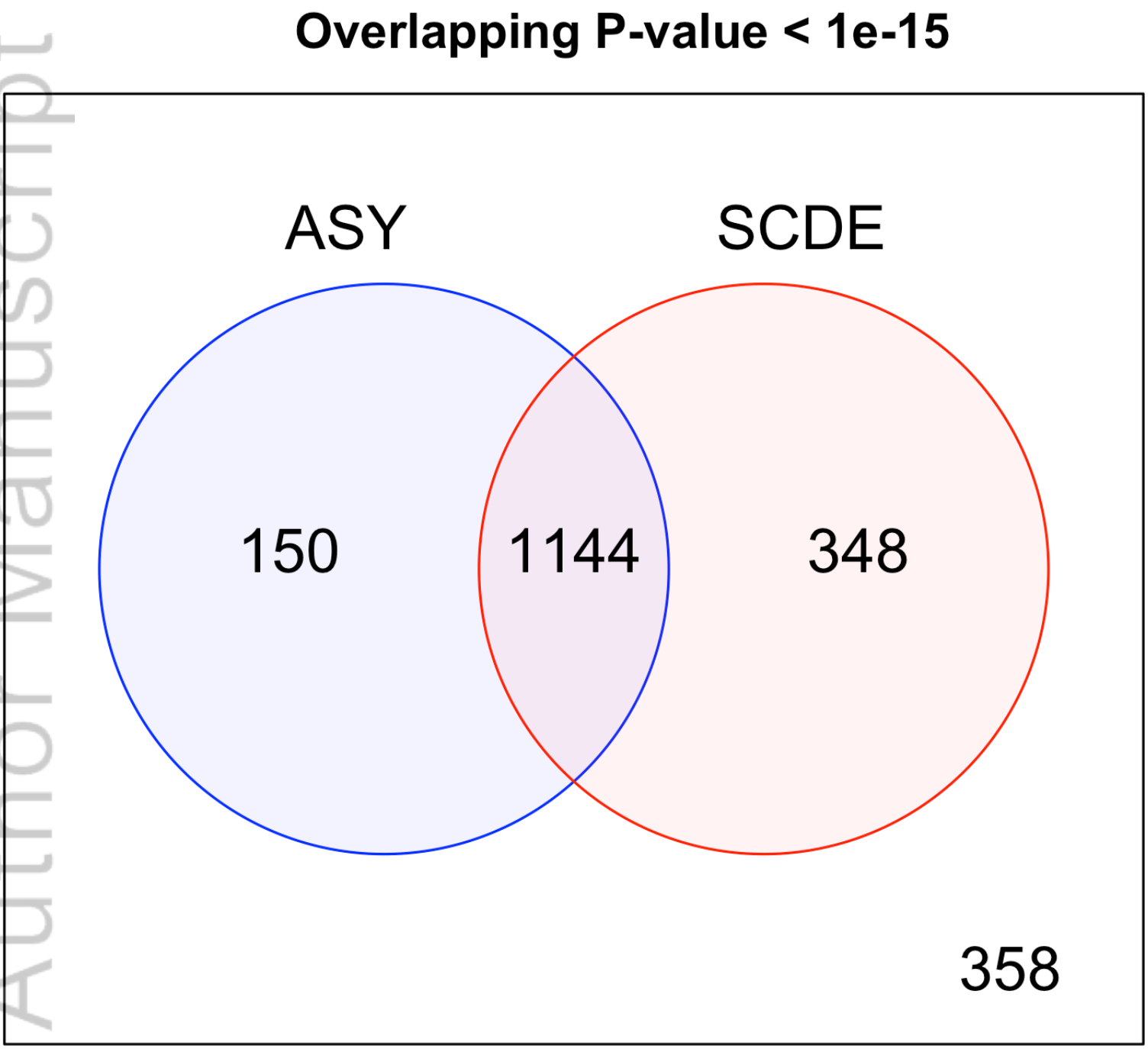
C10.png



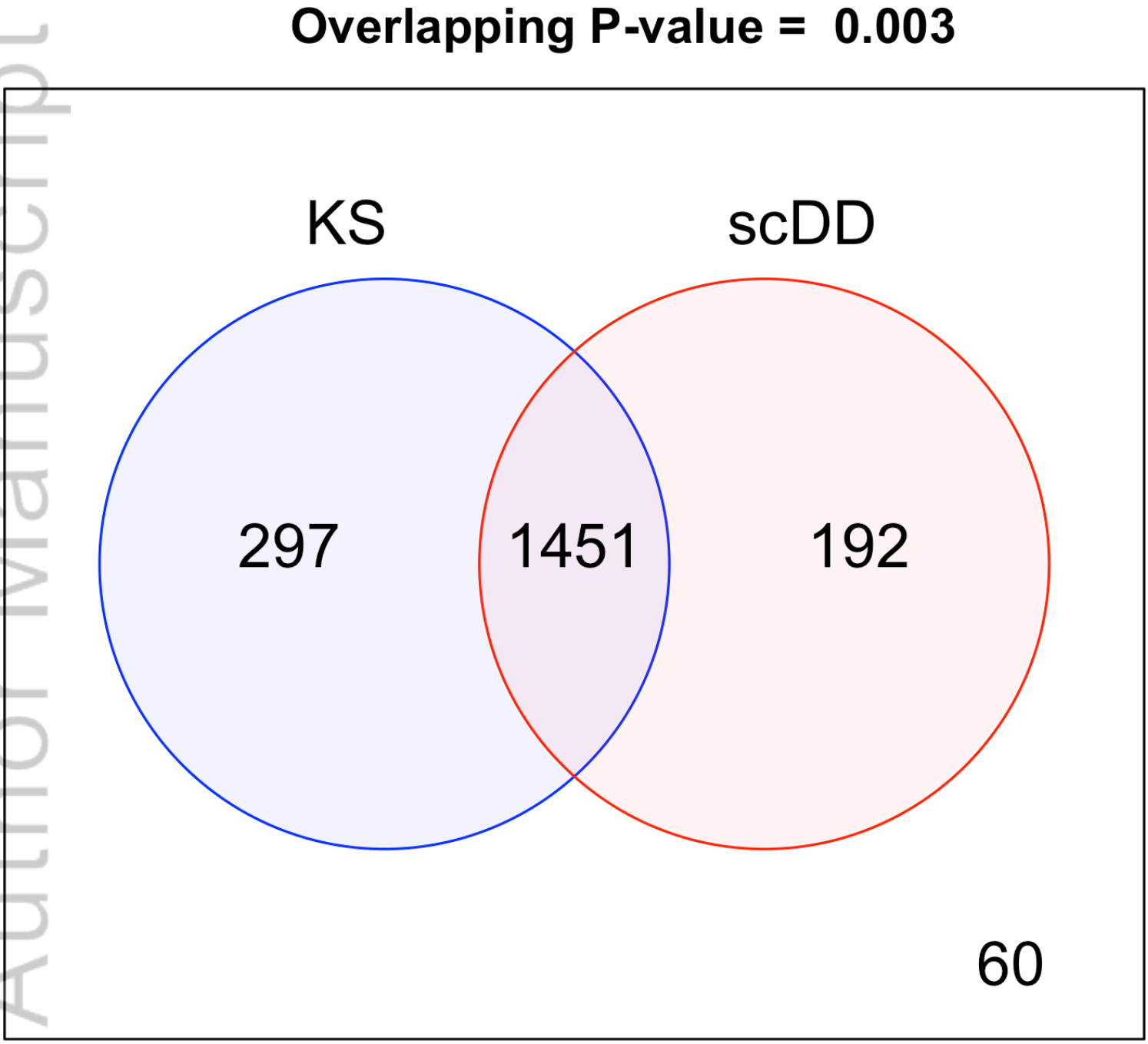
C10.png



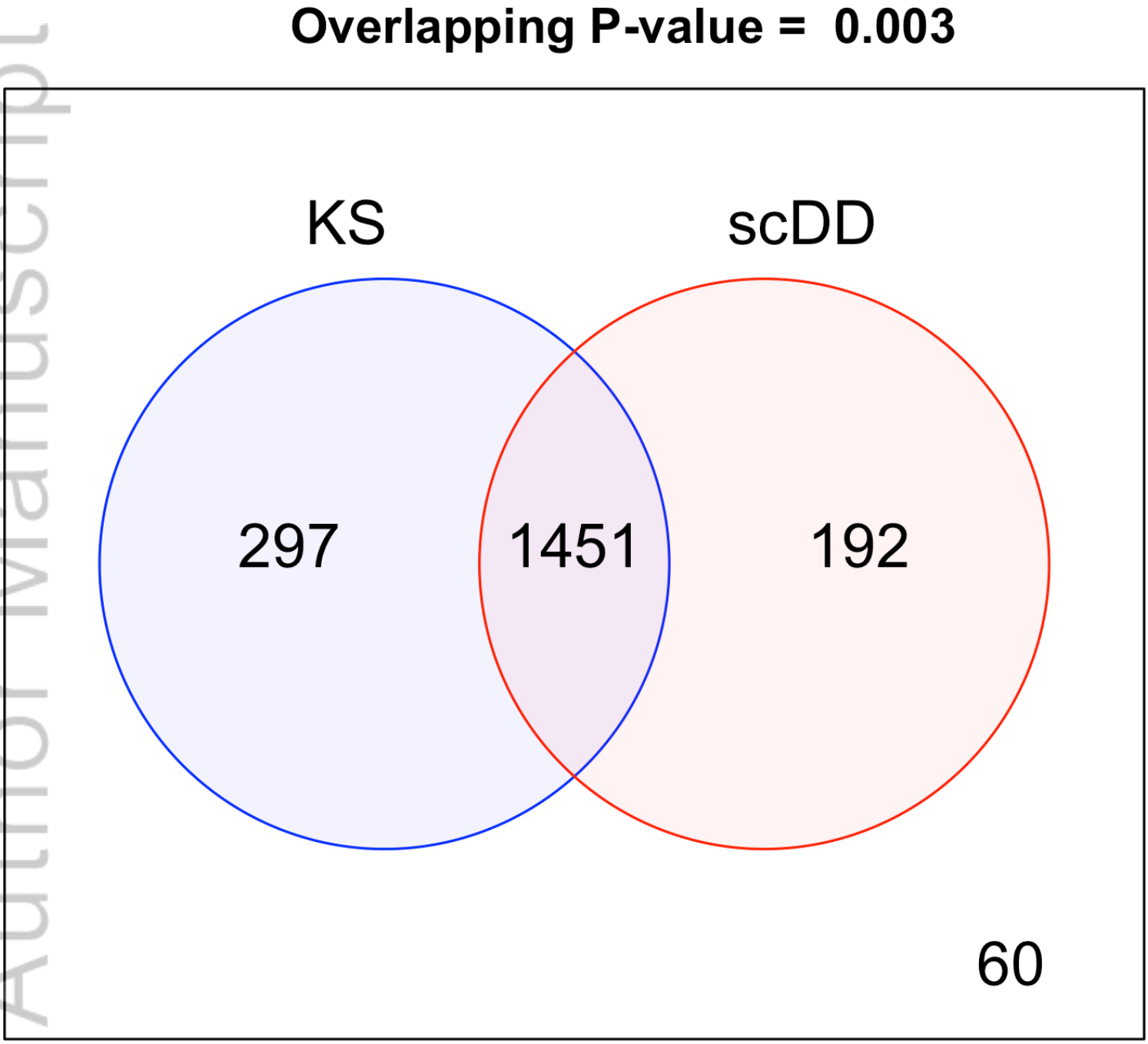
C11.png



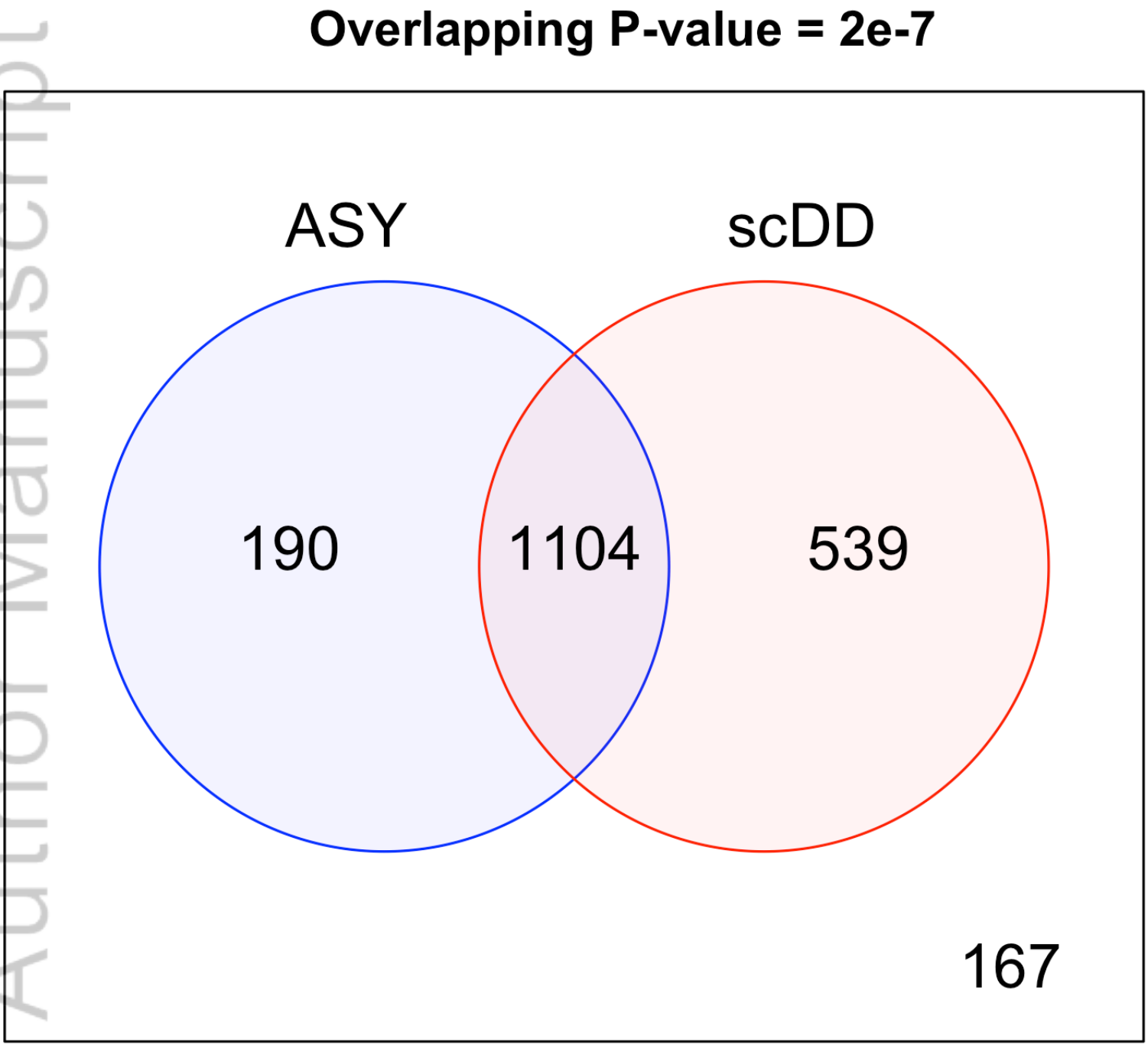
C11.png



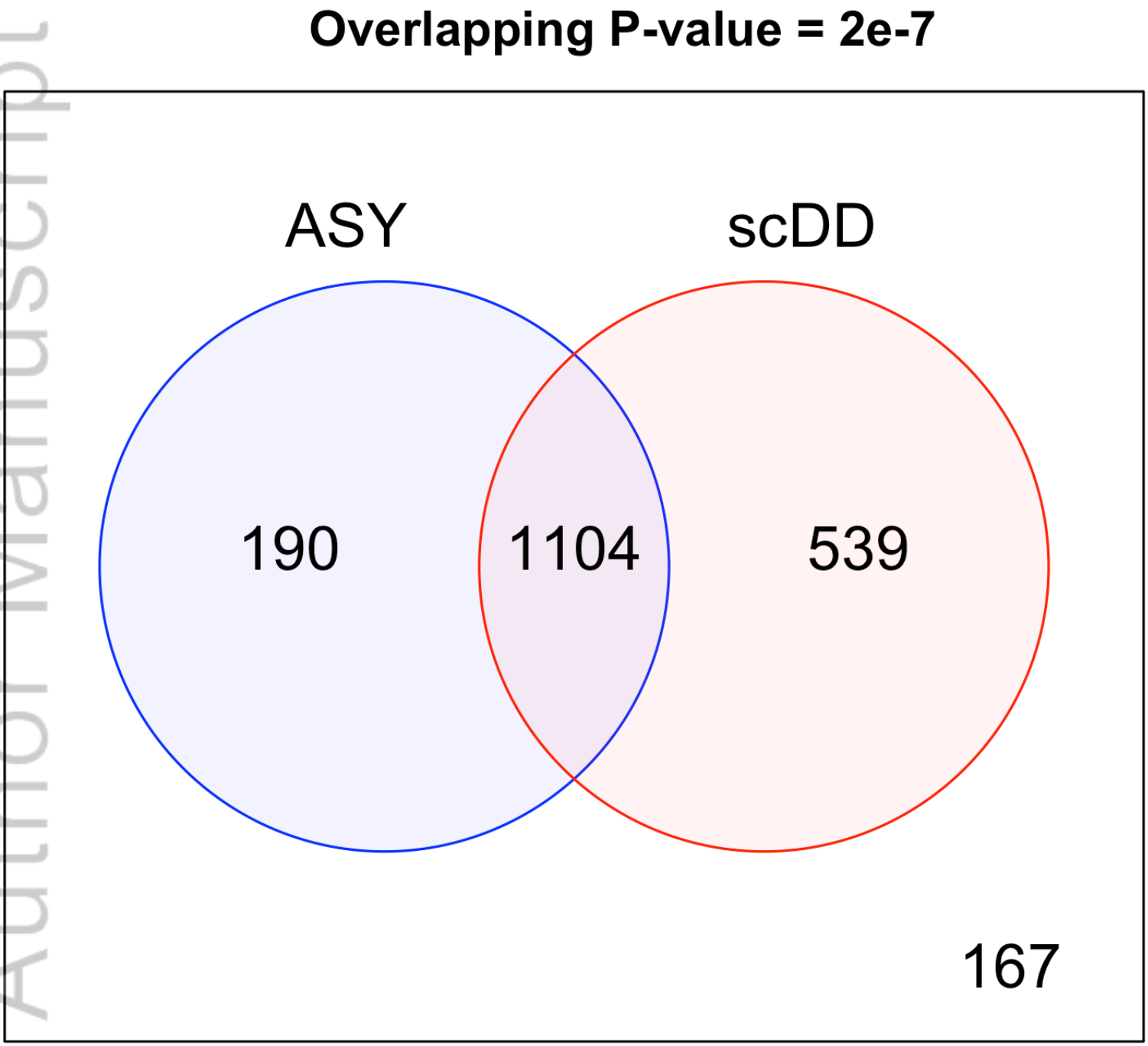
C12.png



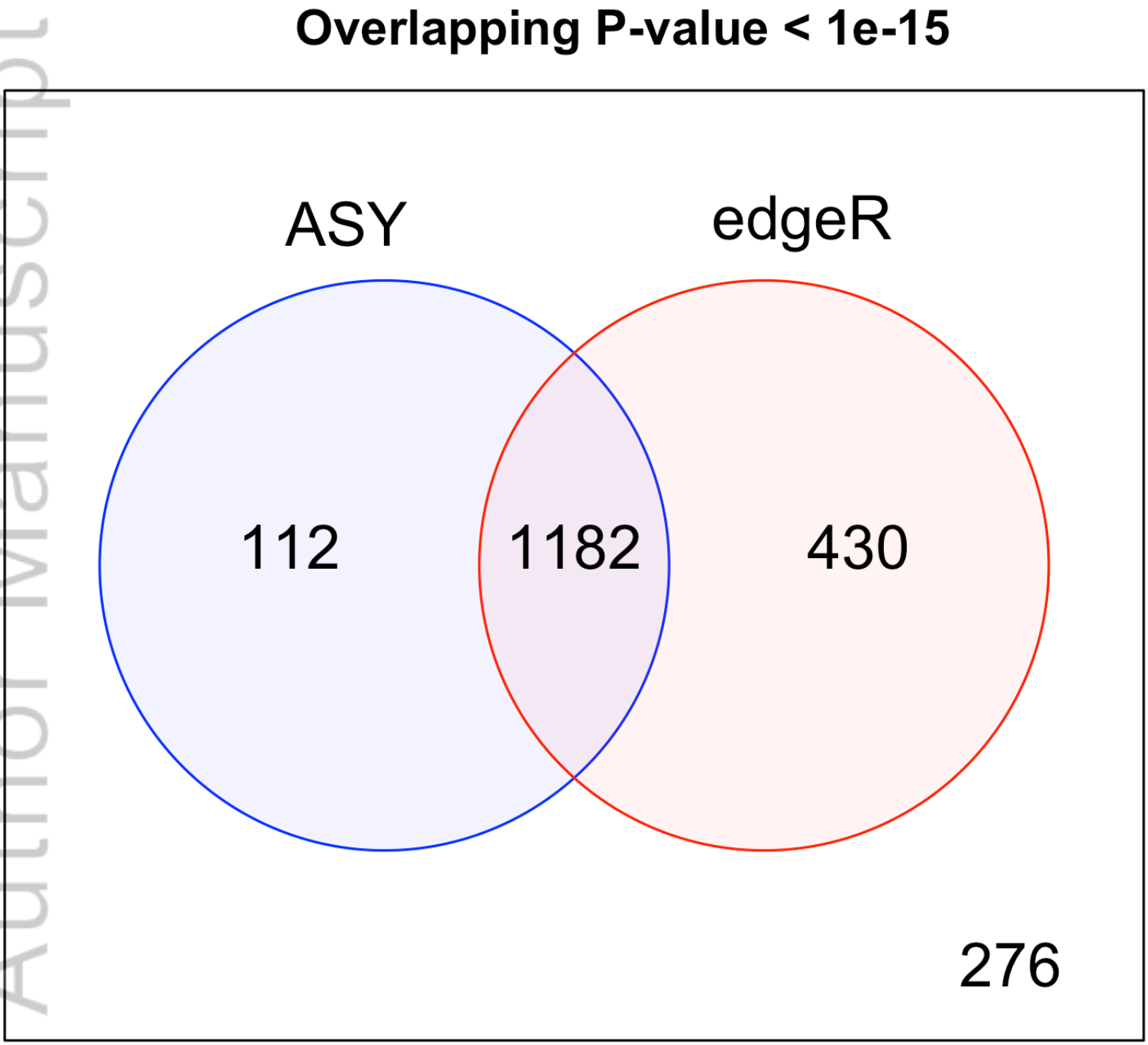
C12.png



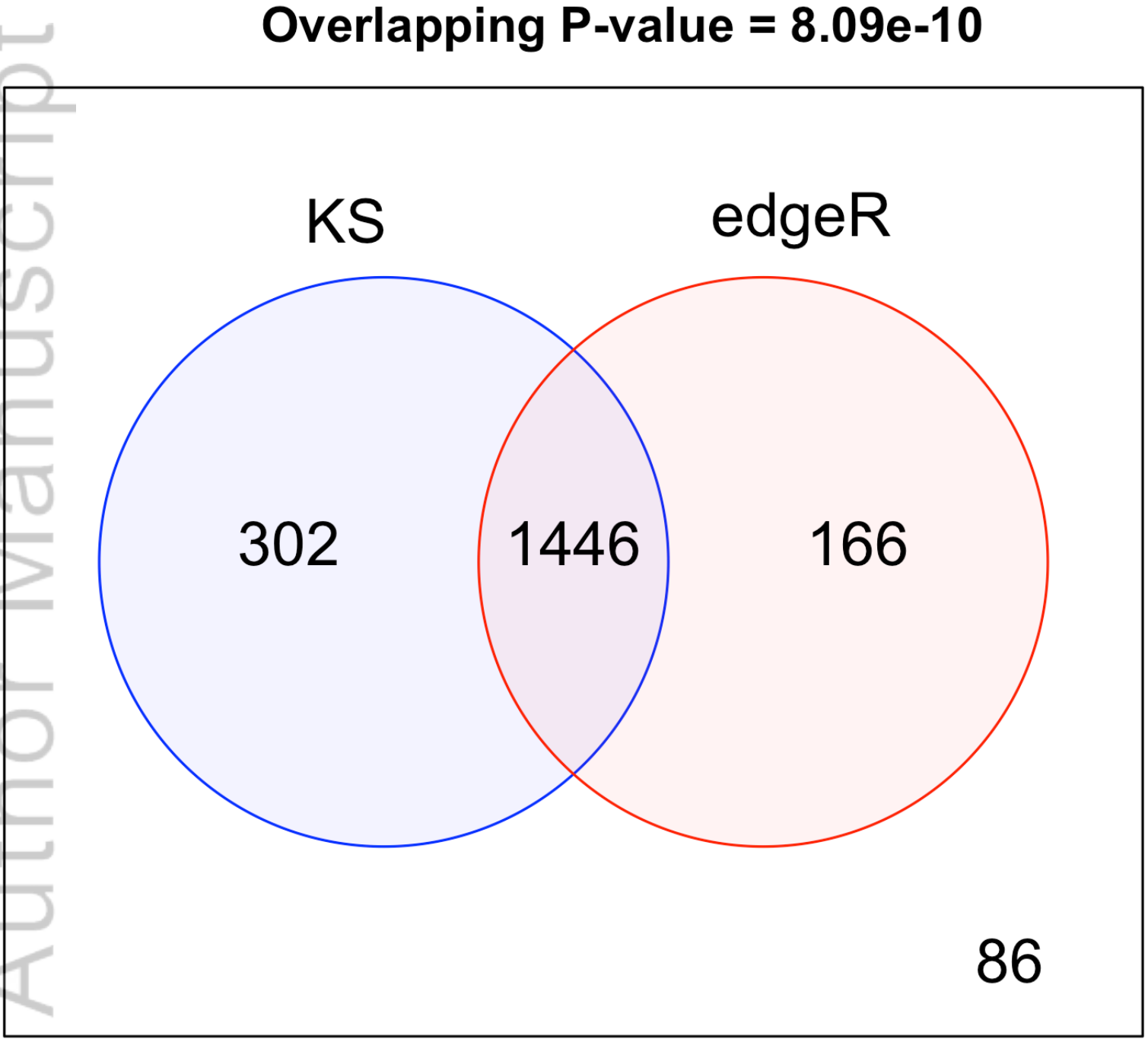
C13.png



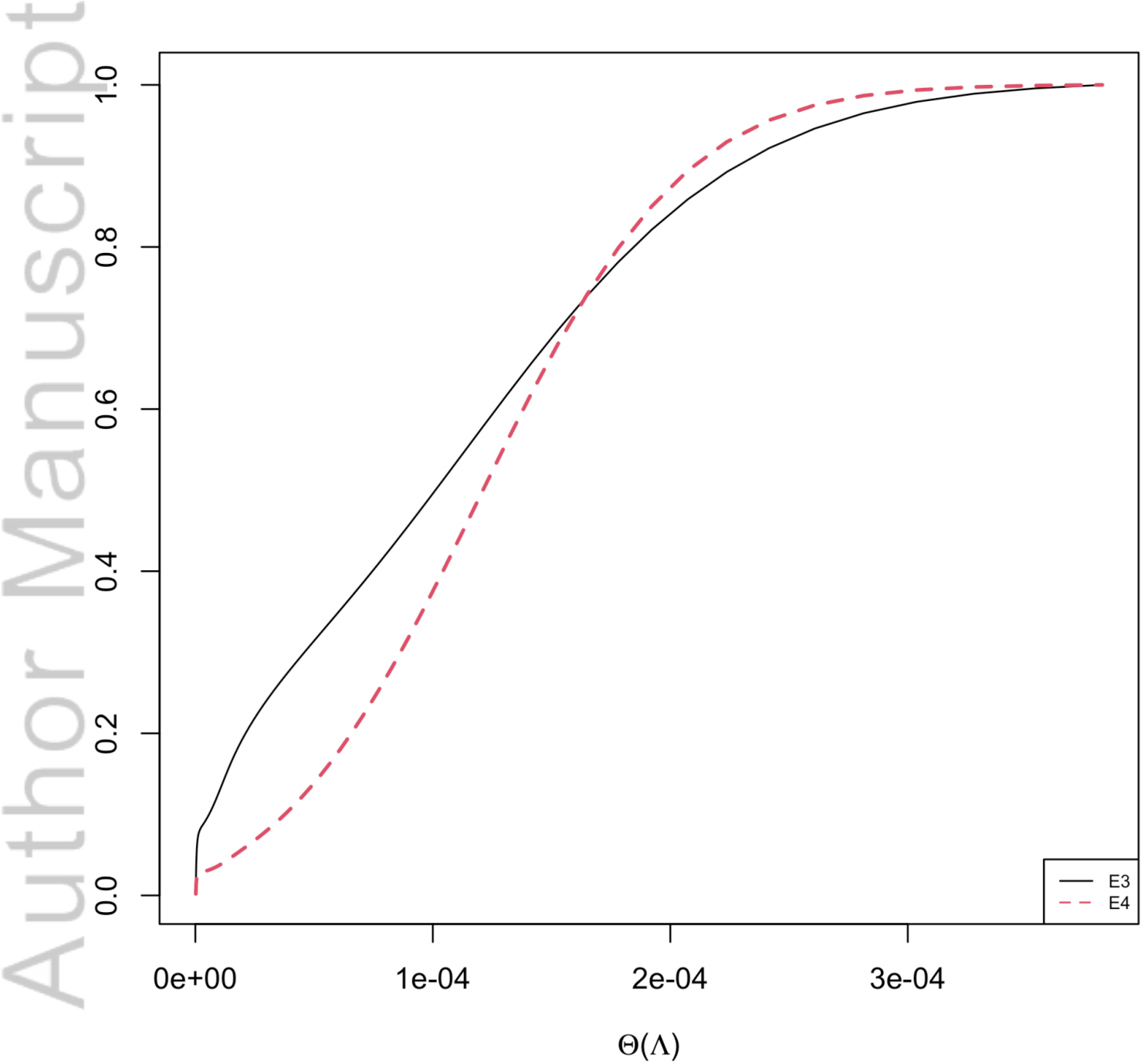
C13.png



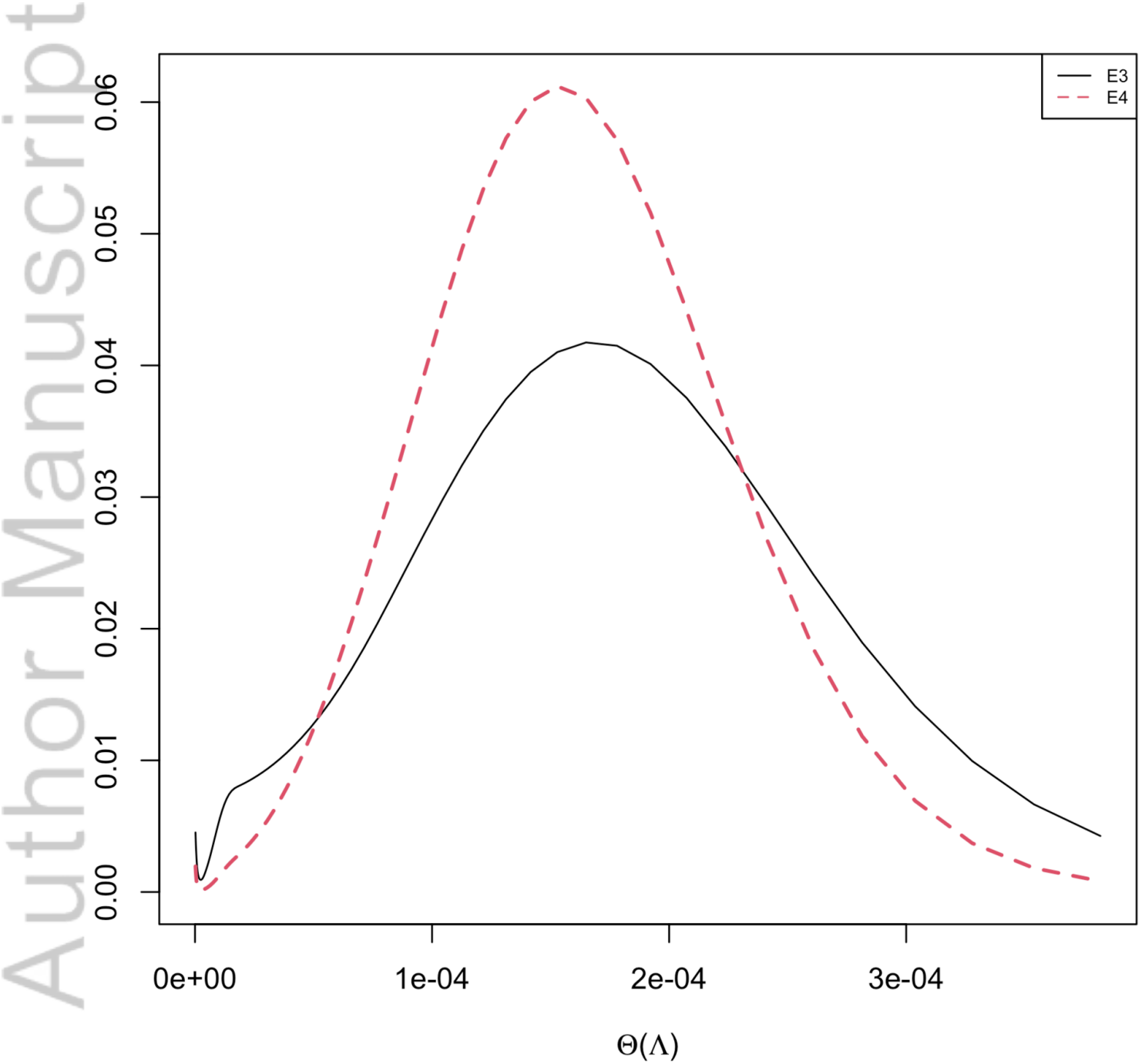
C14.png



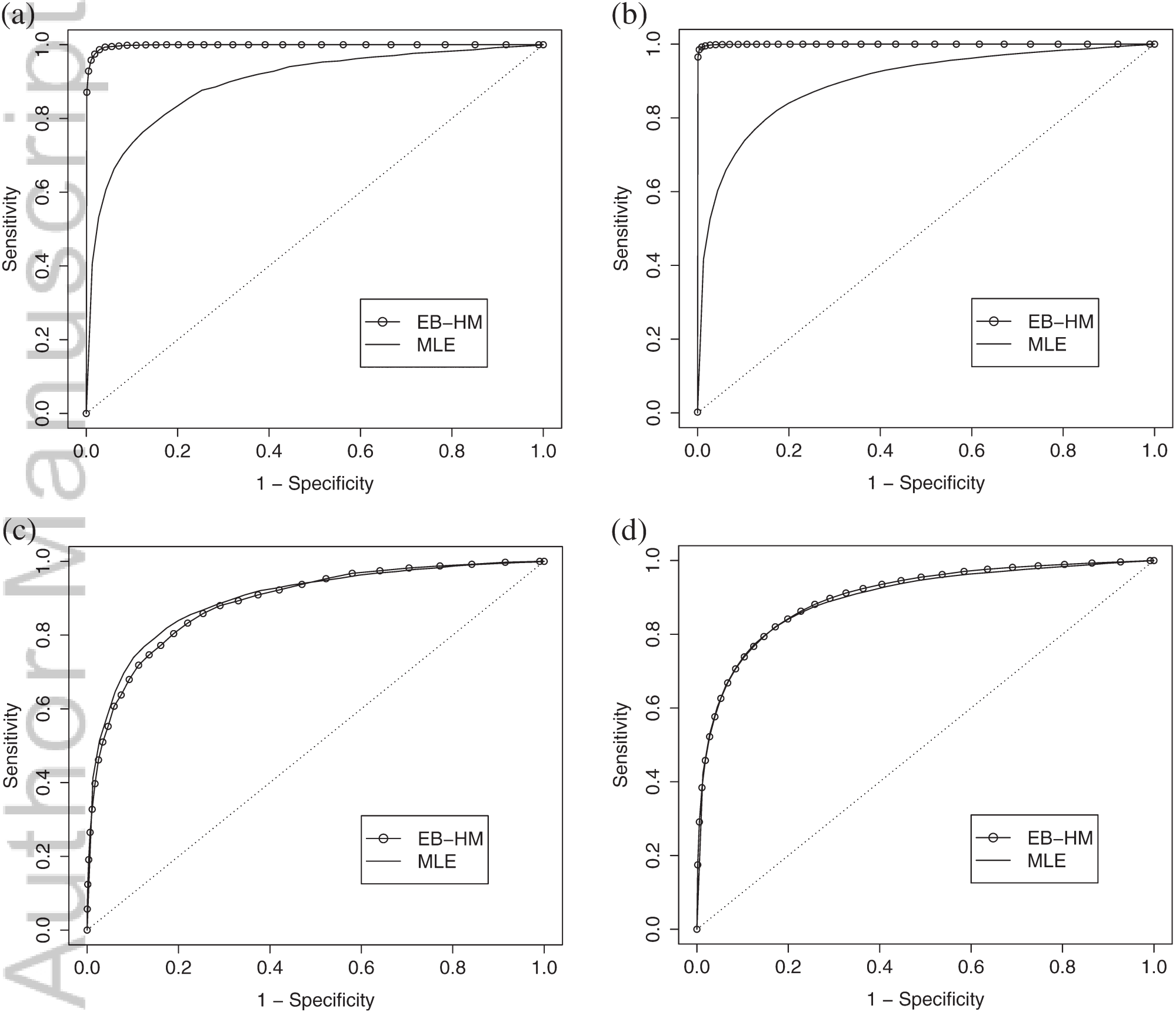
C15.png

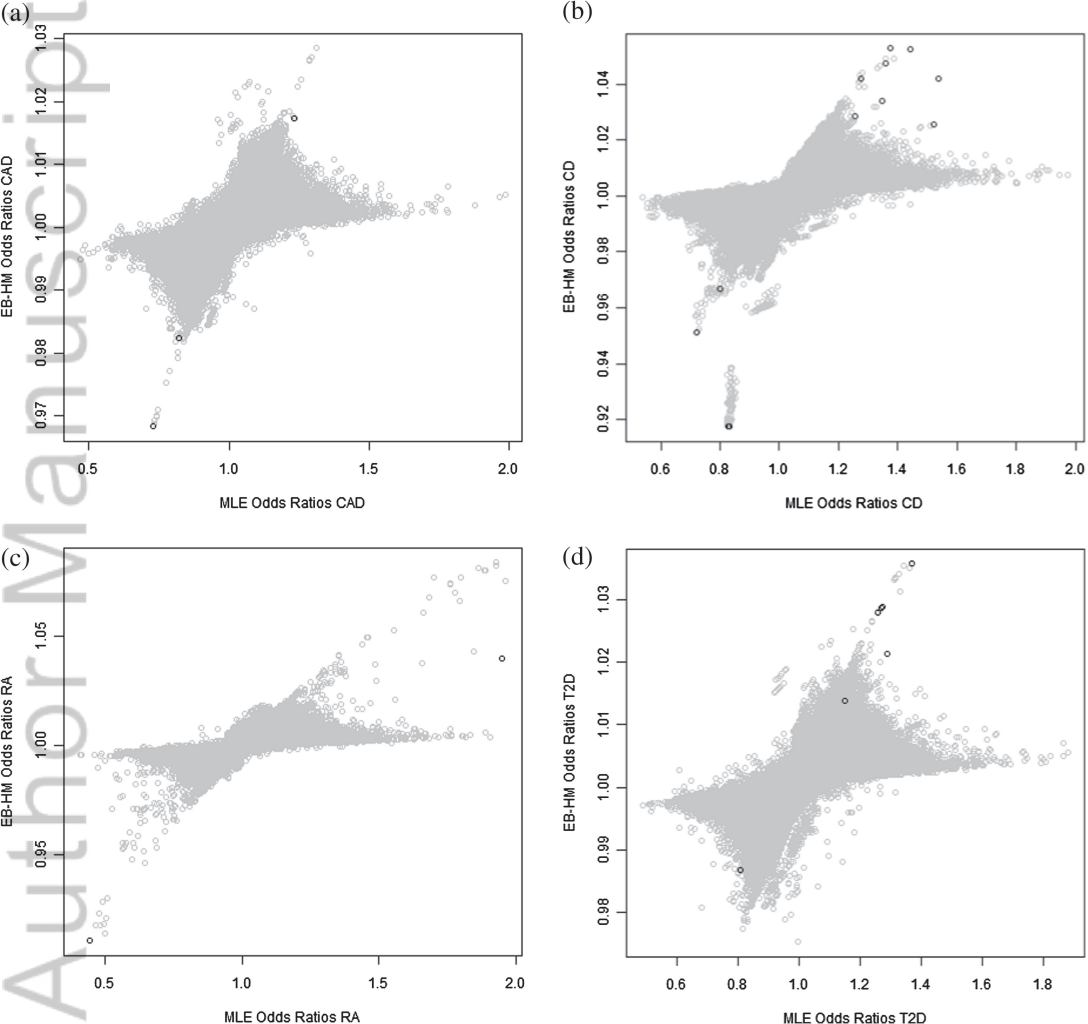


DIMT1 cdf.png



DIMT1 density estimation.png





Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Two-sample test with g-modeling and its applications

Jingyi Zhai | Hui Jiang

Department of Biostatistics, University of Michigan, MI, USA

Correspondence
Hui Jiang, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109. Email: jianghui@umich.edu

Summary

Many real data analyses involve two-sample comparisons in location or in distribution. Most existing methods focus on problems where observations are independently and identically distributed in each group. However, in some applications the observed data are not identically distributed but associated with some unobserved parameters which are identically distributed. To address this challenge, we propose a novel two-sample testing procedure as a combination of the g-modeling density estimation introduced by Efron and the two-sample Kolmogorov-Smirnov test. We also propose efficient bootstrap algorithms to estimate the statistical significance for such tests. We demonstrate the utility of the proposed approach with two biostatistical applications: the analysis of surgical nodes data with binomial models and differential expression analysis of single-cell RNA sequencing (scRNA-seq) data with zero-inflated Poisson model.

KEYWORDS:
Two-sample test, g-modeling; Bootstrap, Differential expression analysis, Single-cell RNA-seq, Zero-inflated Poisson.

1 | INTRODUCTION

Two-sample comparison occurs frequently in statistical analysis, for example, the testing of the drug effect between the control and the treatment groups in clinical trials. Relatively simple comparisons are often conducted to detect the difference in location between two groups where the parametric two-sample t-test or the non-parametric Wilcoxon rank-sum test is widely used. Comparing two samples in distribution is more challenging than comparing them in location. For more complicated and noisy data, the two-sample Kolmogorov-Smirnov (K-S) test is often used to detect the difference between two unknown distributions by comparing the empirical distributions of two groups of observations.[1] All the above tests and most other widely used statistical tests assume that the observations in each group are independently and identically distributed (i.i.d.). Unfortunately, such assumption may be violated in complex real-world problems and the problem becomes harder when the observations are no longer identically distributed. For instance, we may have two groups of independent samples where each observation follows the same type of distribution but with different underlying parameters such as different means. If we assume that these unknown parameters follow certain distributions, the objective of the two-sample comparison becomes the comparison of the distribution of the underlying parameters in the two groups. Specifically, consider the situation where there are two groups of observations $X_1, \ldots, X_{N_X}$ and $Y_1, \ldots, Y_{N_Y}$, where

$$\begin{aligned} X_i &\overset{ind}{\sim} p_i(X_i|\Theta_i), \quad Y_k \overset{ind}{\sim} p_k(Y_k|\Lambda_k), \quad i = 1, \ldots, N_X, \quad k = 1, \ldots, N_Y, \\ &\text{and} \qquad \Theta_i \overset{ind}{\sim} G, \quad \Lambda_k \overset{ind}{\sim} H, \end{aligned} \tag{1}$$

where Θ_i's and Λ_k's are two groups of unknown quantities drawn from two unknown distributions G and H, respectively, and p_i's and p_k's are some known parametric densities (or probability mass functions for discrete data). Our interest is in testing whether G equals H. Under this model, X_i's (and Y_k's) are independently but non-identically distributed, e.g., when X_i is binomial with parameters n_i and Θ_i, where n_i's are fixed but different quantities for different i's. In such case, the standard two-sample K-S test is no longer applicable.

Since we only observe X_i's and Y_k's but not Θ_i's and Λ_k's, we cannot compare the two unknown distributions G and H directly. The related one-sample problem, i.e., the problem of estimating G from X_i's (and similarly to estimate H from Y_k's), a.k.a. empirical Bayes deconvolution,[3] has been well-studied with many methods developed,[4–11] but largely they suffer from intensive computation and slow convergence.[12] Recently, a g-modeling approach was proposed by Efron for the efficient estimation of G (or H) in such case.[2]

To address the above two-sample comparison problem, we propose a novel two-sample testing procedure combining the g-modelling method for density estimation and the two-sample Kolmogorov-Smirnov (K-S) test statistic to detect for differences in distribution between two samples. We also develop efficient bootstrap algorithms to estimate the statistical significance for such tests. Our approach can be applied on a wide range of data types. We apply our approach on simulated data from a surgical nodes dataset under two different scenarios. In this application, the numbers of malignant satellite from patients are assumed to follow binomial distributions. Our test is shown to have high power to detect small difference between the two groups. Moreover, we also applied our test to the differential expression (DE) analysis on a real scRNA-seq dataset. Different from the original purpose of using g-modeling to model the test statistics across all the genes to controlling for multiple testing and false discovery rate,[2] here we model the read counts from each gene across all the samples to detect DE genes individually. Comparing with other existing DE methods, our test can detect more DE genes and has higher accuracy.

The rest of the paper is organized as follows: In Section 2, we introduce the notations and briefly review the g-modelling method for density estimation. We then introduce our proposed approach for two-sample tests in distribution. Section 3 presents two applications: the analysis of surgical nodes data with binomial models and DE analysis of scRNA-seq data with zero-inflated Poisson model. Section 4 concludes the paper with a discussion.

2 | METHODS

Since we build our approach for two-sample comparison based on the g-modeling method, we first review it briefly in the section below.

2.1 | One-sample density estimation with g-modeling

Starting with the one-sample density estimation problem based on the observations $X_1, \dots, X_N$, we follow the same setting as the g-modelling method,[2] where the sample space of Θ is discretized as $\boldsymbol{\tau} = (\theta_1, \dots, \theta_m)$ for computational convenience. The g-modelling framework further assumes that Θ follows a semi-parametric exponential family distribution as follows:

$$Pr(\Theta = \theta_j) = g_j(\boldsymbol{\alpha}) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha} - \phi(\boldsymbol{\alpha})\}, \quad j = 1, \dots, m,$$

where $\boldsymbol{\alpha}$ is a p-dimensional vector of parameters, $\boldsymbol{Q}$ is a fixed and known $m \times p$ matrix taken as the design matrix from natural spline basis,[2] $\boldsymbol{Q}_j$ is the j-th row of $\boldsymbol{Q}$ (as a p-dimensional column vector), and the normalization term $\phi(\boldsymbol{\alpha})$ is $\phi(\boldsymbol{\alpha}) = \log \sum_{j=1}^m \exp(\boldsymbol{Q}_j^T \boldsymbol{\alpha})$. Conditional on Θ_i, the observed X_i follows a known parametric distribution as $X_i \overset{ind}{\sim} p_i(X_i|\Theta_i)$, for $i = 1, \dots, N$, and we define $p_{ij} = p_i(X_i = x_i|\Theta_i = \theta_j)$. Then the marginal probability of X_i and log-likelihood of the observed data can be computed as $Pr(X_i = x_i) = f_i(\boldsymbol{\alpha}) = \sum_{j=1}^m p_{ij} g_j(\boldsymbol{\alpha})$, and $l_i(\boldsymbol{\alpha}) = \log f_i(\boldsymbol{\alpha})$, respectively. Here we assume discrete X_i (or discretized X_i if it was continuous). In order to improve the accuracy for estimation, the log-likelihood is regularized with a ℓ_2 penalty term. Hence, the objective function for maximum likelihood estimation is $m(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - s(\boldsymbol{\alpha})$, where $s(\boldsymbol{\alpha}) = c_0 \|\boldsymbol{\alpha}\|$ with c_0 being a tuning parameter which we take as 1. Now $\boldsymbol{\alpha}$ can be estimated by maximizing the above penalized log-likelihood. We denote the maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}}$ and obtain the estimated $g_j(\boldsymbol{\alpha})$ by $g_j(\hat{\boldsymbol{\alpha}})$ for $j = 1, \dots, m$.

In practice, we find that discretized Θ may introduce numerical problems, especially when we are dealing with scRNA-seq data where the read counts are sparse and have a large range. For instance, it could happen that $p_{ij} = 0$ for all $j = 1, \dots, m$, which leads infinite log-likelihood. Therefore, we make an adjustment in the calculation of p_{ij} to avoid such problem. For each

x_i, we estimate its MLE $\hat{\theta}_i^*$, and calculate p_{ij} for the θ_j that is the closest to $\hat{\theta}_i^*$ as $p_{ij} = p_i(X_i = x_i|\Theta_i = \hat{\theta}_i^*)$. For all other j, we still calculate p_{ij} as $p_{ij} = p_i(X_i = x_i|\Theta_i = \theta_j)$.

2.2 | Two-sample comparison based on g-modeling

Here we propose a two-sample test built on the above one-sample density estimation procedure using g-modelling. With the notations in (1), assume Θ_i and Λ_k have the same discretized sample space $\tau = (\theta_1, \dots, \theta_m)$, $g(\boldsymbol{\alpha}_X)$ and $h(\boldsymbol{\alpha}_Y)$ are the semi-parametric distributions for G and H respectively, and have the following form

$$\begin{aligned} Pr(\Theta_i = \theta_j) &= g_j(\boldsymbol{\alpha}_X) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_X - \phi(\boldsymbol{\alpha}_X)\}, \text{ for } j = 1, \dots, m, \\ Pr(\Lambda_k = \theta_j) &= h_j(\boldsymbol{\alpha}_Y) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_Y - \phi(\boldsymbol{\alpha}_Y)\}, \text{ for } j = 1, \dots, m, \end{aligned} \tag{2}$$

where $\boldsymbol{Q}$ is a fixed and known $m \times p$ structure matrix for both g and h, and $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_Y$ are parameters that can be estimated with g-modeling using the data from the two groups, respectively. To test for the difference between the two distributions G and H against the null hypothesis $H_0 : G = H$, we use the two-sample K-S test statistic which can be calculated as

$$T = \max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, \quad j = 1, \dots, m, \tag{3}$$

where $\hat{G}_j$ and $\hat{H}_j$ are the values of estimated cumulative distribution functions (CDFs) of G and H evaluated at θ_j, respectively.

Different from the generalized linear mixed model introduced originally for two-sample comparison when changes in location between two distributions are of concern,[2] our two-sample test can detect changes not only in location but also in distribution.

2.3 | Simple bootstrap procedure for p-value estimation

To estimate the statistical significance for the test statistic T defined in (3), we can use a simple parametric bootstrap procedure to directly simulate the null distribution of T. Under the null hypothesis, G and H are identical. Therefore, we first pool the two groups together and employ the g-modeling approach to obtain a pooled density estimate $\hat{g}(\hat{\boldsymbol{\alpha}}_p)$. Then, for the b-th bootstrap iteration, $b = 1, \dots, B$, we take the following steps:

1. Sample $\Theta_i^{(b)}, i = 1, \dots, N_X$, and $\Lambda_k^{(b)}, k = 1, \dots, N_Y$, with respect to $\hat{g}(\hat{\boldsymbol{\alpha}}_p)$.
2. Sample $X_i^{(b)}$ from $p_i(X_i|\Theta_i^{(b)}), i = 1, \dots, N_X$, and $Y_k^{(b)}$ from $p_k(Y_k|\Lambda_k^{(b)}), k = 1, \dots, N_Y$.
3. Estimate $\hat{G}^{(b)}$ from $X_1^{(b)}, \dots, X_{N_X}^{(b)}$ and $\hat{H}^{(b)}$ from $Y_1^{(b)}, \dots, Y_{N_Y}^{(b)}$ using g-modeling.
4. Calculate $T^{(b)} = \max_j |\hat{G}_j^{(b)} - \hat{H}_j^{(b)}|, j = 1, \dots, m$.

Finally, we estimate the p-value as $\hat{p} = (\sum_{b=1}^{B} 1_{T^{(b)} \geq T} + 1)/(B+1)$, where we add one to both the numerator and the denominator to avoid a p-value of zero.

Since MLE problems need to be solved for each bootstrap iteration, the above simple bootstrap procedure may be computationally intensive, especially when there are a large number of tests to be performed. For instance, in DE analysis one often needs to test for thousands of genes. In order to reduce the computational load, we employ an early stopping rule in our experiments.[13] Specifically, after the b-th bootstrap iteration, we calculate $\hat{p}_b^* = \sum_{l=1}^{b} 1_{T^{(l)} \geq T}/b$. If $\hat{p}_b^* > (a/b + c)/(1 + c)$ where a and c are some constants, then we stop the bootstrap procedure and output $\hat{p}^s = \hat{p}_b^*$. Otherwise, the bootstrap procedure will continue until $b = B$ and outputs $\hat{p}^s = \hat{p}_B^*$, where $\hat{p}^s$ is the final p-value estimate for our two-sample test. Following recommendation,[13] we take $c = (1 + \delta) \times p_0/(1 - p_0)$, where p_0 is the p-value cutoff, and a and δ are parameters of choice. In our application, the p-value cutoff is chosen as $p_0 = 0.01$, and we set $a = 4$, $\delta = 0.4$. Thus, $c = 0.0141$.

2.4 | Accelerated bootstrap procedure based on asymptotic distribution of the test statistic

From our experiments, we find that the simple bootstrap procedure described in Section 2.3 provides accurate p-value estimates, but at the price of intensive computation, due to the need to estimate $\hat{G}^{(b)}$ and $\hat{H}^{(b)}$ using g-modeling in each bootstrap iteration. In this section, we propose an accelerated bootstrap procedure based on approximating the null distribution of the test statistic using large sample theory.

Suppose X and Y are sampled from exponential family distributions such as normal, Poisson or binomial distributions. With g-modeling, we assume that G and H as semi-parametric exponential distributions as defined in (2). After obtaining the MLE of $\boldsymbol{\alpha}$ denoted as $\hat{\boldsymbol{\alpha}}$, we denote the estimated PDF and CDF of Θ as $g(\hat{\boldsymbol{\alpha}})$ and $G(\hat{\boldsymbol{\alpha}})$. Since both $G(\hat{\boldsymbol{\alpha}})$ and $g(\hat{\boldsymbol{\alpha}})$ are evaluated on the grid $\tau = (\theta_1, \dots, \theta_m)$, w.l.o.g., assuming it is an equally-spaced grid for simplicity, we have $G(\hat{\boldsymbol{\alpha}}) = \boldsymbol{A}g(\hat{\boldsymbol{\alpha}})$, where

$$\boldsymbol{A} = \begin{bmatrix} a & 0 & \dots & 0 \\ a & a & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \dots & a \end{bmatrix},$$

where a is the grid size.

The asymptotic distribution of $G(\hat{\boldsymbol{\alpha}})$ is[2]:

$$G(\hat{\boldsymbol{\alpha}}) - G(\boldsymbol{\alpha}_0) \doteq N[\text{Bias}\{G(\boldsymbol{\alpha}_0)\}, \text{Cov}\{G(\boldsymbol{\alpha}_0)\}],$$

where $\boldsymbol{\alpha}_0$ is the true value of $\boldsymbol{\alpha}$ and

$$\text{Bias}\{G(\boldsymbol{\alpha}_0)\} = \boldsymbol{A}D(\boldsymbol{\alpha}_0)\boldsymbol{Q}\text{Bias}(\boldsymbol{\alpha}_0), \quad \text{Cov}\{G(\boldsymbol{\alpha}_0)\} = \boldsymbol{A}D(\boldsymbol{\alpha}_0)\boldsymbol{Q}\text{Cov}(\boldsymbol{\alpha}_0)\boldsymbol{Q}^T D(\boldsymbol{\alpha}_0)\boldsymbol{A}^T,$$

where

$$D(\boldsymbol{\alpha}_0) = \text{Diag}\{g(\boldsymbol{\alpha}_0)\} - g(\boldsymbol{\alpha}_0)g(\boldsymbol{\alpha}_0)^T, \quad \text{Bias}(\boldsymbol{\alpha}_0) = -\{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1}\dot{s}(\boldsymbol{\alpha}_0),$$
$$\text{Cov}(\boldsymbol{\alpha}_0) = \{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1}\mathcal{I}(\boldsymbol{\alpha}_0)\{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1},$$

and

$$s(\boldsymbol{\alpha}_0) = c_0||\boldsymbol{\alpha}_0||, \quad \dot{s}(\boldsymbol{\alpha}_0) = c_0\frac{\boldsymbol{\alpha}_0}{||\boldsymbol{\alpha}_0||}, \quad \ddot{s}(\boldsymbol{\alpha}_0) = c_0\frac{c_0}{||\boldsymbol{\alpha}_0||}(\boldsymbol{I}_p - \frac{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T}{||\boldsymbol{\alpha}_0||^2}).$$

Here $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and $\text{Diag}\{g(\boldsymbol{\alpha}_0)\}$ is a diagonal matrix with $g(\boldsymbol{\alpha}_0)$ as the diagonal components. $\mathcal{I}(\boldsymbol{\alpha}_0)$ is the corresponding Fisher information matrix calculated as

$$\mathcal{I}(\boldsymbol{\alpha}) = \boldsymbol{Q}^T[W_i(\boldsymbol{\alpha})W_i(\boldsymbol{\alpha})^T + W_i(\boldsymbol{\alpha})g(\boldsymbol{\alpha})^T + g(\boldsymbol{\alpha})W_i(\boldsymbol{\alpha})^T - \text{Diag}\{W_i(\boldsymbol{\alpha})\}]\boldsymbol{Q},$$

where $W_i(\boldsymbol{\alpha})$ is an m-vector and its j-th element is defined as $w_{ij}(\boldsymbol{\alpha}) = g_j(\boldsymbol{\alpha})\{p_{ij}/f_j(\boldsymbol{\alpha}) - 1\}$, and $\text{Diag}\{W_i(\boldsymbol{\alpha})\}$ is an $m \times m$ diagonal matrix with $W_i(\boldsymbol{\alpha})$ on the diagonal. In practice, the true value $\boldsymbol{\alpha}_0$ is unknown, so we replace $\boldsymbol{\alpha}_0$ with $\hat{\boldsymbol{\alpha}}$ in above formulas to estimate the bias and covariance of G.

Now we move on to consider our two-sample test. As the null hypothesis defined where $G = H$, the true values $\boldsymbol{\alpha}_{X0} = \boldsymbol{\alpha}_{Y0}$. Hence, we can pool the two groups X and Y together for estimating the parameter vector similarly as the simple bootstrap procedure and obtain the estimate $\hat{\boldsymbol{\alpha}}_p$. Then we use $\hat{\boldsymbol{\alpha}}_p$ in the calculation of the biases and the covariance matrices. Since the two groups are independent, the asymptotic null distribution of $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ is

$$G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \doteq N[\text{Bias}\{G(\hat{\boldsymbol{\alpha}}_p)\} - \text{Bias}\{H(\hat{\boldsymbol{\alpha}}_p)\}, \text{Cov}\{G(\hat{\boldsymbol{\alpha}}_p)\} + \text{Cov}\{H(\hat{\boldsymbol{\alpha}}_p)\}]. \tag{4}$$

Our two-sample K-S test statistics T defined in (3) is the maximum absolute difference between $G(\hat{\boldsymbol{\alpha}}_p)$ and $H(\hat{\boldsymbol{\alpha}}_p)$. Although it is difficult to derive the null distribution of T analytically, we can use parametric bootstrap to simulate its null distribution based on (4) and then estimate the p-value of the test. By doing so, since we only need to sample from a multivariate normal distribution and therefore have avoided the estimation of $\hat{G}^{(b)}$ and $\hat{H}^{(b)}$ using g-modeling in each bootstrap iteration as in the simple bootstrap procedure, the computational burden is greatly reduced. See Section S1.1 in the supplementary materials for more details.

The accelerated bootstrap procedure based on the asymptotic null distribution derived in this section is computationally more efficient than the simple bootstrap procedure described in Section 2.3, as we directly obtain bootstrap samples from the multivariate normal distribution. However, it only provides an approximated p-value estimate which requires relatively large sample sizes in both groups for the approximation to be accurate, especially in settings where the convergence of the asymptotic distribution may be slow.

3 | APPLICATIONS

3.1 | Malignant nodes analysis on intestinal surgery

We apply our method on a satellite nodes dataset of intestinal surgery in Gholami and others (2002).[14] There are 844 cancer patients who have removed satellite nodes for later testing. Each patient has a pair of $(n_i, X_i), i = 1, \dots, 844$, where n_i is the number of removed satellites and X_i is the number of malignant satellites found. Binomial distribution is assumed as $X_i \sim Bin(n_i, \Theta_i)$, where Θ_i denotes the probability of any one satellite being malignant for the patient i.

To extend the real dataset into a two-group comparison setting, we simulate data based on these 844 pairs in three cases. We add a binary group indicator C ($C = 0$ or $C = 1$) and randomly split data into two groups with equal size. Now we have

$$\Theta_i | C_i = 0 \sim G, \qquad \Theta_i | C_i = 1 \sim H.$$

G is estimated with all pairs with $C = 0$ and H is estimated with all pairs with $C = 1$.

In the first case – null hypothesis, we directly apply g-modeling to estimate G and H, and then perform our two-sample test with 99 bootstrap iterations. We repeat the random assignment procedure for $1,000$ times and obtain the histogram and Q-Q plot of empirical p-values (see Figure 1a and Figure 1b). Both the histogram and the Q-Q plot shows a uniform distribution the p-values which indicated insignificant difference in the distribution of the malignant probability of satellites between the two randomly assigned groups, which is expected.

In the second case – alternative hypothesis 1, we keep $\hat{G}$ as the final estimated density for the group with $C = 0$, and for the group with $C = 1$, we sample $\hat{\Theta}_i$ from $\hat{H}$ and implement the transformation $\Theta_i^* = (1 - w)\hat{\Theta}_i + wI(\hat{\Theta}_i > 0)$ where $0 < w < 1$ is a tuning parameter. Θ_i^* is more distinct from $\hat{\Theta}_i$ with larger w. Here Θ_i^* is always greater than Θ_i.

In the third case – alternative hypothesis 2, we also maintain $\hat{G}$ as the final estimated density for the group with $C = 0$, and for the group with $C = 1$, we sample $\hat{\Theta}_i$ from $\hat{H}$ and implement the transformation $\Theta_i^* = (1 - w)\hat{\Theta}_i + w\{0.5I(0 < \hat{\Theta}_i < 1) + I(\hat{\Theta}_i = 1)\}$ where $0 < w < 1$ is a tuning parameter. Θ_i^* is more distinct from $\hat{\Theta}_i$ with larger w. Θ_i^* can be different from Θ_i in two directions – either larger than or smaller than Θ_i.

In both alternative hypotheses, X_i^* is generated from $Bin(n_i, \Theta_i^*)$. Then we obtain the final estimated density for the group $C = 1$ from the generated pairs (n_i, X_i^*). Through 99 bootstrap iterations, we compute the empirical p-values. The data generation procedure is repeated 100 times with the same w, which ranges from 0 to 0.05 for the power plot in the alternative hypothesis 1, while $w \in (0, 0.1)$ for the power plot in the alternative hypothesis 2. From Figure 2a and Figure 2b, we can see that the power increases as w increases in both cases. However, the alternative case 1 has higher power than the alternative case 2 with the same w. In the alternative case 1, the change can only occur in one direction. However the transformation in the alternative case 2 involves two directions of change which reduces the final combined transformation. Therefore, with the same w, the transformation in the alternative case 1 is more dramatic than that in the alternative case 2.

3.2 | Differential expression analysis of single-cell RNA-seq data

Another application of our proposed two-sample test is the differential expression (DE) analysis of RNA sequencing (RNA-Seq) or the more recent single-cell RNA sequencing (scRNA-seq) data. Such data is often sparse, i.e. have lots of zero counts,[15–17] and the gene distribution of scRNA-seq data is complex since there is substantial heterogeneity among different cell samples.[18,19] Most importantly, since each measurement (read count for a given gene) from a sample depends on some sample-specific factors (e.g. library size) of that sample, the measurements across samples are not identically distributed.

Since the true expression level of a gene can vary across samples (e.g., cells in scRNA-seq), we model it as the underlying unknown parameters of interest (i.e., Θ_i and Λ_k) with g-modeling. The observed read counts (i.e., X_i and Y_k) are generated from these underlying parameters via a parametric distribution (e.g., zero-inflated Poisson) with the library size as a known covariate. Many existing DE methods assume that the gene expression level follows some parametric distribution (e.g., normal or gamma distributions) and are interested in detecting its locational changes between conditions. Our approach, however, use the semi-parametric g-modeling framework and can detect distributional changes between conditions.

In the literature, a number of methods have been introduced to detect DE genes from scRNA-seq data. Model-based analysis of single-cell transcriptomics (MAST) and single-cell differential expression (SCDE) fit two-stage models to handle inflated zero counts.[20,21] Nonparametric methods such as SigEMD, EMDomics and D3E address the multimodality issue in scRNA-seq data[22–24]. However, as pointed out by Jaakkola and others (2016) and Wang and others(2019),[25,26] none of these methods is able to handle inflated zero counts and multimodality issues simultaneously. We compare our method on a scRNA-seq dataset with seven other existing methods designed for differential expression analysis: (i) Model-based analysis of single-cell transcriptomics (MAST) models scRNA-seq data with a mixture of two components. One component describes the unobserved or dropout

measurements, and the other component explains the observed gene expression in cells.[20] A two-part generalized linear model is used to fit the data. MAST only considers the differences in location between the two groups. (ii) Single-cell differential expression (SCDE) is a three-step Bayesian approach, including data filtering, finding an error model and differential expression test.[21] (iii) Single-cell Differential Distributions (scDD) is a mixture modelling method based on a Bayesian framework to detect genes with expression changes in distributions between conditions.[27] MAST, SCDE and scDD are all designed specifically for scRNA-seq data. Four other methods designed for differential expression analysis in microarray or bulk RNA-seq data are also included in our comparison. (iv) Differential expression analysis for sequence count data (DESeq) is based on a negative binomial model with mean and variance linked through local regression.[28] (v) Linear models for microarray and RNA-Seq data (Limma-Voom) fit a gene-wise linear model on the gene expression values and applied modified t-statistical to test for differential expressed genes.[29] (vi) Reproducibility-optimized test statistic (ROTS) uses a modified t-statistic by maximizing the reproducibility of top-ranked features across group-preserving bootstrap samples.[30] (vii) A negative binomial generalized log-linear model is fitted with likelihood ratio test on read counts by edgeR to detect DE genes.[31]

Accounting for the excess number of zero counts in scRNA-seq data, we modify our test statisitcs and the associated two bootstrap procedures. Specifically, given a gene, we assume that the reads count (X) for that gene from each cell follow a Zero-Inflated Poisson distribution $ZIP(\lambda, \pi)$ where λ is the usual Poisson rate parameter and π is the excessive probability of zero counts with the following probability mass function

$$p(X = k|\pi, \lambda) = \begin{cases} \pi + (1-\pi)e^{\lambda}, & k = 0, \\ (1-\pi)\dfrac{\lambda^k e^{\lambda}}{k!}, & k > 0. \end{cases} \tag{5}$$

However, the Poisson expression rates and are different for different cells. These Poisson expression rates cannot be observed but assumed to follow a certain distribution which can be estimated from the reads counts with g-modeling density estimation. We denote $G(\boldsymbol{\alpha}_X)$ and $H(\boldsymbol{\alpha}_Y)$ as the distributions of Poisson expression rates in two groups. We define π_X and π_Y as the excessive probability of zeros in two groups. Then, the problem of DE analysis is to detect the difference between $G(\boldsymbol{\alpha}_X)$ and $H(\boldsymbol{\alpha}_Y)$ as well as the difference between π_X and π_Y. In such case, since π_X and π_Y also need to be estimated together with $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_Y$, we use a modified K-S test statistic defined as

$$T = \max\{\max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, |\hat{\pi}_X - \hat{\pi}_Y|\}. \tag{6}$$

Note that here we apply the g-modeling method for a purpose different from that in Efron's original paper. Instead of treating the test statistics across all the genes as observations,[2] here we model the read counts from a given gene across all the cells as observations.

For the modified test statistic T defined in (6), a similar simple parametric bootstrap procedure with slight modifications can be used. We first estimate $\hat{\pi}_p$ together with $\hat{\boldsymbol{\alpha}}_p$ by pooling two groups together. Then, in each bootstrap iteration, we simulate data with the zero-inflated Poisson distribution according to (5) and estimate $(\hat{G}^{(b)}, \hat{\pi}_X^{(b)})$ and $(\hat{H}^{(b)}, \hat{\pi}_Y^{(b)})$ using MLE from the simulated data, respectively, and calculate $T^{(b)}$ as $T^{(b)} = \max(\max_j |\hat{G}_j^{(b)} - \hat{H}_j^{(b)}|, |\hat{\pi}_X^{(b)} - \hat{\pi}_Y^{(b)}|)$.

The accelerated bootstrap procedure can also be adjusted accordingly. Specifically, we now have two parts in the modified test statistics, so we define an augmented vector of parameters as $\boldsymbol{\alpha}' = \begin{pmatrix} \pi \\ \boldsymbol{\alpha} \end{pmatrix}$. Given $\Theta_i = \theta_j$ and π, X_i follows a zero-inflated Poisson distribution $ZIP(\theta_j, \pi)$. We can obtain the joint asymptotic distribution of $\hat{\pi}_X - \hat{\pi}_Y$ and $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ in a similar manner as the general case in Section 2.4. Here we also pool the two groups of sample together to estimate the augmented vector of parameters as $\hat{\boldsymbol{\alpha}}'_p = \begin{pmatrix} \hat{\pi}_p \\ \hat{\boldsymbol{\alpha}}_p \end{pmatrix}$ for the asymptotic null distribution. Defining two transformation functions $C(\hat{\boldsymbol{\alpha}}') = \begin{bmatrix} \hat{\pi} \\ G(\hat{\boldsymbol{\alpha}}) \end{bmatrix}$ and $K(\boldsymbol{\alpha}') = \begin{bmatrix} \pi \\ g(\boldsymbol{\alpha}) \end{bmatrix}$, we can then obtain the joint asymptotic multivariate normal distribution of $\hat{\pi}_X - \hat{\pi}_Y$ and $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ as

$$\begin{bmatrix} \hat{\pi}_X - \hat{\pi}_Y \\ G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \end{bmatrix} \doteq N[\text{Bias}\{C(\hat{\boldsymbol{\alpha}}'_p)\} - \text{Bias}\{C(\hat{\boldsymbol{\alpha}}'_p)\}, \text{Cov}\{C(\hat{\boldsymbol{\alpha}}'_p)\} + \text{Cov}\{C(\hat{\boldsymbol{\alpha}}'_p)\}],$$

where

$$\text{Bias}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Bias}\{K(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\dot{\boldsymbol{K}}^T\text{Bias}(\hat{\boldsymbol{\alpha}}'),$$
$$\text{Cov}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Cov}\{K(\hat{\boldsymbol{\alpha}}')\}\boldsymbol{B}^T = \boldsymbol{B}\dot{\boldsymbol{K}}^T\text{Cov}(\hat{\boldsymbol{\alpha}}')\dot{\boldsymbol{K}}\boldsymbol{B}^T,$$

and

$$\dot{\boldsymbol{K}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \dot{g} & \\ 0 & & & \end{bmatrix}, \quad \dot{g} = \boldsymbol{Q}^T D, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \boldsymbol{A} & \\ 0 & & & \end{bmatrix}.$$

Here $\dot{\boldsymbol{K}}$ is the matrix of the first derivatives of K with respect to $\boldsymbol{\alpha}'$ when $\boldsymbol{\alpha}'$ takes the value $\hat{\boldsymbol{\alpha}}'_p$. $\dot{g}$ is the matrix of the first derivatives of g with respect to $\boldsymbol{\alpha}$ when we use $\hat{\boldsymbol{\alpha}}$ to approximate $\boldsymbol{\alpha}$. The bias and covariance of $\hat{\boldsymbol{\alpha}}'$ can be derived similarly as in the simple case. See Section S1.2 in the supplementary materials for more details.

To combine the above asymptotic null distribution with bootstrap for p-value estimation, we can generate bootstrap samples for the difference in π and the difference in cumulative density function jointly from using the above asymptotic multivariate normal distribution.

In practical settings such as the zero-inflated Poisson case, the data is skewed with an excess number of zero counts, and outliers are more likely to occur. We find that the asymptotic null distribution-based accelerated bootstrap procedure may fail occasionally when there are extreme outliers and consequently the estimated covariance matrix becomes singular. We use numerical remedies to stabilize the covariance matrix (e.g., adding a constant diagonal matrix to it) to handle such issues.

3.2.1 | Simulations with the zero-inflated Poisson model

We first evaluate our proposed approach with simulated data from the ZIP model. In particular, we simulate two samples as follows

$$\begin{gathered} X_i \sim ZIP(d_{X_i}\Theta_i, \pi_X), \quad Y_k \sim ZIP(d_{Y_k}\Lambda_k, \pi_Y), \quad i,k = 1, \dots N, \\ d_{X_i} \sim Unif(0.5, 1), \quad d_{Y_k} \sim Unif(0.5, 1), \quad \pi_X = \pi_Y = 0.5, \end{gathered} \tag{7}$$

where d_{X_i} and d_{Y_k} are known constants modeling the sequencing depths (or library sizes), Θ_i and Λ_k are unknown parameters modeling gene expression levels, and π_X and π_Y are unknown parameters modeling the excessive probabilities of zeros. X_i's and Y_k's are the simulated reads counts for a gene from two groups of cells.

Under the null hypothesis, we simulate both Θ_i and Λ_k from chi-square distribution with ten degrees of freedom, that is, $\Theta_i \sim \chi^2_{10}$ and $\Lambda_k \sim \chi^2_{10}$. Under the alternative hypothesis, we simulate Θ_i and Λ_j from two chi-square distributions with different degrees of freedom, respectively, $\Theta_i \sim \chi^2_{10}$ and $\Lambda_k \sim \chi^2_{10+\Delta}$, where we vary the effect size Δ from 1 to 5. We use the modified K-S statistic (6) to capture the difference between the distributions of Θ_i and Λ_k as well as the difference between π_X and π_Y.

We run 1000 simulations and use $B = 99$ bootstrap samples for each simulation to save computational cost. Figures 3a and 3b show the histogram and Q-Q plot of the p-values estimated from data simulated under the null hypothesis, which is uniformly distributed on $(0, 1)$ as expected. Figure 3c shows the estimated distributions of Θ and Λ against the true distribution χ^2_{10}. We see that there are some small but noticeable biases due to the semi-parametric g-modeling and penalized MLE at the tail and peak points, which is expected.[2] The estimated π_X and π_Y also seem reasonable as their histograms are centered around the true value 0.5 as Figures 3d and 3e show.

Figure 4 shows the statistical power of the simple and accelerated bootstrap procedures with varying differences in degrees of freedom between two chi-square distributions ($\Delta = 0$ simulated under the null hypothesis, and $\Delta = 1, \dots, 5$ simulated under the alternative hypothesis), and with varying sample sizes ($N = 100, 300, 500, 2000$) in each group. We can see that the power of both bootstrap procedures increases with larger differences between the two distributions. When $\Delta = 0$, both procedures have controlled type-I error rate at the predefined significance level $\alpha = 0.05$. As Δ increases, the accelerated bootstrap procedure has substantially lower power than the simple bootstrap procedure when the sample size is 100, but the power is similar for the two methods when the sample size is 300 or larger.

Additional simulations based on normal, binomial and Poisson distributions are provided in Section S2 in the supplementary materials. Furthermore, to compare the speed of the two bootstrap approaches, Figure 5 and Table S1 in the supplementary materials show the computational time of the two bootstrap procedures with varying sample sizes for normal-based model. We can see that the time increases linearly with sample size for both procedures. However, the accelerated bootstrap procedure is about 30 to 50 times faster than the simple bootstrap procedure. Hence, the accelerated bootstrap procedure is more suitable for datasets with more observations (e.g., $N > 300$ in each group) for being faster and reasonably powerful, while the simple bootstrap procedure is more suitable for datasets with smaller sample sizes for being more powerful and reasonably fast.

3.2.2 | Real data analysis

We run our proposed two-sample test on a real scRNA-seq dataset on human embryonic cells in early development.[32] We compare 81 cells in embryonic day 3 (E3) to 190 cells embryonic day 4 (E4), similarly as performed in.[33] There are dramatic changes between these two days, so this subset is suitable for running differential expression analysis. We select 2000 genes with the highest mean read counts across all cell lines and pick the genes with higher standard deviation when there are ties. We use the total read counts for each cell line as the constants d_{X_i} and d_{Y_k} adjusted in the zero-inflated Poisson model (7). Due to the relatively small sample size of this dataset, i.e., less than 200 cells in each group, the accelerated bootstrap procedure based on asymptotic distributions (noted as ASY) will have lower power in detecting DE genes. Therefore, we focus on the simple bootstrap procedure with our two-sample test based on the modified K-S statistic (6) (noted as KS) in this experiment. We also apply MAST, ROTS, DESeq, Limma-Voom, SCDE, scDD and edgeR for comparison, based on the code in[25]. Since we test on these 2000 genes simultaneously, we use Benjamini-Hochberg(BH) method to adjust the p-values and control for multiple testing by calculating the false discovery rate (FDR) for each gene.[34] Since most expressed genes have relatively small FDR, we use 999 bootstrap samples for our test with early stopping rule and consider genes with FDR < 0.01 as differentially expressed (DE). From Table 1, out test with the simple bootstrap procedure (KS) detect the largest number of DE genes among the 2000 selected genes across all methods. An example where a DE gene can only detected by our KS method is provided in Section S3 in the supplementary materials. The average expression of this gene is similar in the two groups but the shape of gene distribution differs. The Venn diagrams in Figure 6 show that our KS method has significantly overlapping DE genes with MAST, DESeq, SCDE, scDD and edgeR, while the overlapping DE gene list between our test and Limma-Voom or ROTS is not significant as both overlapping p-values are greater than 0.05. Limma-Voom and ROTS were not specifically designed for scRNA-seq data which might the reason for the poor agreement between our method and these two methods. ASY method also has similar agreement with other existing methods (See Figure S2 in the supplementary materials for details). Moreover, the two p-value estimation procedures for our test are highly consistently in the detection of DE genes.

3.2.3 | Validation with real data-based simulation

Since we do not know the ground truth in the real dataset, to further assess the comparison between our method and the other methods, we simulate data based on the real dataset. First, we restrict the data to a subset of 100 genes selected with the highest mean and standard deviation of read counts across all cell lines. Second, based on the results from our KS method, we select those genes with FDR < 0.01 as true DE genes and the remaining ones as true null genes. For the true null genes, we pool the two groups together to estimate the distribution of gene expression and the probability of excessive zero counts using g-modelling. For the true DE genes, we estimate the distribution of the gene expression and the probability of excessive zero counts separately for the two groups using g-modelling. Third, based on the estimated distributions of gene expression and the estimated probability of excessive zero counts, we simulate the raw reads counts from our zero-inflated Poisson model. Finally, we apply our KS method, ASY method and the other methods to our simulated data to identify DE genes. To account for the randomness of data generation, we repeat the above simulation five times.

With our KS method, 86 DE genes and 14 null genes are detected from the real data, and therefore they are used as true DE genes and true null genes in our simulations. On average, our KS method detects 84 true DE genes and mis-identifies 3 genes. However our ASY method shows poorer performance with 70 true DE genes found and mis-identifies 16 genes. However, there are no null genes mis-identified as DE genes by ASY method. Due to the small sample size of this experiment, the ASY method is more conservative than the KS method, and therefore the ASY method detected fewer DE genes. Among other methods, scDD detects 81 true DE genes as the most but mis-identifies 5 genes. Limma-Voom detects 76 DE genes among which 6 are misidentified. Similarly as Limma-Voom, there is 79 DE genes found by edgeR with 7 of them mis-identified. MAST does not detect any true null genes as DE genes, but it only detects 70 DE genes. ROTS has the poorest performance with both low power and high FDR. SCDE has similar performance as DESeq, where their power is similar to MAST but FDR is higher. Table 2 shows the observed FDR and AUC of all the methods in all five simulations, with the average ROC curves shown in Figure 7. We can see that our KS method, ASY method and MAST have controlled FDRs at level 0.01, while ROTS, DESeq, Limma-Voom, SCDE, scDD and edgeR have inflated FDRs. In terms of AUC, our KS method and ASY methods rank the top two, followed by MAST and scDD, and ROTS ranks the last. When ignoring the choice of p-value cutoff, the rankings of the p-values from our KS and ASY methods were very similar and therefore the AUC of the ASY method is similar with the KS method. Although scDD has high power in detecting DE genes, its relatively high FDR may compromise its performance of real data application. Overall, both our KS method and ASY method outperform other methods in this comparison.

4 | DISCUSSION

In many statistical problems, we observe two unknown distributions indirectly and aim to investigate the difference between them.[35] The unknown distribution can be estimated through deconvolution, in accordance with existing methods.[12] However, these methods are only designed for one-sample estimation. Thus, we combine the existing g-modelling method with two-sample K-S test statistic for two-sample density comparison. In terms of p-value estimation, we propose two versions of bootstrap procedures to cover the wide range of sample sizes and balance the needs for accuracy and speed. For small sample size, the simple bootstrap procedure has higher power than the accelerated bootstrap procedure for detecting the difference in distributions. For large sample size, the accelerated bootstrap procedure based on the asymptotic null distribution provides similar statistical power while being much more computationally efficient than the simple bootstrap procedure.

Our approach can be applied to a wide range of areas, as our approach is capable of handling various types of data including count or continuous outcomes. In the analysis of surgical nodes data with binomial models, our proposed test has controlled type I error and sufficient power in several cases of differences in distribution. In terms of scRNA-seq data application, the existing parametric methods all assumed that the unknown gene distributions follow a particular parametric family, and only detect changes in one (or a few) parameters (usually just the location parameter) of that family, which is restrictive. On the contrary, our proposed test assume that the underlying parameters (i.e., true expression levels of a gene in different samples) follow unknown distributions, and g-modeling allows us to model these unknown distributions and detect changes in distribution from one condition to the other. Compared with other existing methods for differential expression analysis on scRNA-seq data, our approach can detect more DE genes and has well-controlled false discovery rate.

There are several directions for potential future research. The current choice of grid for discretizing the sample space of θ is subjective, and data adaptive approaches could be considered to obtain more efficient density estimation. The computation efficiency might be further improved by the deriving the null distribution the our test statistics directly to avoid the use of any bootstrap procedure. The accelerated bootstrap procedure might be calibrated to boost its power for data with small sample size. The two-group comparison can be further extended to multi-group comparison with a k-sample test statistics. The adjustment for additional covariates could be included in the model to control for confounding effects and to improve the the statistical power. We can also extend our two-sample test to generalized linear mixed models to account for correlated or clustered observations.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data and computer codes that support the findings in this paper are available at at https://github.com/kkttzjy/gmodeltest. These data were derived from the following resources available in the public domain: https://cran.rstudio.com/web/packages/deconvolveR/index.html and https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3929/.

References

1. Pratt JW, Gibbons JD. *Kolmogorov-Smirnov Two-Sample Tests*: 318-344; New York, NY: Springer New York . 1981
2. Efron B. Empirical Bayes deconvolution estimates. *Biometrika* 2016; 103(1): 1-20. doi: 10.1093/biomet/asv068
3. Efron B, Hastie T. *Computer age statistical inference : algorithms, evidence, and data science*. New York, NY, USA : Cambridge University Press, 2016 Cambridge University Press . 2016.
4. Laird N. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association* 1978; 73(364): 805–811.
5. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* 1983; 78(381): 47-55. doi: 10.1080/01621459.1983.10477920

6. Zhang C. Empirical Bayes and compound estimation of normal means. *Statistica Sinica* 1997; 7(1): 181–193.

7. Jiang W, Zhang CH. General Maximum Likelihood Empirical Bayes Estimation of Normal Means. *The Annals of Statistics* 2009; 37(4): 1647–1684.

8. Robbins H. An Empirical Bayes Approach to Statistics. In: University of California Press; 1956; Berkeley, Calif.: 157–163.

9. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics MonographsCambridge University Press . 2010

10. Brown LD, Greenshtein E, Ritov Y. The Poisson Compound Decision Problem Revisited. *Journal of the American Statistical Association* 2013; 108(502): 741–749.

11. Efron B. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association* 2011; 106(496): 1602-1614. doi: 10.1198/jasa.2011.tm11181

12. Efron B. Two Modeling Strategies for Empirical Bayes Estimation. *Statist. Sci.* 2014; 29(2): 285–301. doi: 10.1214/13-STS455

13. Jiang H, Salzman J. Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika* 2012; 99(4): 973-980. doi: 10.1093/biomet/ass051

14. Gholami S, Janson L, Worhunsky DJ, et al. Number of Lymph Nodes Removed and Survival after Gastric Cancer Resection: An Analysis from the US Gastric Cancer Collaborative. *Journal of the American College of Surgeons* 2015; 221(2): 291—299. doi: 10.1016/j.jamcollsurg.2015.04.024

15. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology* 2016; 17(1): 63. doi: 10.1186/s13059-016-0927-y

16. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* 2015; 58(4): 610-620. doi: 10.1016/j.molcel.2015.04.005

17. Alam M, Al Mahi N, Begum M. Zero-Inflated Models for RNA-Seq Count Data. *Journal of Biomedical Analytics* 2018; 1: 55-70. doi: 10.30577/jba.2018.v1n2.23

18. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 2015; 16: 133. doi: 10.1038/nrg3833

19. Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013; 498: 236. doi: 10.1038/nature12172

20. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* 2015; 16(1): 278. doi: 10.1186/s13059-015-0844-5

21. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nature Methods* 2014; 11: 740. doi: 10.1038/nmeth.2967

22. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods* 2018; 145: 25-32. doi: https://doi.org/10.1016/j.ymeth.2018.04.017

23. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* 2015; 32(4): 533-541. doi: 10.1093/bioinformatics/btv634

24. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 2016; 17(1): 110. doi: 10.1186/s12859-016-0944-6

25. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics* 2016; 18(5): 735-743. doi: 10.1093/bib/bbw057

26. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* 2019; 20(1): 40. doi: 10.1186/s12859-019-2599-6

27. Korthauer KD, Chu LF, Newton MA, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* 2016; 17(1): 222. doi: 10.1186/s13059-016-1077-y

28. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology* 2010; 11(10): R106. doi: 10.1186/gb-2010-11-10-r106

29. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* 2004; 3(1): 3-25. doi: 10.2202/1544-6115.1027

30. Seyednasrollah F, Rantanen K, Jaakkola P, Elo LL. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Research* 2015; 44(1): e1-e1. doi: 10.1093/nar/gkv806

31. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 2012; 40(10): 4288-4297. doi: 10.1093/nar/gks042

32. Petropoulos S, Edsgärd D, Reinius B, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* 2016; 165(4): 1012-1026. doi: 10.1016/j.cell.2016.03.023

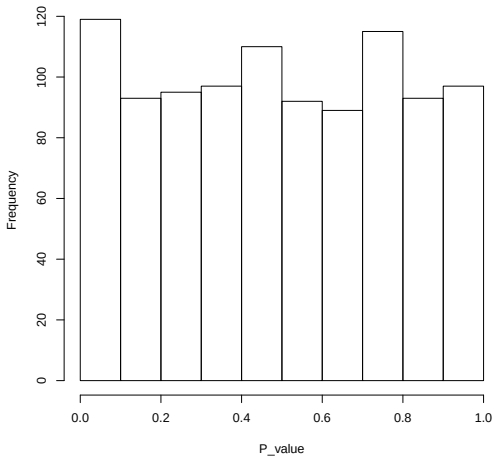
33. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 2018; 34(18): 3223-3224. doi: 10.1093/bioinformatics/bty332

34. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; 57(1): 289–300.

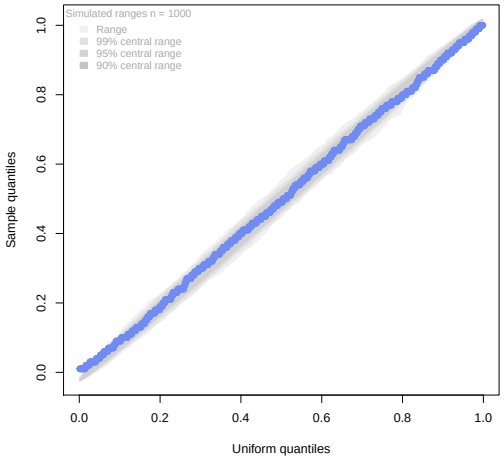
35. Madrid-Padilla OH, Polson NG, Scott J. A deconvolution path for mixtures. *Electron. J. Statist.* 2018; 12(1): 1717-1751. doi: 10.1214/18-EJS1430

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

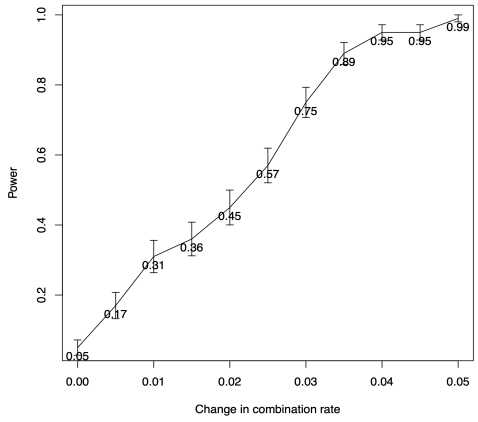


(a) Histogram of p-values

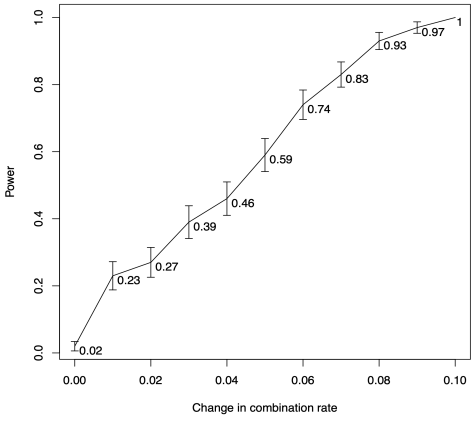


(b) Q-Q plot of p-values

FIGURE 1 (a) Histogram for estimated *p*-values from 1000 simulations under the null hypothesis and (b) Q-Q plot for estimated p-values against $U(0, 1)$.



(a) Alternative case 1

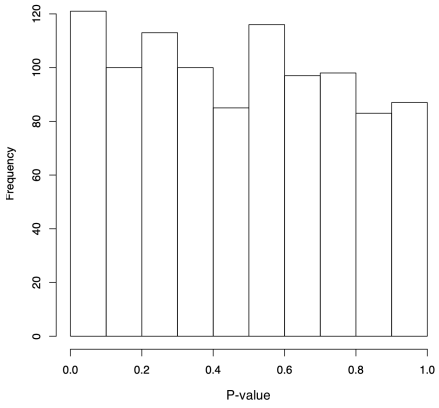


(b) Alternative case 2

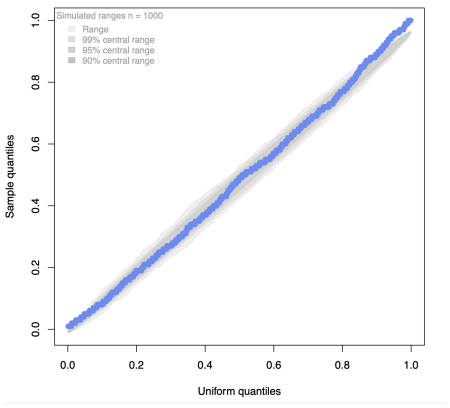
FIGURE 2 Power plots for alternative cases (a) 1 and (b) 2. For each line, the power changes with the difference in π varying from 0.1 to 1. The vertical bar represents the mean $\pm$ sd for each power value.

TABLE 1 Number of DE genes found among 2000 selected genes by proposed methods and other methods.

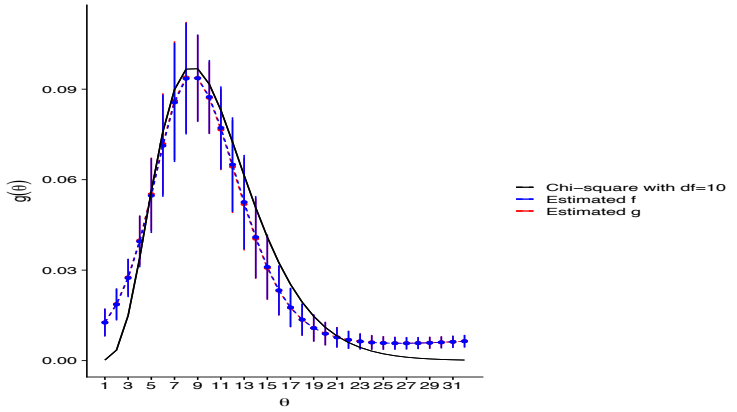
Method	KS	ASY	MAST	ROTS	DESeq	Limma-Voom	SCDE	scDD	edgeR
DE genes	1748	1294	1570	1647	1398	1546	1492	1642	1612



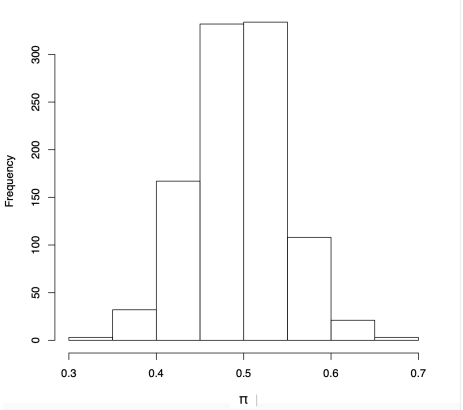
(a) Histogram of p-values



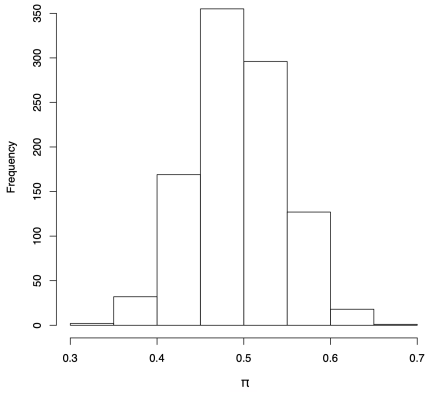
(b) Q-Q plot of p-values



(c) Density comparison

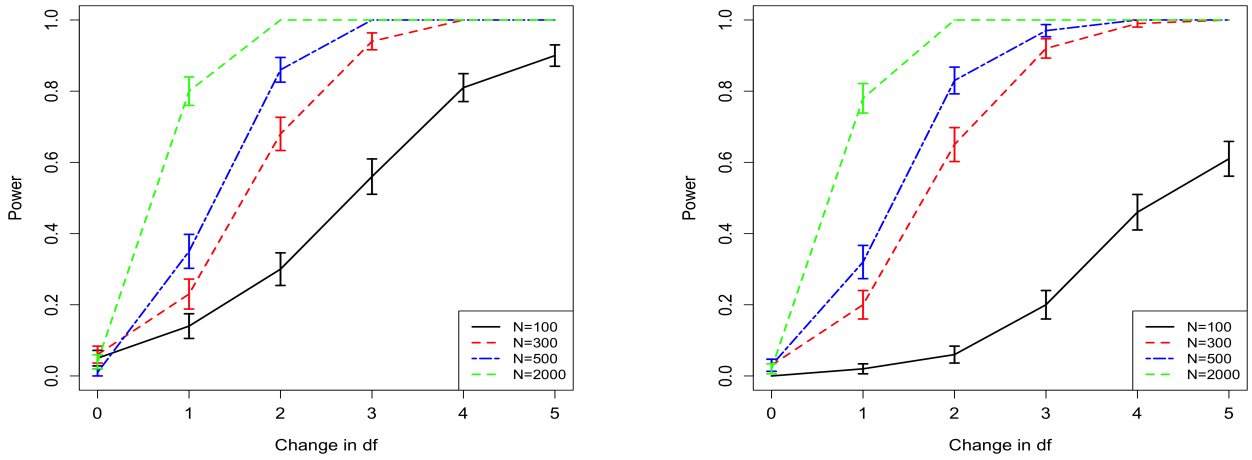


(d) Histogram of π_X



(e) Histogram of π_Y

FIGURE 3 (a) Histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density χ^2_{10}, and histograms for estimated (d) π_X and (e) π_Y.



(a) Simple bootstrap

(b) Accelerated bootstrap

FIGURE 4 Power plots of proposed (a) simple and (b) accelerated bootstrap procedures with varying sample sizes. For each line, the power changes with the difference in degrees of freedom varying from 0 to 5. The vertical bars represent the mean ± sd for each power value.

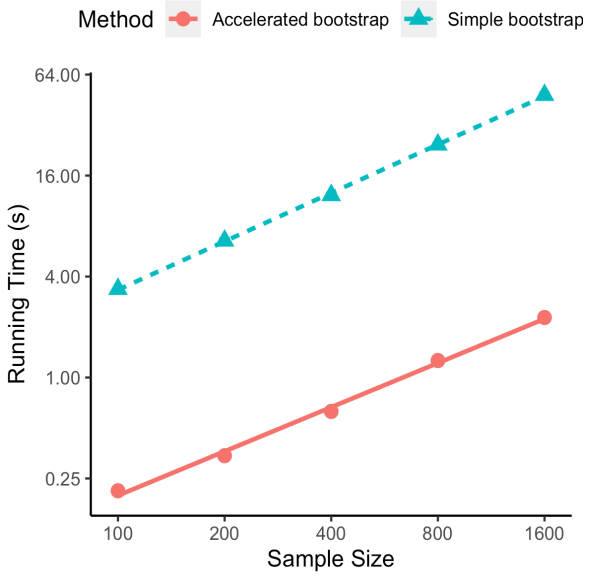


FIGURE 5 Log-log plot of Running time (in seconds) for proposed simple and accelerated bootstrap procedures against sample size in each group under null hypothesis for normal-based model.

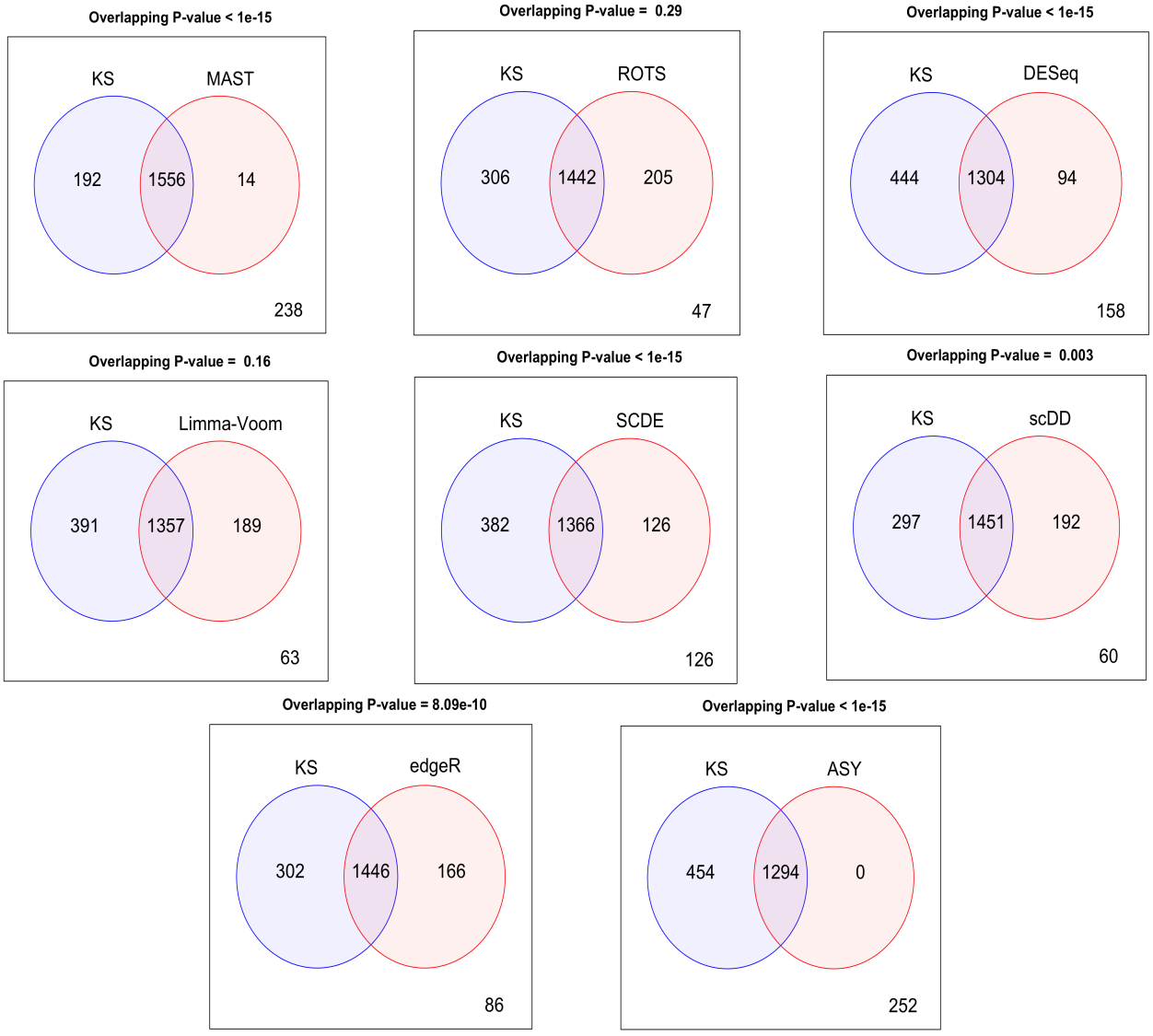


FIGURE 6 Venn diagrams for numbers of DE genes found by proposed method (KS) versus other methods with p-values from hypergeometric tests.

TABLE 2 Observed FDRs and AUCs of proposed methods (KS and ASY) and other methods in five simulations.

Simulation		1	2	3	4	5	Mean (SD)
Observed FDR	KS	0.00	0.07	0.00	0.00	0.00	0.01 (0.03)
	ASY	0.00	0.00	0.00	0.00	0.00	0.00 (0.00)
	MAST	0.00	0.00	0.00	0.00	0.00	0.00 (0.00)
	ROTS	0.07	0.00	0.00	0.64	0.79	0.30 (0.38)
	DESeq	0.07	0.29	0.36	0.07	0.14	0.19 (0.13)
	Limma-Voom	0.29	0.36	0.79	0.36	0.36	0.43 (0.20)
	SCDE	0.00	0.36	0.43	0.21	0.07	0.21 (0.18)
	scDD	0.29	0.14	0.36	0.14	0.29	0.24 (0.10)
	edgeR	0.43	0.57	0.79	0.43	0.50	0.54 (0.15)
AUC	KS	1.00	1.00	1.00	1.00	1.00	1.00 (0.00)
	ASY	0.98	1.00	0.98	0.99	0.99	0.99 (0.01)
	MAST	0.98	0.97	0.98	0.94	0.94	0.96 (0.02)
	ROTS	0.81	0.80	0.71	0.42	0.48	0.64 (0.18)
	DESeq	0.84	0.83	0.80	0.86	0.85	0.84 (0.02)
	Limma-Voom	0.84	0.81	0.74	0.85	0.82	0.81 (0.04)
	SCDE	0.96	0.81	0.77	0.80	0.91	0.85 (0.08)
	scDD	0.95	0.99	0.93	0.97	0.95	0.96 (0.02)
	edgeR	0.85	0.82	0.76	0.82	0.82	0.81 (0.03)

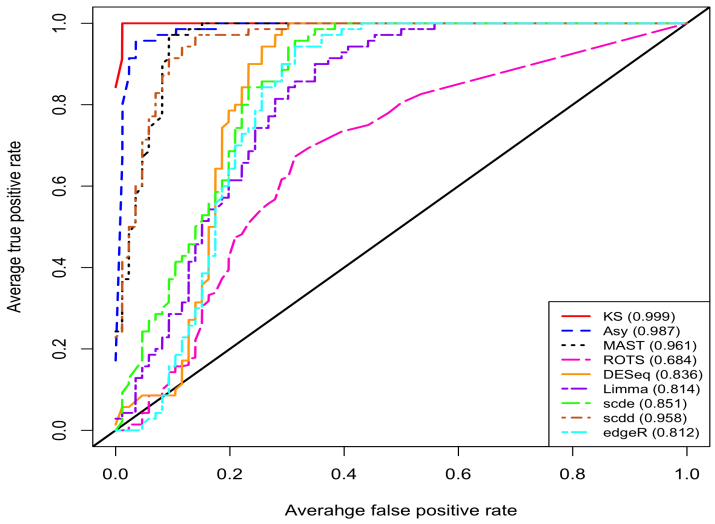
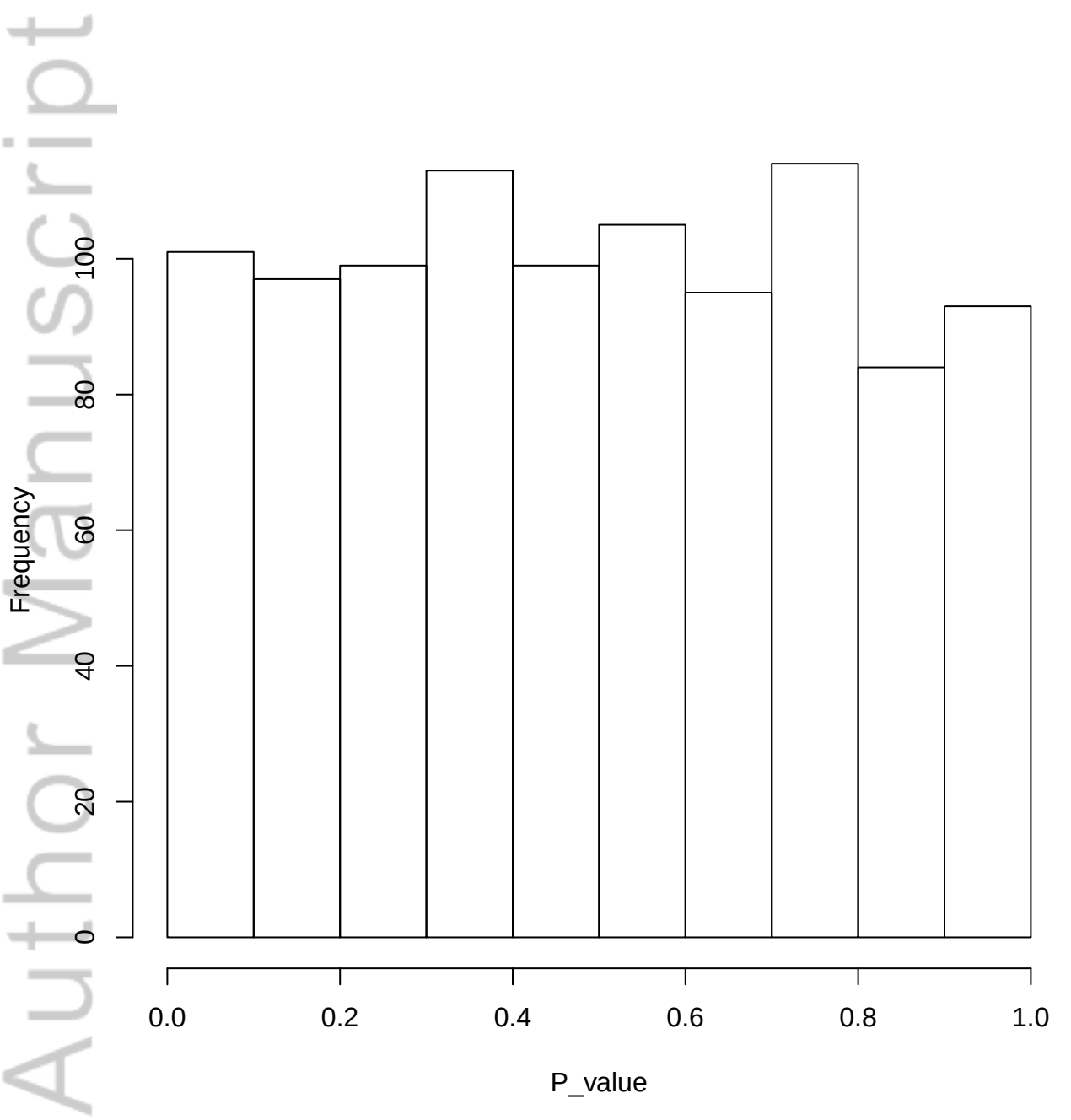
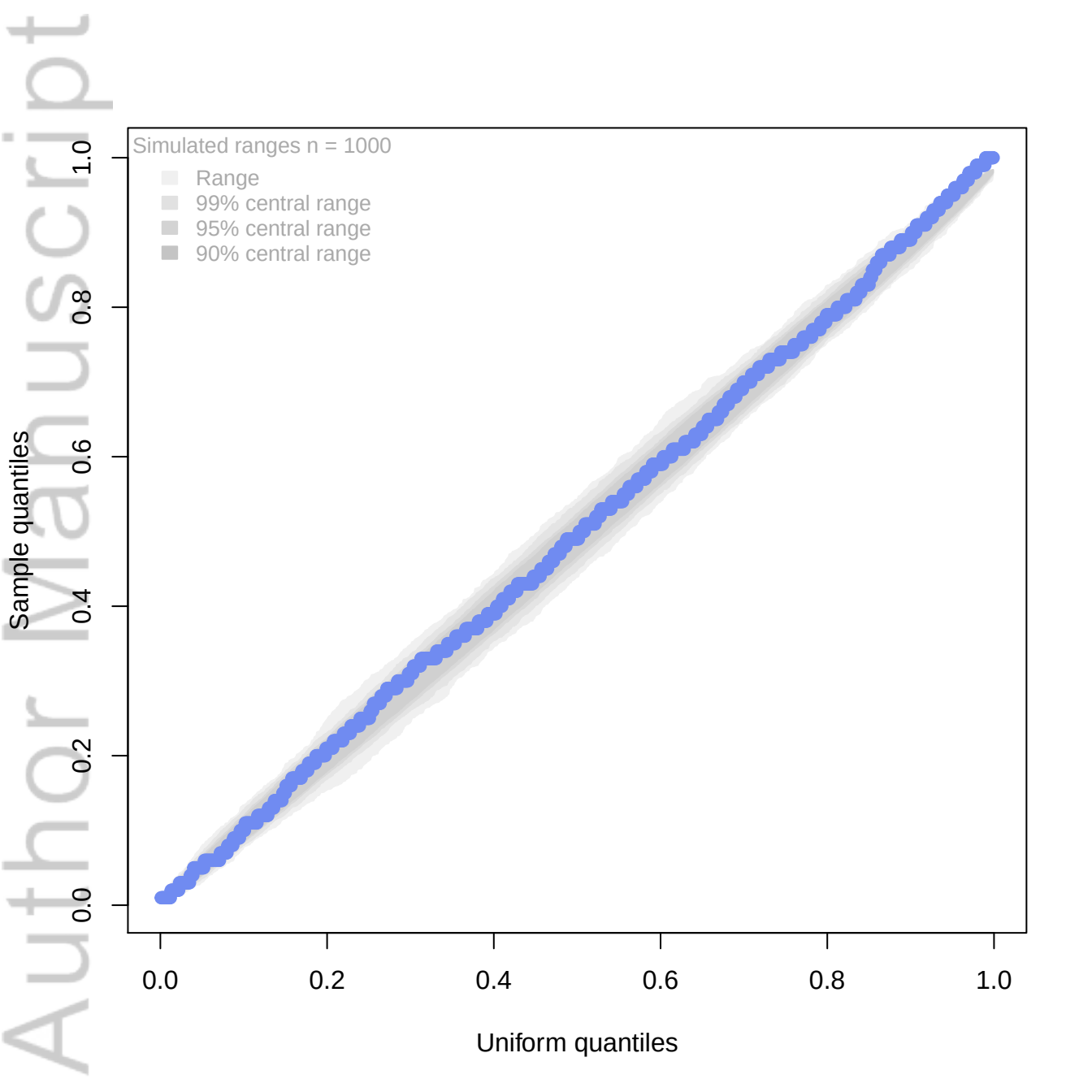
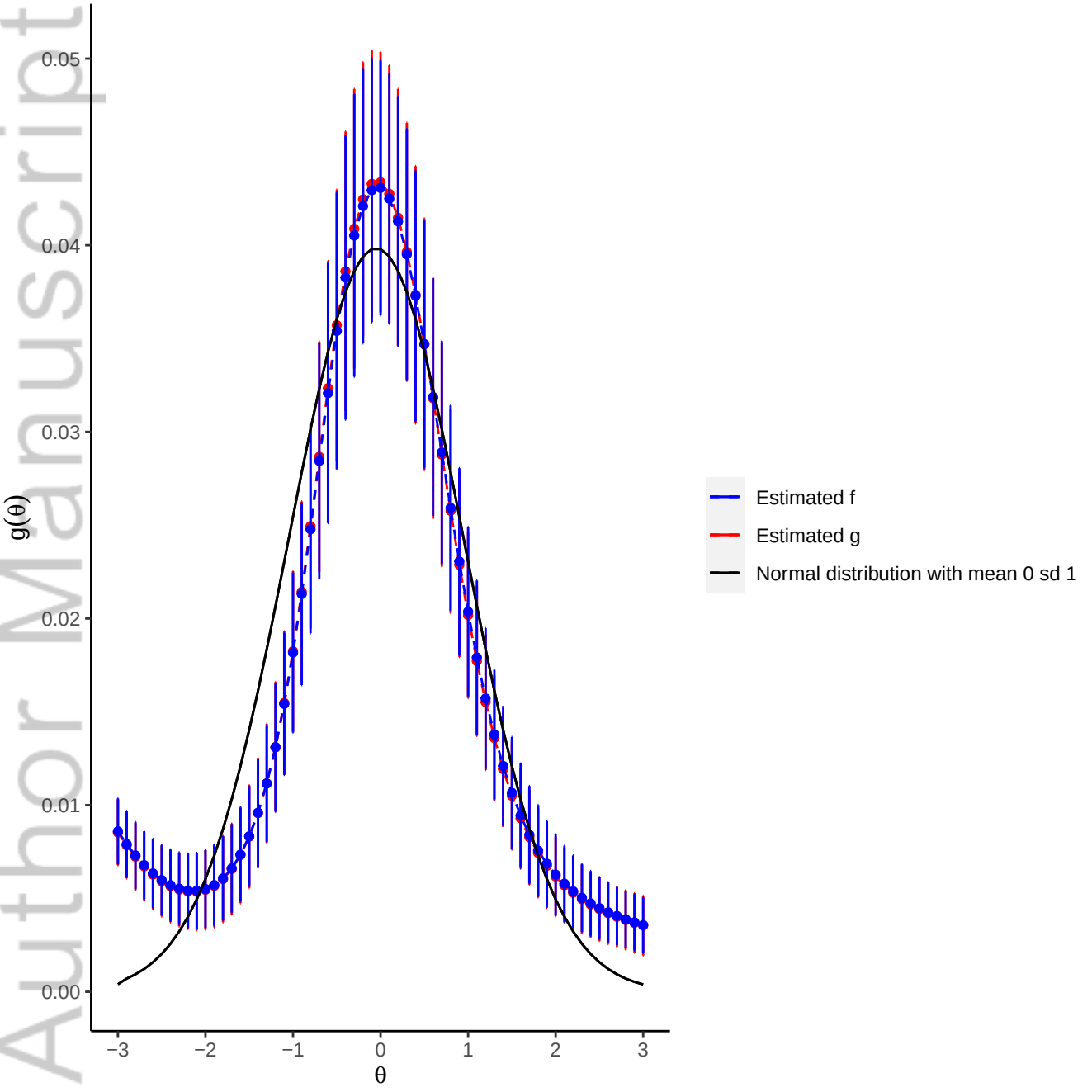


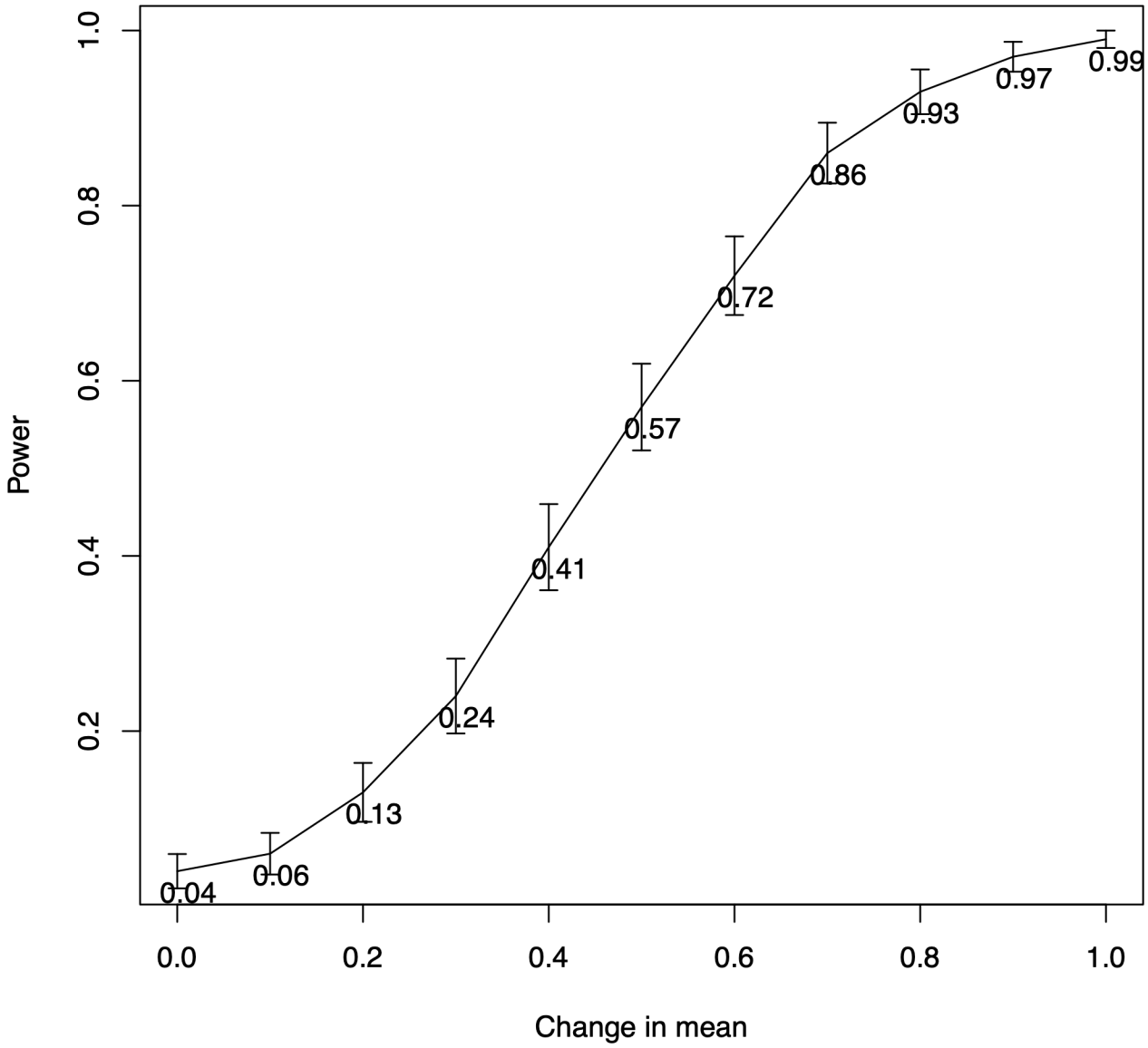
FIGURE 7 Mean ROC curves across five simulations for eight methods.

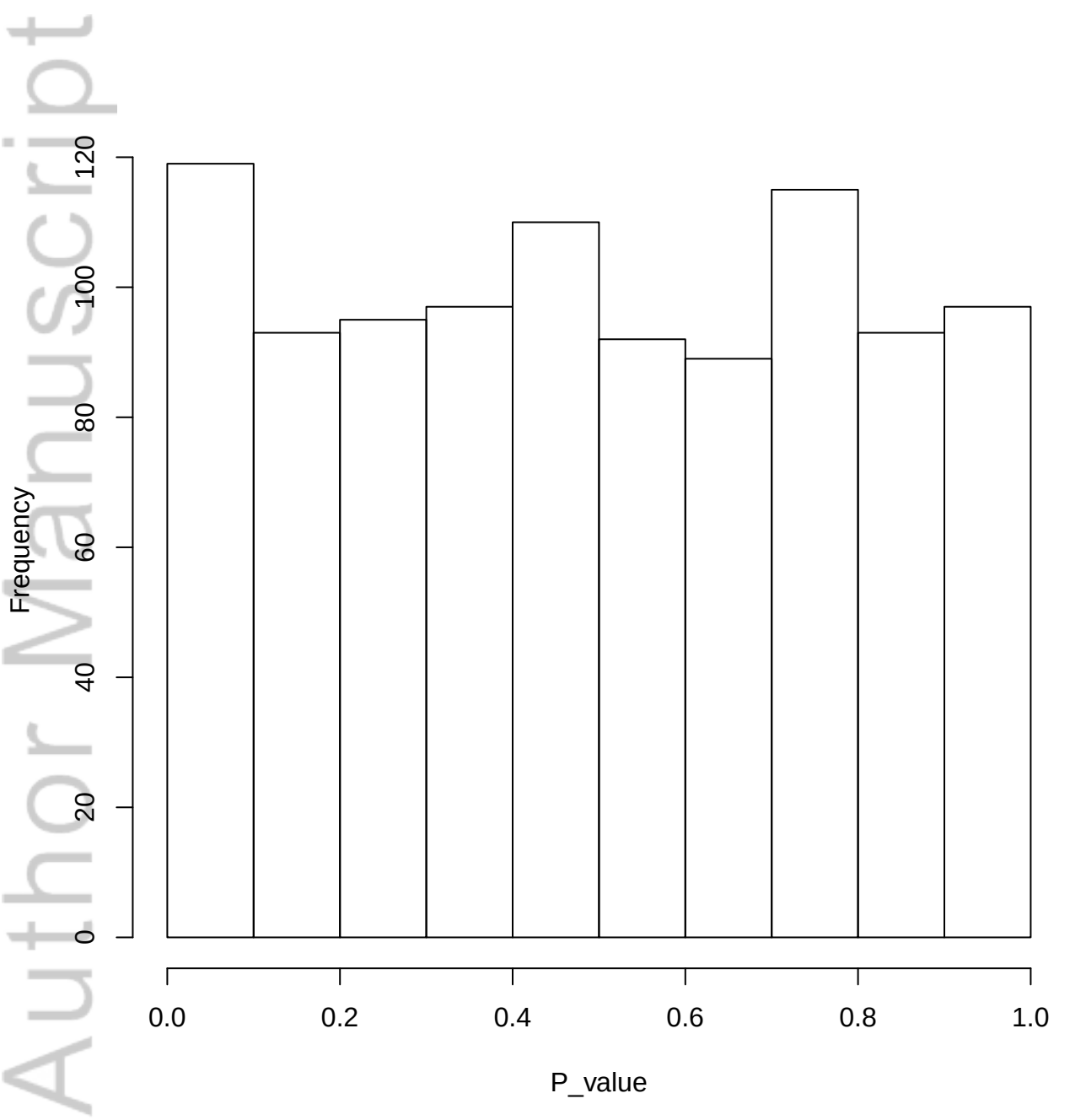
How to cite this article:

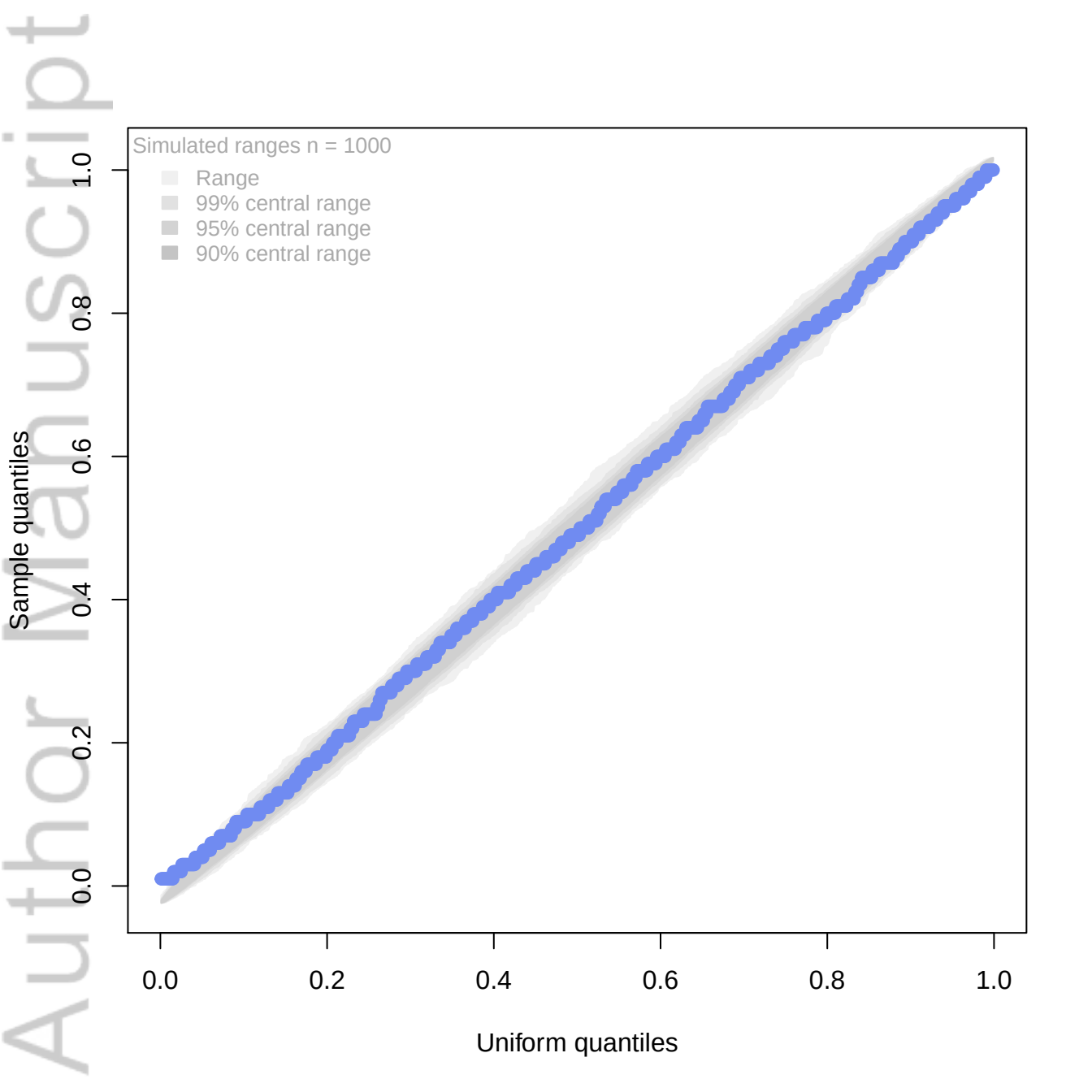


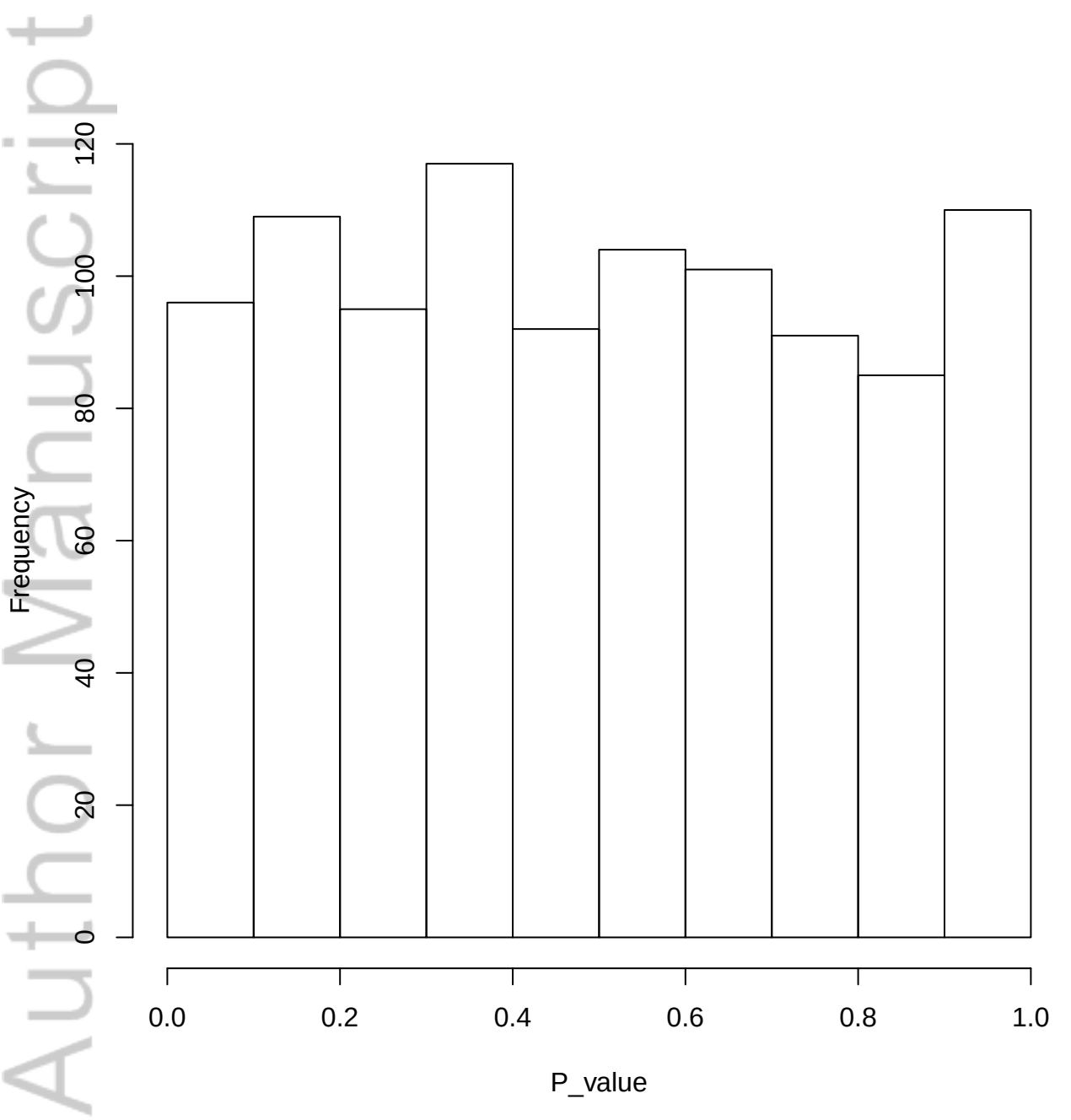


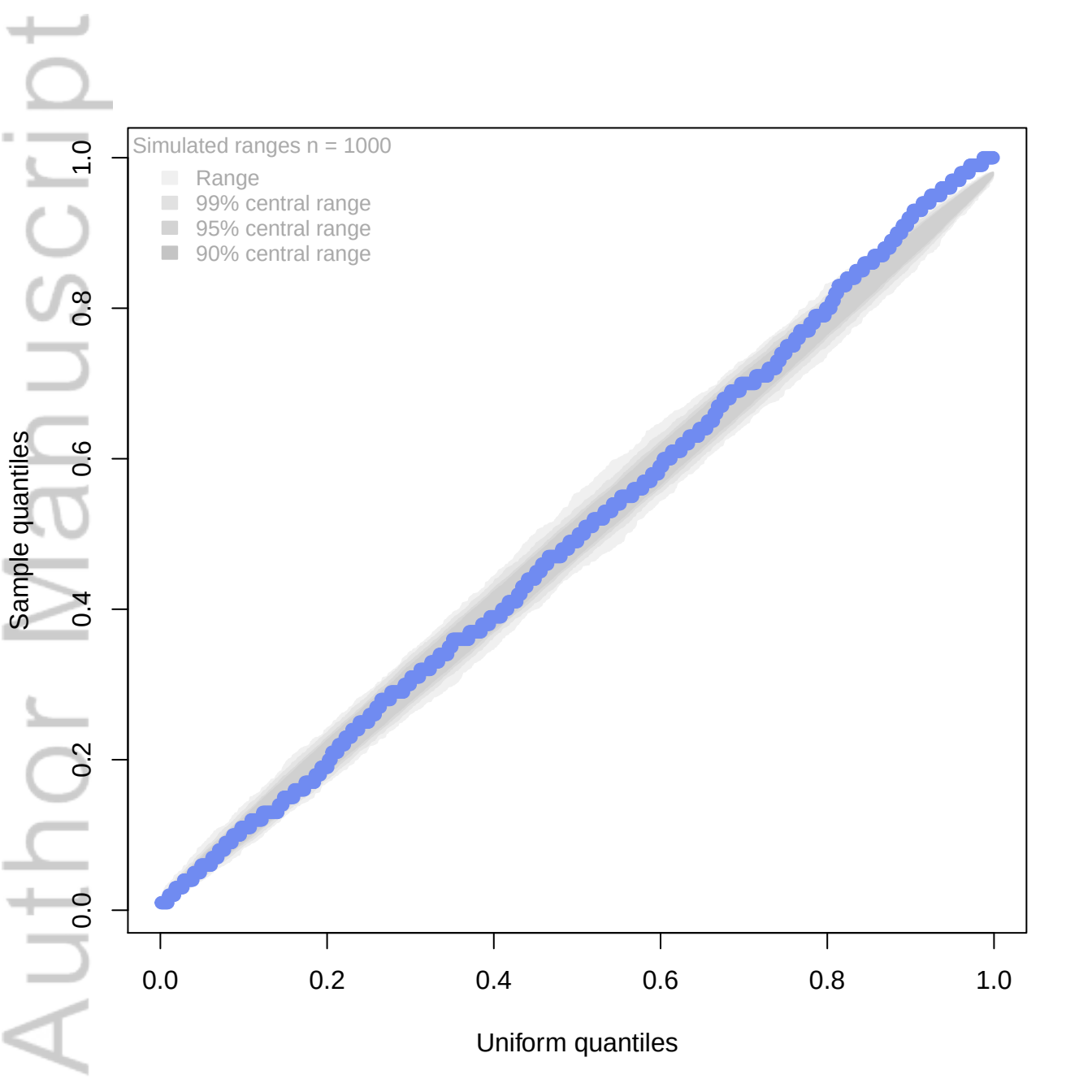


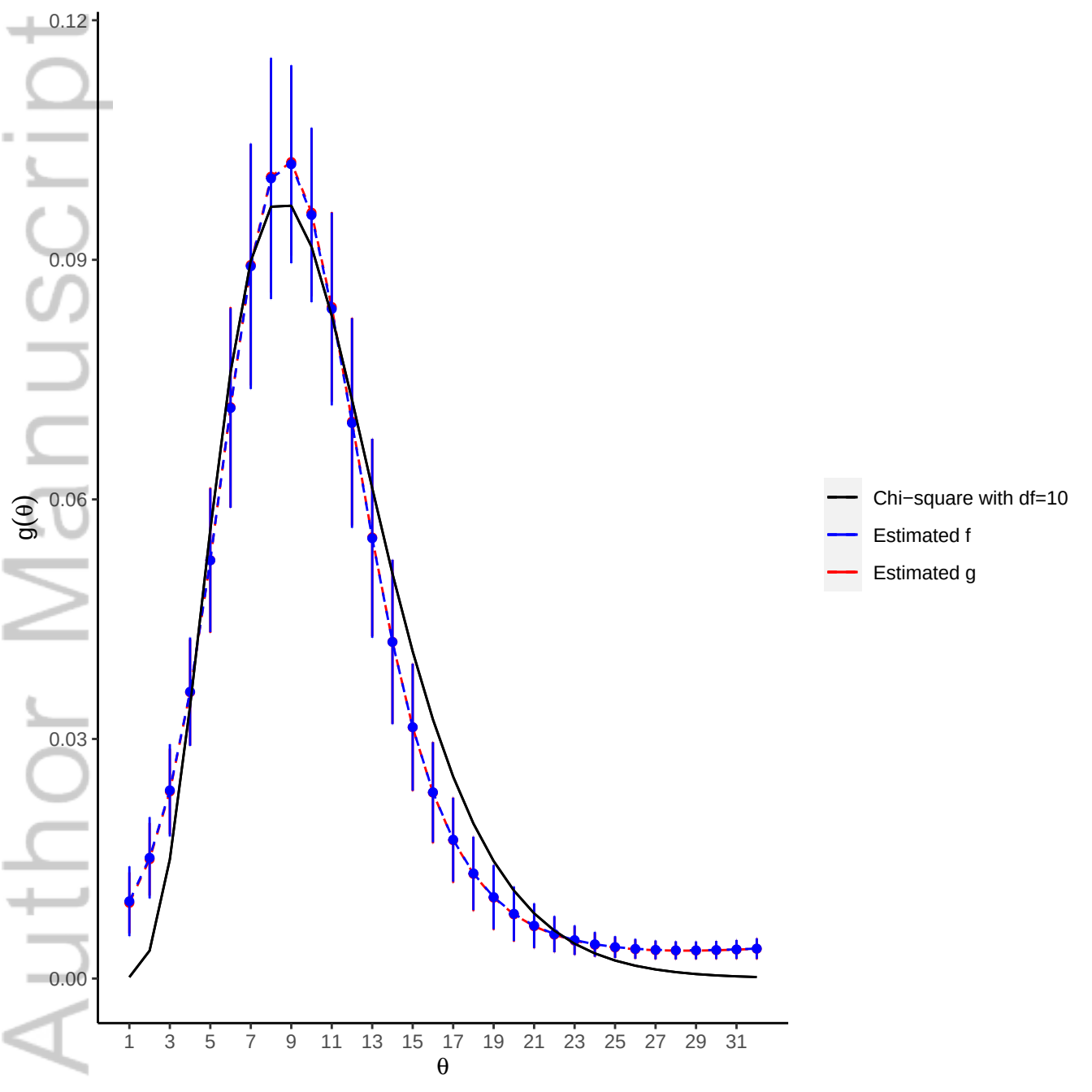


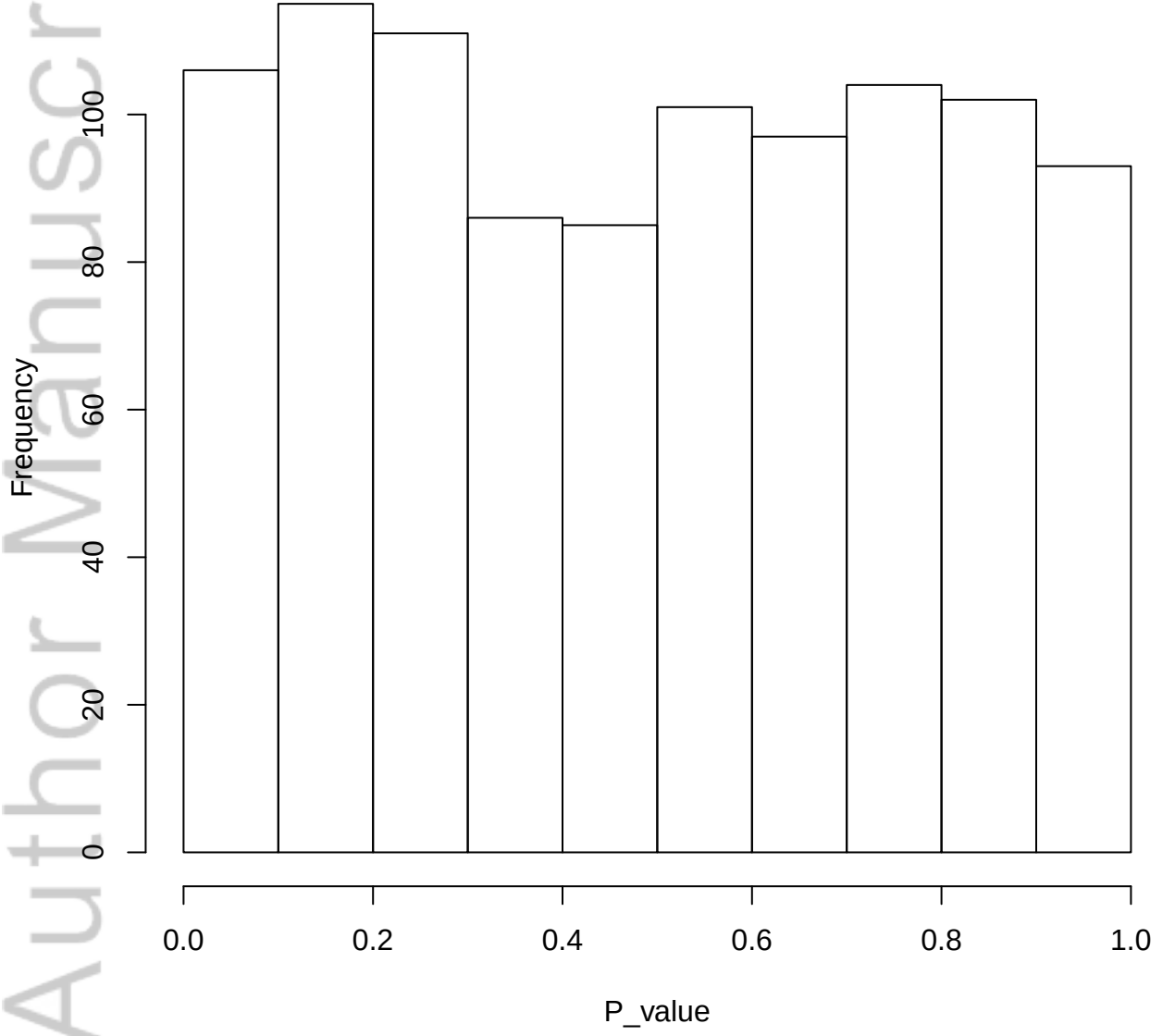


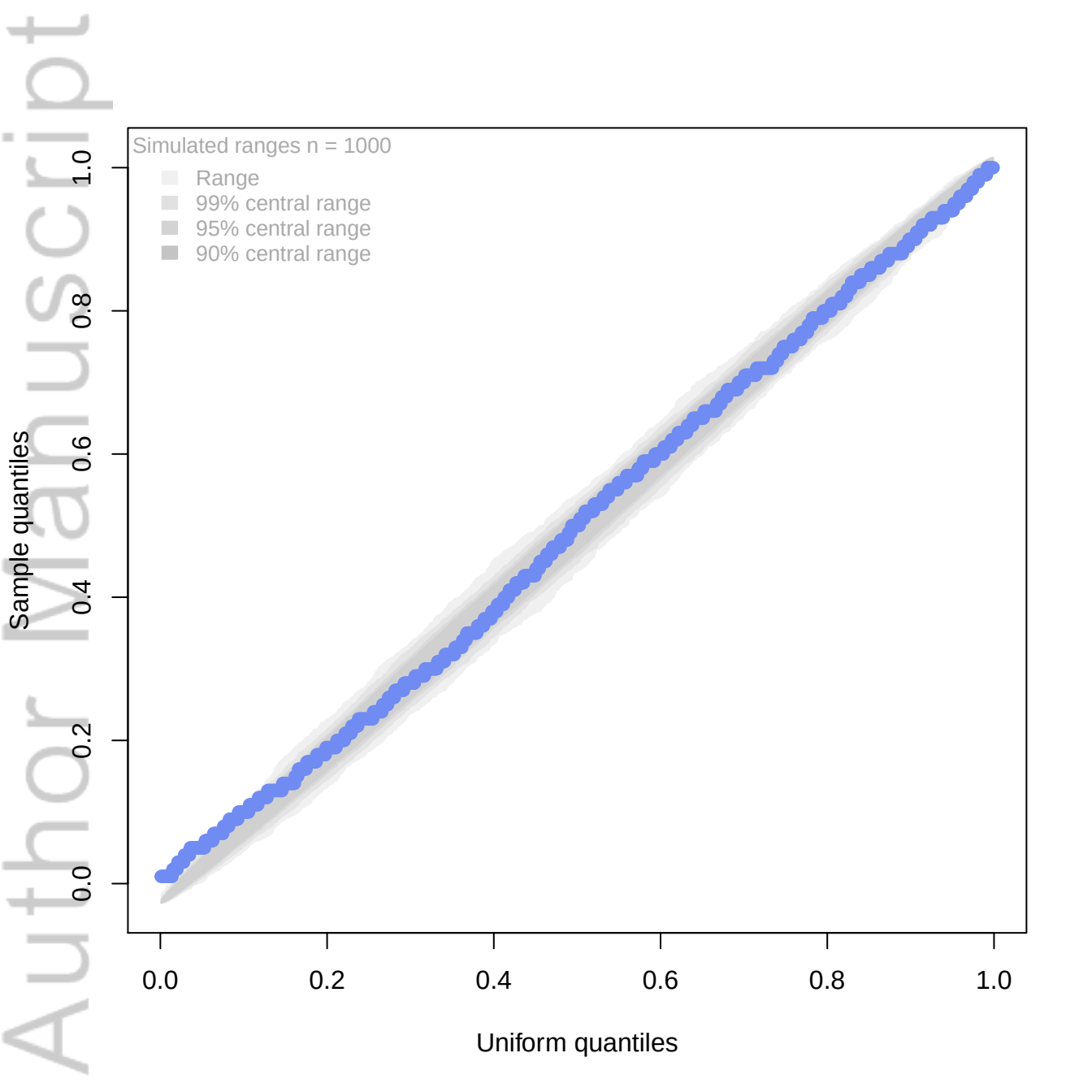


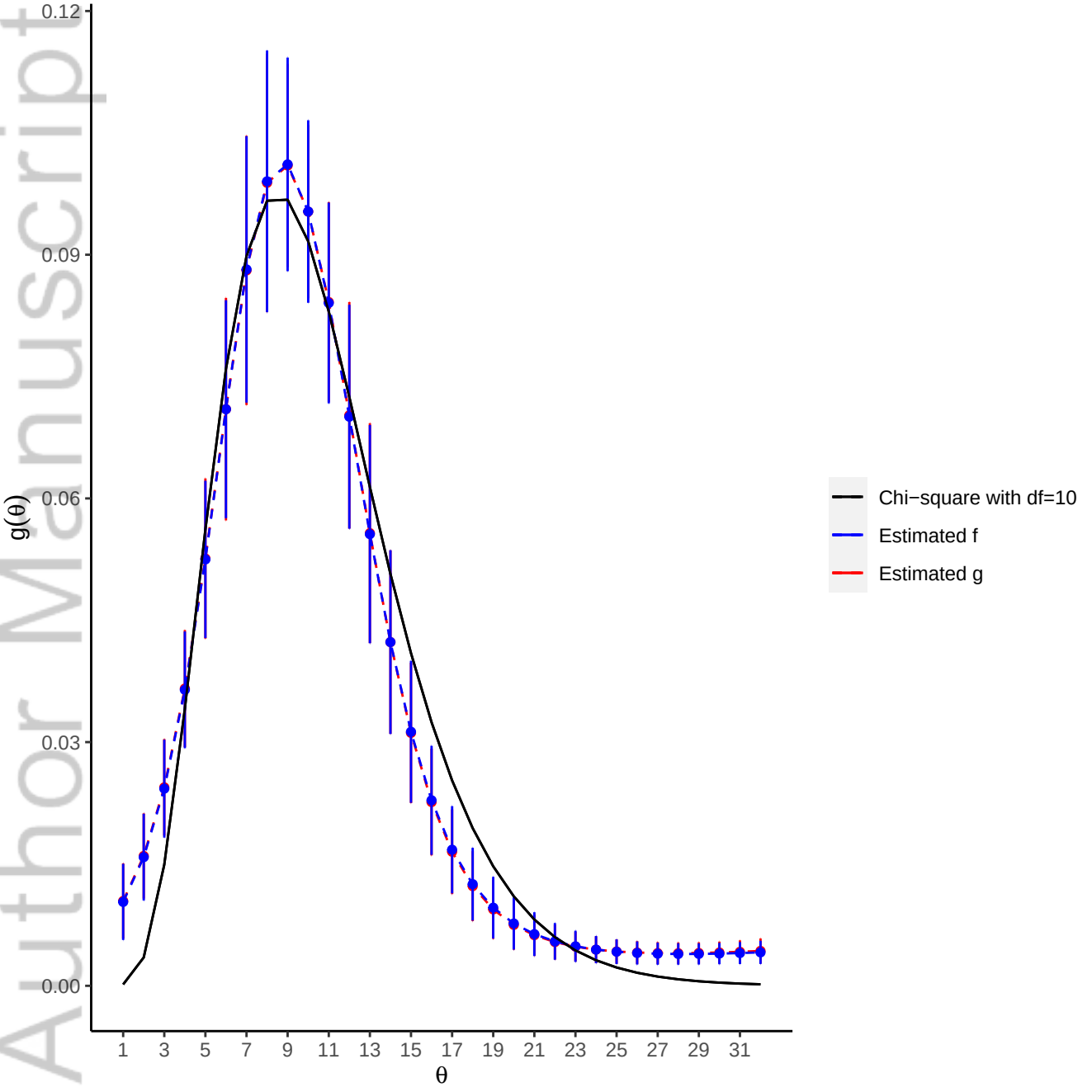


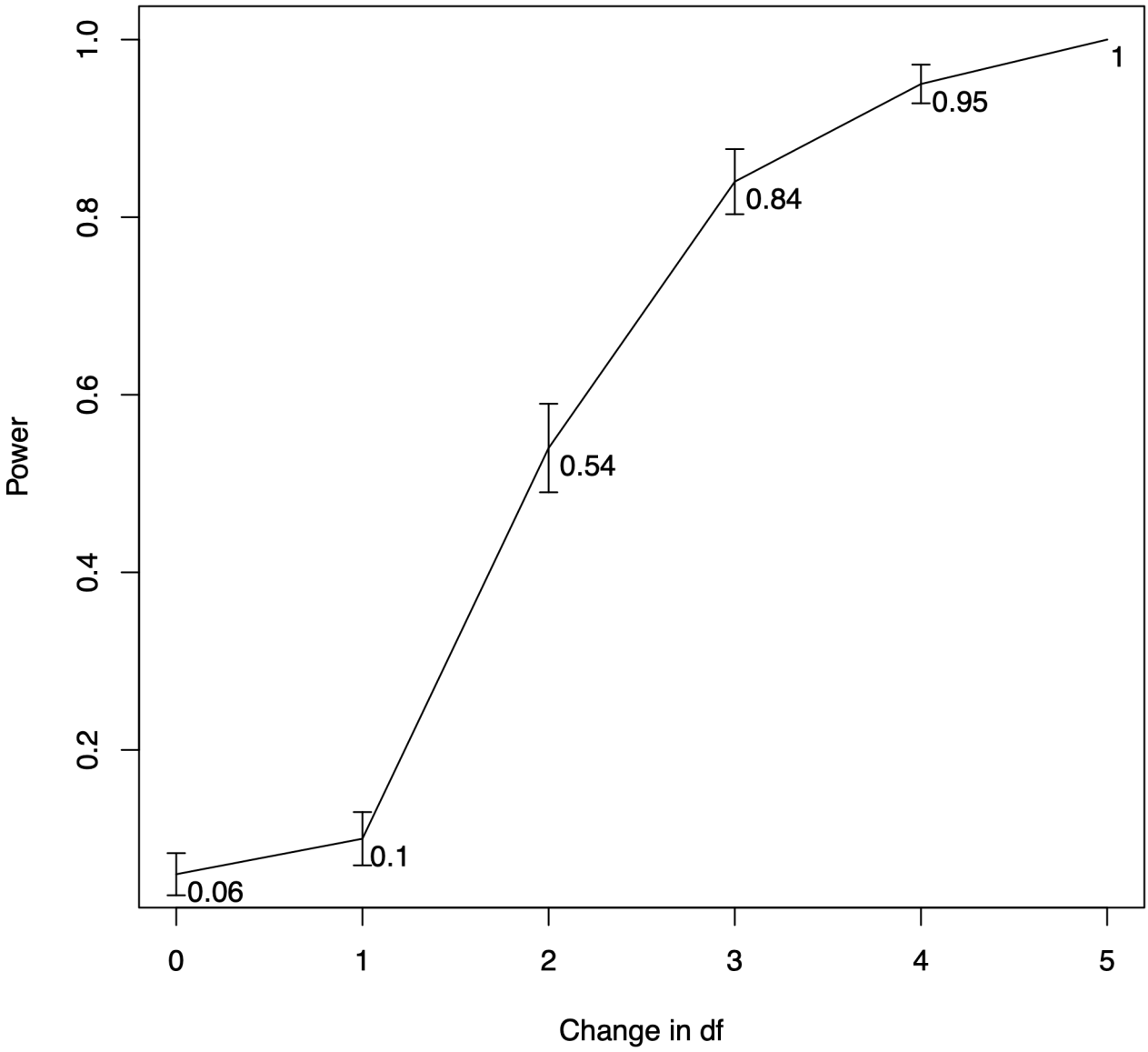


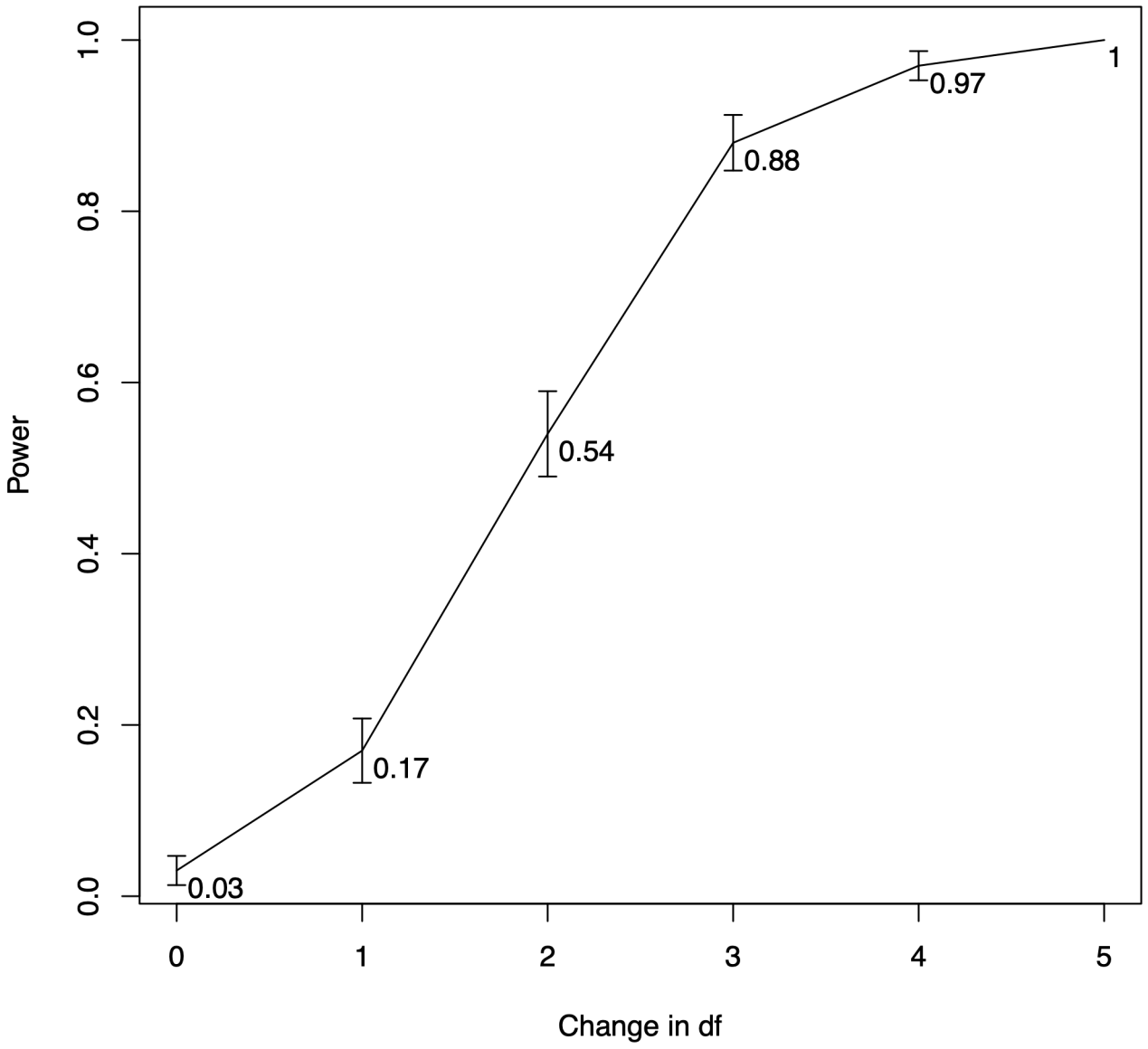


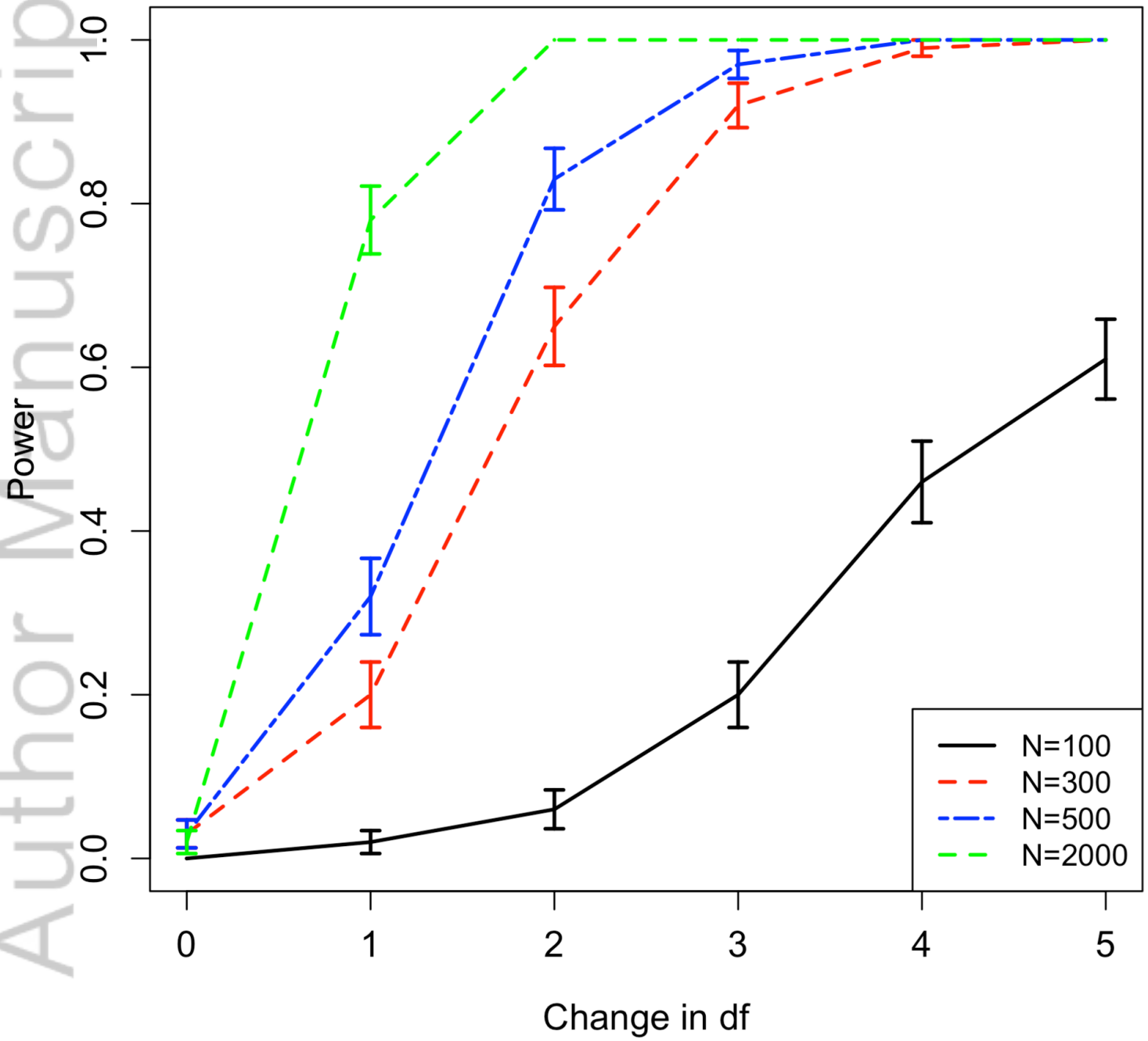




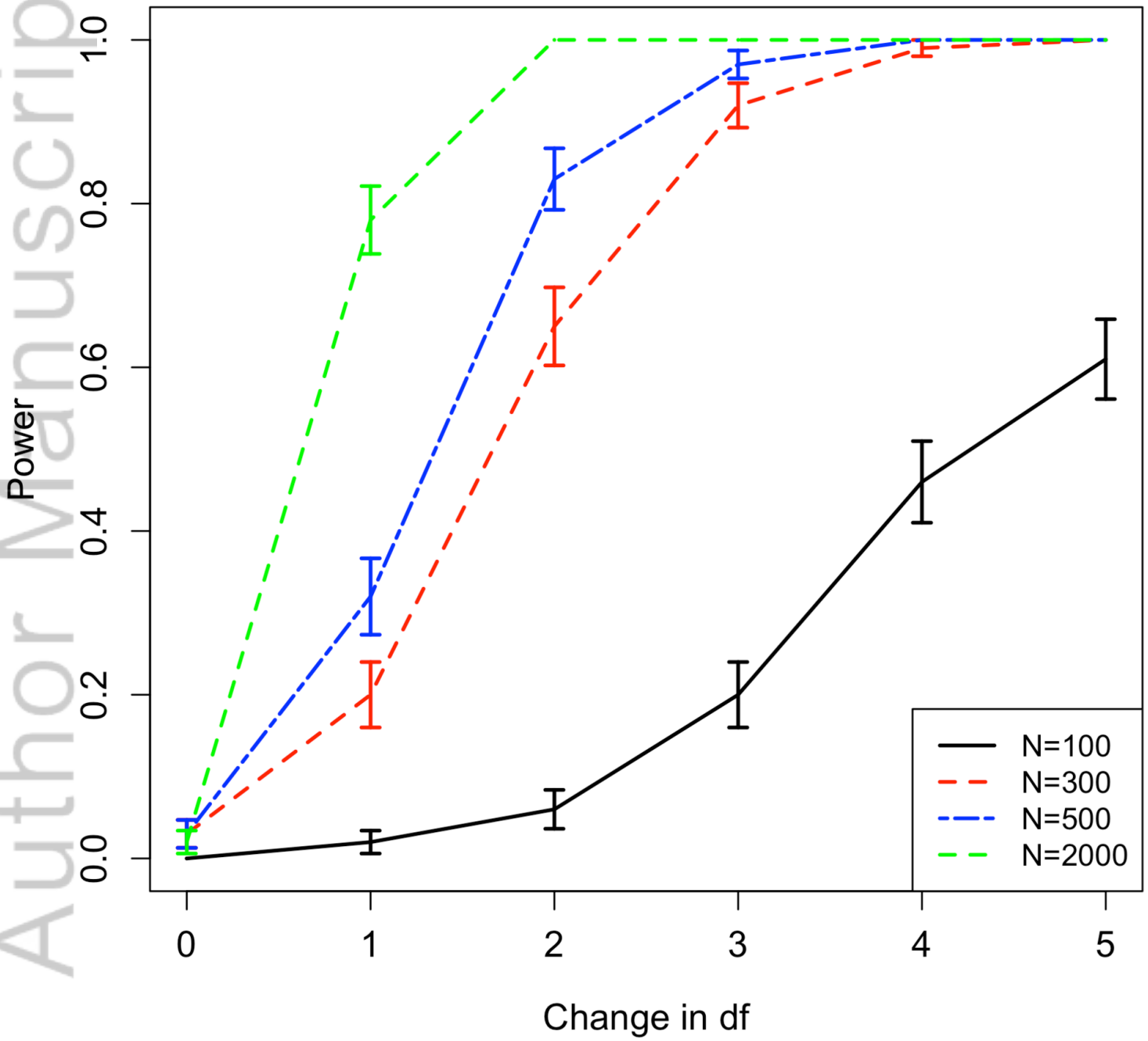




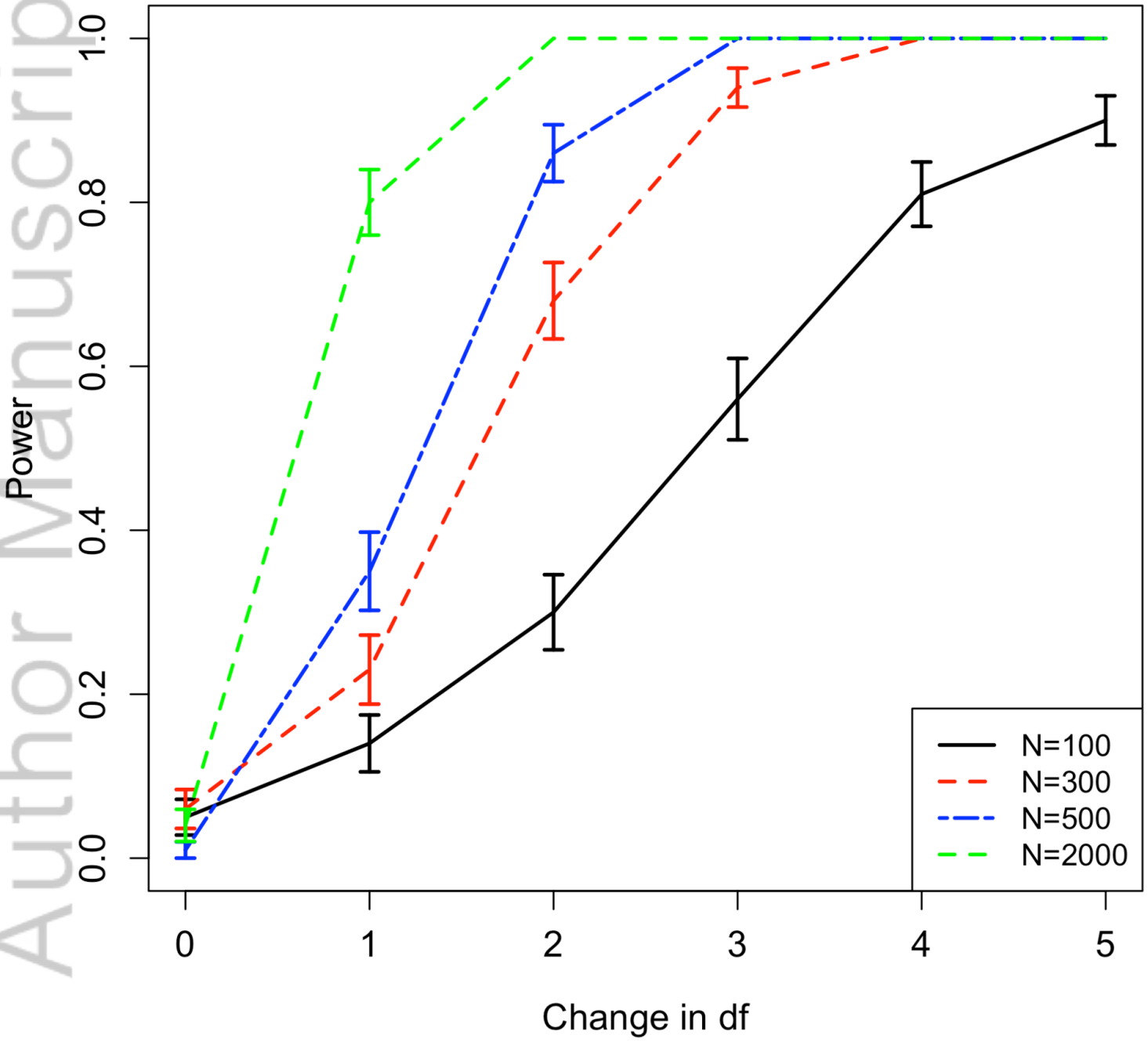




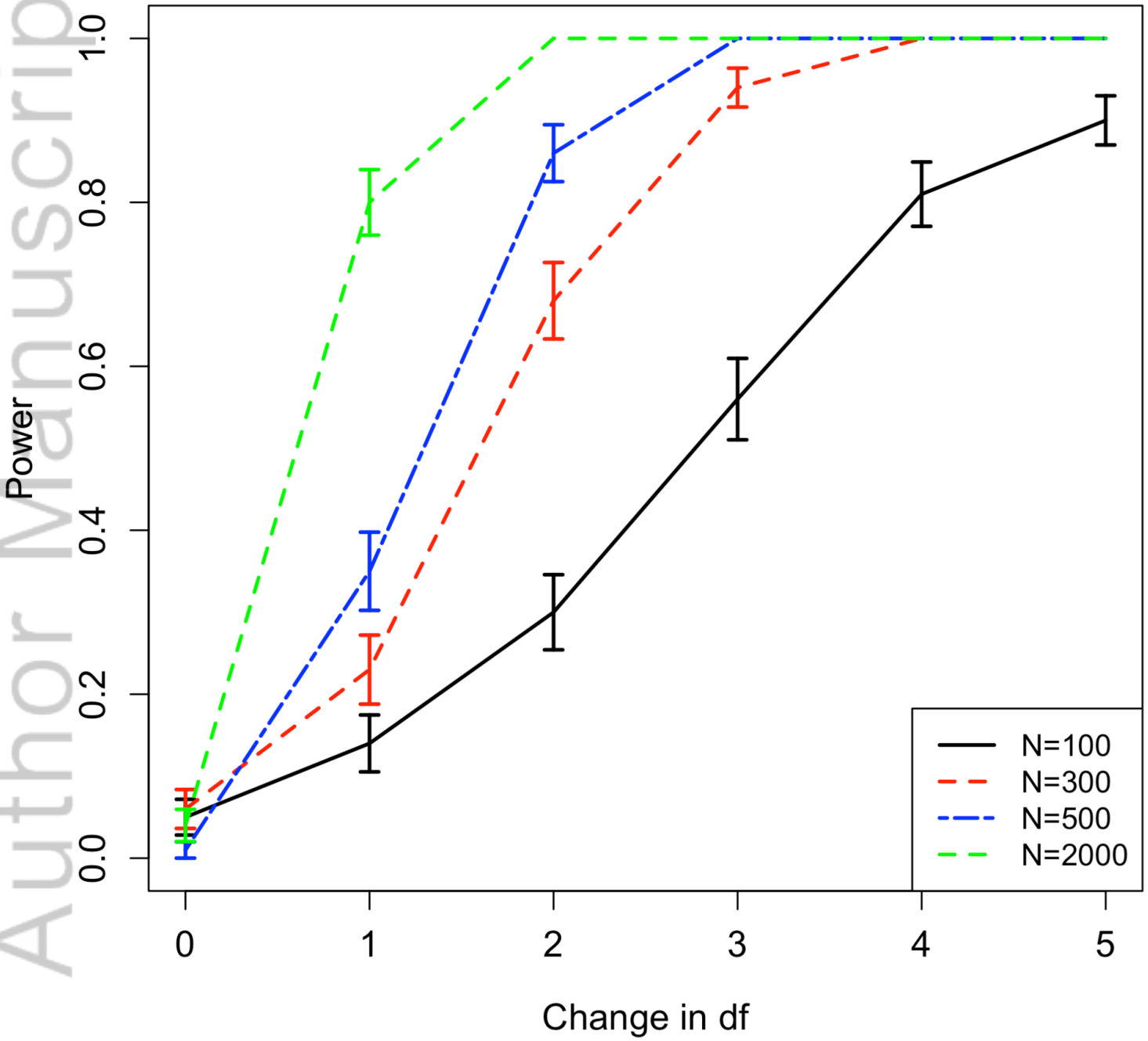
ppasy.png



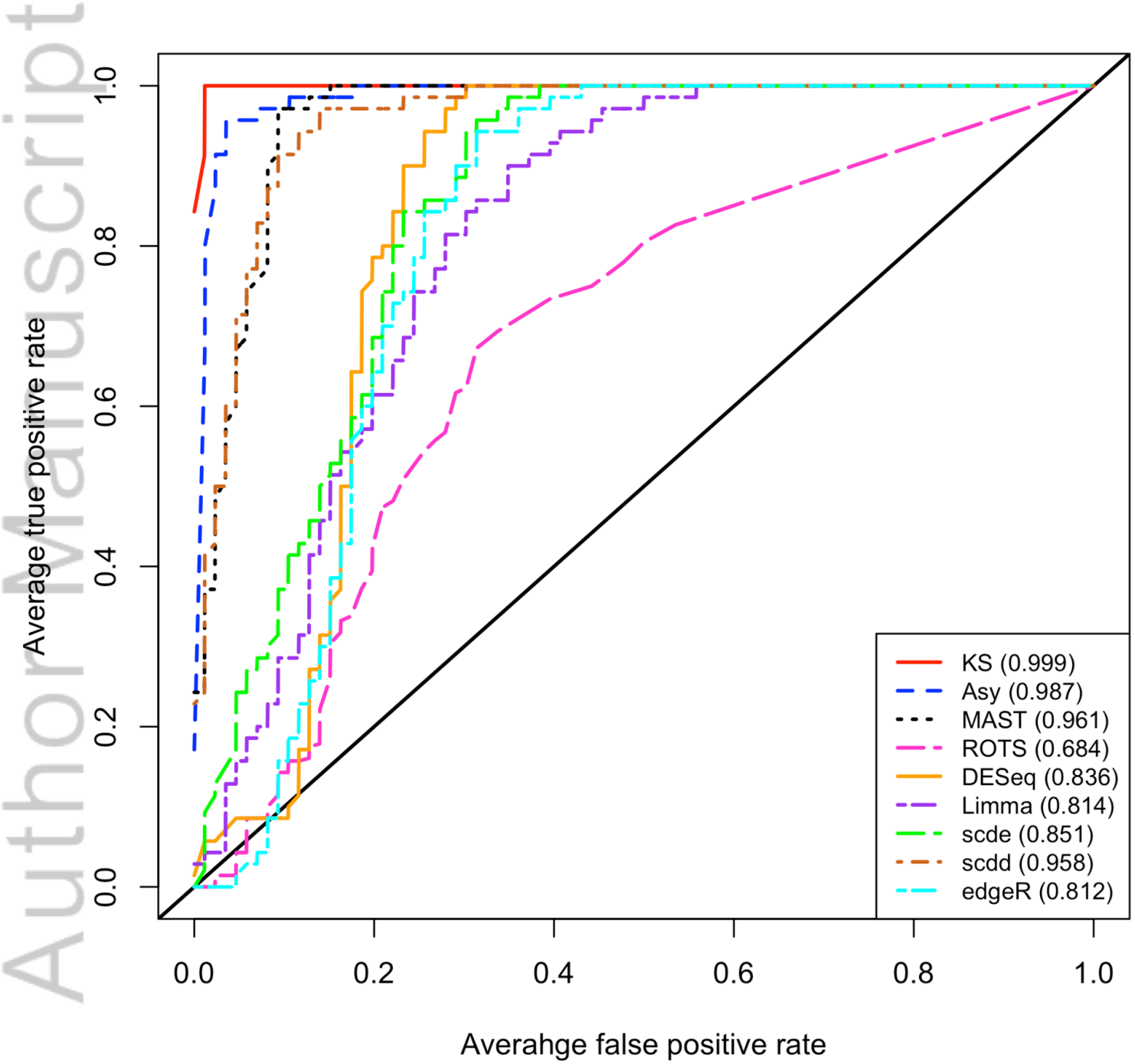
ppasy.png



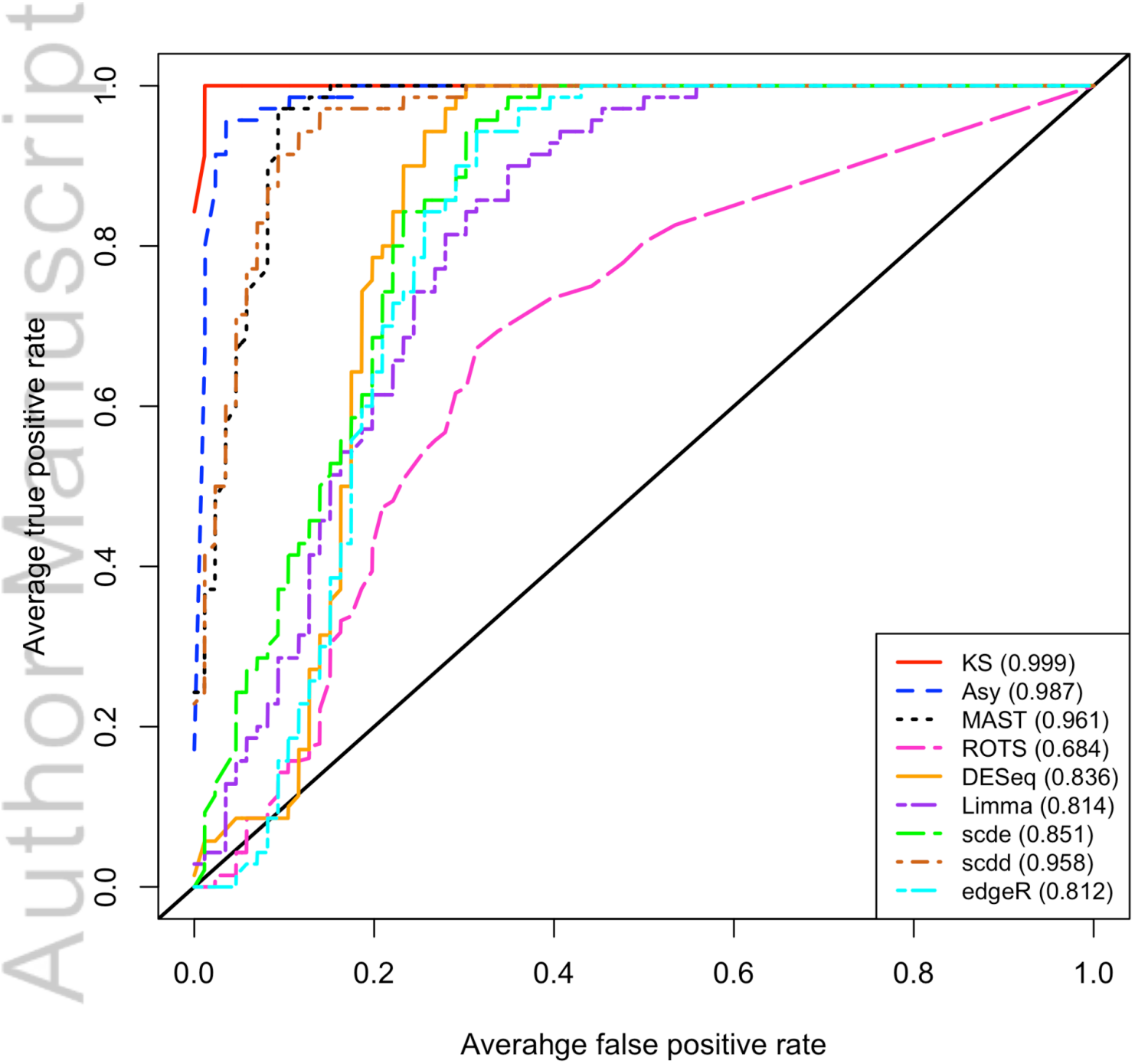
ppks.png



ppks.png



ROCcompare.png



ROCcompare.png