Clipper: p-value-free FDR control on high-throughput data from two conditions

Xinzhou Ge[1,†], Yiling Elaine Chen[1,†], Dongyuan Song[2], MeiLu McDermott[3,4], Kyla Woyshner[3], Antigoni Manousopoulou[3], Wei Li[5], Leo D. Wang[3], and Jingyi Jessica Li[1,2,6,7,8,*]

[1]Department of Statistics, University of California, Los Angeles, CA 90095
[2]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, CA 90095
[3]Beckman Research Institute, City of Hope National Medical Center, Duarte, CA 91010
[4]The Quantitative and Computational Biology section, University of Southern California, Los Angeles, CA 90089
[5]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, CA 92697
[6]Department of Human Genetics, University of California, Los Angeles, CA 90095
[7]Department of Computational Medicine, University of California, Los Angeles, CA 90095
[8]Department of Biostatistics, University of California, Los Angeles, CA 90095
[†]These authors contributed equally to this work.
[*]To whom correspondence should be addressed. Email: jli@stat.ucla.edu

Abstract

High-throughput biological data analysis commonly involves identifying "interesting" features (e.g., genes, genomic regions, and proteins), whose values differ between two conditions, from numerous features measured simultaneously. The most widely-used criterion to ensure the analysis reliability is the false discovery rate (FDR), the expected proportion of uninteresting features among the identified ones. Existing bioinformatics tools primarily control the FDR based on p-values. However, obtaining valid p-values relies on either reasonable assumptions of data distribution or large numbers of replicates under both conditions, two requirements that are often unmet in biological studies. To address this issue, we propose Clipper, a general statistical framework for FDR control without relying on p-values or specific data distributions. Clipper is applicable to identifying both enriched and differential features from high-throughput biological data of diverse types. In comprehensive simulation and real-data benchmarking, Clipper outperforms existing generic FDR control methods and specific bioinformatics tools designed for various tasks, including peak calling from ChIP-seq data, differentially expressed gene identification from RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data. Notably, our benchmarking results for peptide identification are based on the first mass spectrometry data standard with a realistic dynamic range. Our results demonstrate Clipper's flexibility and reliability for FDR control, as well as its broad applications in high-throughput data analysis.

Significance Statement

The reproducibility crisis has been increasingly alarming in biomedical research, which often involves high-throughput data analysis to identify targets for downstream experimental validation. False discovery rate (FDR) is the state-of-the-art criterion to guard reproducibility in such biological data analysis. Existing bioinformatics tools control FDR using p-values, which are usually ill-posed, leading to failed FDR control or poor power. Clipper is a flexible, powerful FDR-control framework that removes the need for high-resolution, well-calibrated p-values. Applicable to various bioinformatics analyses, Clipper

outperforms popular bioinformatics tools, including identifying peaks from ChIP-seq data, differentially expressed genes from RNA-seq data, and differentially interacting chromatin regions from Hi-C data. Clipper is a significant computational advance to addressing the reproducibility crisis in biomedical research.

Introduction

High-throughput technologies are widely used to measure system-wide biological features, such as genes, genomic regions, and proteins ("high-throughput" means the number of features is large, at least in thousands). The most common goal of analyzing high-throughput data is to contrast two conditions so as to reliably screen "interesting features," where "interesting" means "enriched" or "differential." "Enriched features" are defined to have higher expected measurements (without measurement errors) under the experimental (i.e., treatment) condition than the background (i.e., the negative control) condition. The detection of enriched features is called "enrichment analysis." For example, typical enrichment analyses include calling protein-binding sites in a genome from chromatin immunoprecipitation sequencing (ChIP-seq) data [1, 2] and identifying peptides from mass spectrometry (MS) data [3]. In contrast, "differential features" are defined to have different expected measurements between two conditions, and their detection is called "differential analysis." For example, popular differential analyses include the identification of differentially expressed genes (DEGs) from genome-wide gene expression data (e.g., microarray and RNA sequencing (RNA-seq) data [4–10]) and differentially interacting chromatin regions (DIRs) from Hi-C data [11–13] (Fig. 1a). In most scientific research, the interesting features only constitute a small proportion of all features, and the remaining majority is referred to as "uninteresting features."

The identified features, also called the "discoveries" from enrichment or differential analysis, are subject to further investigation and validation. Hence, to reduce experimental validation that is often laborious or expensive, researchers demand reliable discoveries that contain few false discoveries. Accordingly, the false discovery rate (FDR) [14] has been developed as a statistical criterion for ensuring discoveries' reliability. The FDR technically is defined as the expected proportion of uninteresting features among the discoveries under the frequentist statistical paradigm. In parallel, under the Bayesian paradigm, other criteria have been developed, including the Bayesian false discovery rate [15], the local false discovery rate (local fdr) [16], and the local false sign rate [17]. Among all these frequentist and Bayesian criteria, the FDR is the dominant criterion for setting thresholds in biological data analysis [1, 10, 18–24] and is thus the focus of this paper.

FDR control refers to the goal of finding discoveries such that the FDR is under a pre-specified threshold (e.g., 0.05). Existing computational methods for FDR control primarily rely on p-values, one per feature. Among the p-value-based methods, the most classic and popular ones are the Benjamini-Hochberg (BH) procedure [14] and the Storey's q-value [25]; later development introduced methods that incorporate feature weights [26] or covariates (e.g., independent hypothesis weighting (IHW) [27], adaptive p-value thresholding [28], and Boca and Leek's FDR regression [29]) to boost the detection power. All these methods set a p-value cutoff based on the pre-specified FDR threshold. However, the calculation of p-values requires either distributional assumptions, which are often questionable, or large numbers of replicates, which are often unachievable in biological studies (see Results). Due to these limitations of p-value-based methods in high-throughput biological data analysis, bioinformatics tools often output ill-posed p-values. This issue is evidenced by serious concerns about the widespread miscalculation and misuse of p-values in the scientific community [30]. As a result, bioinformatics tools using questionable p-values either cannot reliably control the FDR to a target level [23] or lack power

to make discoveries [31]; see Results. Therefore, p-value-free control of FDR is desirable, as it would make data analysis more transparent and thus improve the reproducibility of scientific research.

Although p-value-free FDR control has been implemented in the MACS2 method for ChIP-seq peak calling [1] and the SAM method for microarray DEG identification [32], these two methods are restricted to specific applications and lack theoretical guarantee for FDR control[1]. More recently, the Barber-Candès (BC) procedure has been proposed to achieve theoretical FDR control without using p-values [35], and it has been shown to perform comparably to the BH procedure with well-calibrated p-values [36]. The BC procedure is advantageous because it does not require well-calibrated p-values, so it holds tremendous potential in various high-throughput data analyses where p-value calibration is challenging [37]. For example, a recent paper has implemented a generalization of the BC procedure to control the FDR in peptide identification from MS data [38].

Inspired by the BC procedure, we propose a general statistical framework Clipper to provide reliable FDR control for high-throughput biological data analysis, without using p-values or relying on specific data distributions. Clipper is a robust and flexible framework that applies to both enrichment and differential analyses and that works for high-throughput data with various characteristics, including data distributions, replicate numbers (from one to multiple), and outlier existence.

Results

Clipper consists of two main steps: construction and thresholding of contrast scores. First, Clipper defines a contrast score for each feature, as a replacement of a p-value, to summarize that feature's measurements between two conditions and to describe the degree of interestingness of that feature. Second, as its name suggests, Clipper establishes a cutoff on features' contrast scores and calls as discoveries the features whose contrast scores exceed the cutoff (see Online Methods and Supplementary). Clipper is a flexible framework that only requires a minimal input: all features' measurements under two conditions and a target FDR threshold (e.g., 5%) (Fig. 1b).

Clipper only relies on two fundamental statistical assumptions of biological data analysis: (1) measurement errors (i.e., differences between measurements and their expectations, with the expectations including both biological signals and batch effects) are independent across all features and replicates; (2) every uninteresting feature has measurement errors identically distributed across all replicates under both conditions. These two assumptions are used in almost all bioinformatics tools and are commonly referred to as the "measurement model" in statistical genomics [39]. With these two assumptions, Clipper has a theoretical guarantee for FDR control under both enrichment and differential analyses with any number of replicates (see Online Methods and Supp. Section S2).

To verify Clipper's performance, we designed comprehensive simulation studies to benchmark Clipper against existing generic FDR control methods (Supp. Section S1). We also benchmarked Clipper against bioinformatics tools in studies including peak calling from ChIP-seq data, peptide identification from mass spectrometry data, DEG identification from RNA-seq data, and DIR identification from Hi-C data. Notably, our benchmarking results for peptide identification are based on the first MS data standard with a realistic dynamic range.

Clipper has verified FDR control and power advantage in simulation

Simulation is essential because we can generate numerous datasets from the same distribution with known truths to calculate the FDR, which is not observable from real data. Our simulation covers both

[1] Although later works have studied some theoretical properties of SAM, they are not about the exact control of the FDR [33, 34].

enrichment and differential analyses. In enrichment analysis, we consider four "experimental designs": 1vs1 design (one replicate per condition), 2vs1 design (two and one replicates under the experimental and background conditions, respectively), 3vs3 design (three replicates per condition), and 10vs10 design (ten replicates per condition). In differential analysis, since Clipper requires that at least one condition has two replicates, we only consider the 2vs1 and 3vs3 designs. For each analysis and design, we simulated data from three "distributional families"—Gaussian, Poisson, and negative binomial—for individual features under two "background scenarios" (i.e., scenarios of the background condition): homogeneous and heterogeneous. Under the homogeneous scenario, all features' measurements follow the same distribution under the background condition; otherwise, we are under the heterogeneous scenario, which is ubiquitous in applications, e.g., identifying DEGs from RNA-seq data and calling protein-binding sites from ChIP-seq data. The detailed simulation settings, each of which is a combination of an experimental design, a distributional family, and a background scenario, are described in Supp. Section S4.

For both enrichment and differential analyses and each simulation setting, we compared Clipper against generic FDR control methods, including p-value-based methods and local-fdr-based methods. The p-value-based methods include BH-pair, BH-pool, qvalue-pair, and qvalue-pool, where "BH" and "qvalue" stand for p-value thresholding procedures, and "pair" and "pool" represent the paired and pooled p-value calculation approaches, respectively. The local-fdr-based methods include locfdr-emp and locfdr-swap, where "emp" and "swap" represent the empirical null and swapping null local-fdr calculation approaches, respectively. See Online Methods for detail.

The comparison results in Fig. 2 and Supp. Figs. S1–S11 show that Clipper controls the FDR and is overall more powerful than other methods, excluding those that fail to control the FDR, under all settings. Clipper is also shown to be more robust to the number of features and the existence of outliers than other methods. In detail, in both enrichment analyses (1vs1, 2vs1, 3vs3, and 10vs10 designs) and differential analyses (2vs1 and 3vs3 designs), Clipper consistently controls the FDR, and it is more powerful than the generic methods in most cases under the realistic, heterogeneous background, where features do not follow the same distribution under the background condition. Under the idealistic, homogeneous background, Clipper is still powerful and only second to BH-pool and qvalue-pool, which, however, cannot control the FDR under the heterogeneous background.

Here we summarize the performance of the generic FDR control methods. First, the two p-value-based methods using the pooled approach, BH-pool and qvalue-pool, are the most powerful under the idealistic, homogeneous background, which is their inherent assumption; however, they cannot control the FDR under the heterogeneous background (Fig. 2b). Besides, they cannot control the FDR when the number of features is small (Fig. 2a and Supp. Fig. S1). These results show that the validity of BH-pool and qvalue-pool requires a large number of features and the homogeneous background assumption, two requirements that rarely hold in biological applications. Second, the four p-value-based methods using the paired approach with misspecified models or misformulated tests (BH-pair-mis, qvalue-pair-mis, BH-pair-2as1, and qvalue-pair-2as1; see Online Methods) fail to control the FDR by a large margin in most cases, and rarely when they control the FDR, they lack power (Fig. 2c–d and Supp. Figs. S1–S8). These results confirm that the BH-pair and qvalue-pair rely on the correct model specification to control the FDR; however, the correct model specification is hardly achievable with no more than three replicates per condition. Third, even when models are correctly specified (an idealistic scenario), the p-value-based methods that use the paired approach—BH-pair-correct and qvalue-pair-correct (see Online Methods)—fail to control the FDR in the existence of outliers (Fig. 2e and Supp. Figs. S3 and S7) or for the negative binomial distribution with unknown dispersion (Fig. 2f and Supp. Fig. S9). It is worth noting that even when they control the FDR, they are less powerful than Clipper in most cases except for the 3vs3 differential analysis with the Poisson distribution (Fig. 2d and Supp.

Figs. S4 and S8). Fourth, the two local-fdr-based methods—locfdr-emp and locfdr-swap—achieve the FDR control under all designs and analyses; however, they are less powerful than Clipper in most cases (Supp. Figs. S1–S4). Fifth, when the numbers of replicates are large (10vs10 design), non-parametric tests become applicable. We compared Clipper with three BH-pair methods that use different statistical tests: BH-pair-Wilcoxon (the non-parametric Wilcoxon rank-sum test), BH-pair-permutation (the non-parametric permutation test), and BH-pair-parametric (the parametric test based on the correct model specification, equivalent to BH-pair-correct). Although all the three methods control the FDR, they are less powerful than Clipper (Supp. Fig. S10).

The above results are all based on simulations with independent features. To examine the robustness of Clipper, we introduced feature correlations to our simulated data, on which we compared Clipper with other generic FDR control methods. The comparison results in Supp. Fig. S11 show that even when the feature independence assumption is violated, Clipper still demonstrates strong performance in both FDR control and power.

Clipper has broad applications in omics data analyses

We then demonstrate the use of Clipper in four omics data applications: peak calling from ChIP-seq data, peptide identification from MS data, DEG identification from RNA-seq data, and DIR identification from Hi-C data. The first two applications are enrichment analyses, and the last two are differential analyses. In each application, we compared Clipper with mainstream bioinformatics methods to demonstrate Clipper's superiority in FDR control and detection power.

Peak calling from ChIP-seq data (enrichment analysis I)

ChIP-seq is a genome-wide experimental assay for measuring binding intensities of a DNA-associated protein [40], often a transcription factor that activates or represses gene expression [41, 42]. ChIP-seq data are crucial for studying gene expression regulation, and the indispensable analysis is to identify genomic regions with enriched sequence reads in ChIP-seq data. These regions are likely to be bound by the target protein and thus of biological interest. The identification of these regions is termed "peak calling" in ChIP-seq data analysis.

As the identified peaks are subject to experimental validation that is often expensive [43], it is essential to control the FDR of peak identification to reduce unnecessary costs. The two most highly-cited peak-calling methods are MACS2 [1] and [2], both of which claim to control the FDR for their identified peaks. Specifically, both MACS2 and HOMER assume that the read counts for each putative peak (one count per sample/replicate) follow the Poisson distribution, and they use modified paired approaches to assign each putative peak a p-value and a corresponding Storey's q-value. Then given a target FDR threshold $0 < q < 1$, they call the putative peaks with q-values $\leq q$ as identified peaks. Despite being popular, MACS2 and HOMER have not been verified for their FDR control, to our knowledge.

To verify the FDR control of MACS2 and HOMER (Supp. Section S5.1), we used ENCODE ChIP-seq data of cell line GM12878 [44] and generated semi-synthetic data with spiked-in peaks (Supp. Section S6.1). We examined the actual FDR and power of MACS2 and HOMER in a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3a shows that MACS2 and HOMER cannot control the FDR as standalone peak-calling methods. However, with Clipper as an add-on (Supp. Section S7.1), both MACS2 and HOMER can guarantee the FDR control. This result demonstrates the flexibility and usability of Clipper for reducing false discoveries in peak calling analysis.

Technically, the failed FDR control by MACS2 and HOMER is attributable to the likely model misspecification and test misformulation in their use of the paired approach. Both MACS2 and HOMER assume

the Poisson distribution for read counts in a putative peak; however, it has been widely acknowledged that read counts are over-dispersed and thus better modeled by the negative binomial distribution [45]. Besides, MACS2 uses one-sample tests to compute p-values when two-sample tests should have been performed. As a result, the p-values of MACS2 and HOMER are questionable, so using their p-values for FDR control would not lead to success. (Note that MACS2 does not use p-values to control the FDR but instead swaps experimental and background samples to calculate the empirical FDR; yet, we emphasize that controlling the empirical FDR does not guarantee the FDR control.) As a remedy, Clipper strengthens both methods to control the FDR while maintaining high power.

Peptide identification from MS data (enrichment analysis II)

The state-of-the-art proteomics studies use MS experiments and database search algorithms to identify and quantify proteins in biological samples. In a typical proteomics experiment, a protein mixture sample is first digested into peptides and then measured by tandem MS technology as mass spectra, which encode peptide sequence information. "Peptide identification" is the process that decodes mass spectra and converts mass spectra into peptide sequences in a protein sequence database via search algorithms. The search process matches each mass spectrum to peptide sequences in the database and outputs the best match, called a "peptide-spectrum match" (PSM). The identified PSMs are used to infer and quantify proteins in a high-throughput manner.

False PSMs could occur when mass spectra are matched to wrong peptide sequences due to issues such as low-quality spectra, data-processing errors, and incomplete protein databases, causing problems in the downstream protein identification and quantification [46]. Therefore, a common goal of database search algorithms is to simultaneously control the FDR and maximize the number of identified PSMs, so as to maximize the number of proteins identified in a proteomics study [3, 47, 48]. A widely used FDR control strategy is the target-decoy search, where mass spectra of interest are matched to peptide sequences in both the original (target) database and a decoy database that contains artificial false protein sequences. The resulting PSMs are called the target PSMs and decoy PSMs, respectively. The decoy PSMs, i.e., matched mass spectrum and decoy peptide pairs, are known to be false and thus used by database search algorithms to control the FDR. Mainstream database search algorithms output a q-value for each PSM, target or decoy. Discoveries are the target PSMs whose q-values are no greater than the target FDR threshold q.

We used the first comprehensive benchmark dataset from an archaea species *Pyrococcus furiosus* to examine the FDR control and power of a popular database search algorithm Mascot [3] (Supp. Section S5.2). Using this benchmark dataset (Supp. Section S6.2), we demonstrate that, as an add-on, Clipper improves the power of Mascot. Specifically, Clipper treats mass spectra as features. For each mass spectrum, Clipper considers its measurement under the experimental condition as the $-\log_{10}$-transformed q-value of the target PSM that includes it, and its measurement under the background condition as the $-\log_{10}$-transformed q-value of the decoy PSM that includes it. Then Clipper decides which mass spectra and their corresponding target PSMs are discoveries (Supp. Section S7.2). Based on the benchmark dataset, we examined the empirical FDR, i.e., the FDP calculated based on the true positives and negatives, and the power of Mascot with or without Clipper as an add-on, for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3b shows that although Mascot and Mascot+Clipper both control the FDR, Mascot+Clipper consistently improves the power, thus enhancing the peptide identification efficiency of proteomics experiments.

While preparing this manuscript, we found a recent work [38] that used a similar idea to identify PSMs without using p-values. Clipper differs from this work in two aspects: (1) Clipper is directly applicable as an add-on to any existing database search algorithms that output q-values; (2) Clipper is not restricted

to the peptide identification application.

DEG identification from RNA-seq data (differential analysis I)

RNA-seq data measure genome-wide gene expression levels in biological samples. An important use of RNA-seq data is the DEG analysis, which aims to discover genes whose expression levels change between two conditions. The FDR is a widely used criterion in DEG analysis [4–9].

We compared Clipper with two popular DEG identification methods DESeq2 [5] and edgeR [4] (Supp. Section S5.3). Specifically, we applied Clipper to two gene expression matrices, one per condition, to identify DEGs (Supp. Section S7.3). To verify the FDR control of Clipper, DESeq2, and edgeR, we generated realistic semi-synthetic data from real RNA-seq data of classical and non-classical human monocytes [49], i.e., two conditions (Supp. Section S6.3). With ground truths (true DEGs and non-DEGs), the semi-synthetic data allow evaluating each method's FDR and power for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3c shows that Clipper consistently controls the FDR and achieves high power. In contrast, edgeR controls the FDR but has lower power than Clipper does, while DESeq2 fails to control the FDR. To explain this DESeq2 result, we examined the p-value distributions of 16 non-DEGs that were most frequently identified from 200 synthetic datasets by DESeq2 at the target FDR threshold $q = 0.05$. Our results in Fig. S12 show that the 16 non-DEGs' p-values are non-uniformly distributed with a mode close to 0. Such unusual enrichment of overly small p-values makes these non-DEGs mistakenly called discoveries by DESeq2.

IHW is a popular procedure for boosting the power of p-value-based FDR control methods by incorporating feature covariates [27]. We used IHW as an add-on for DESeq2 and edgeR by adding every gene's mean expression across replicates and conditions as that gene's covariate, as suggested in [50]. The result in Supp. Fig. 3c shows that DESeq2+IHW and edgeR+IHW are similar to DESeq2 and edgeR, respectively, in FDR control and power.

We further compared Clipper with DESeq2 and edgeR on the real RNA-seq data of classical and non-classical human monocytes [49]. In this dataset, gene expression changes are expected to be associated with the immune response process. We input three classical and three non-classical samples into Clipper, DESeq2, and edgeR for DEG identification. Fig. 4a shows that edgeR identifies the fewest DEGs, while DESeq2 identifies the most DEGs, followed by Clipper. Notably, most DEGs identified by DESeq2 are not identified by Clipper or edgeR. To investigate whether DESeq2 makes too many false discoveries and whether the DEGs found by Clipper but missed by DESeq2 or edgeR are biologically meaningful, we performed functional analysis on the set of DEGs identified by each method. We first performed the gene ontology (GO) analysis on the three sets of identified DEGs using the R package `clusterProfiler` [51]. Fig. 4b ("Total") shows that more GO terms are enriched (with enrichment q-values ≤ 0.01) in the DEGs identified by Clipper than in the DEGs identified by DESeq2 or edgeR. For the GO terms enriched in all three sets of identified DEGs, Fig. 4c shows that they are all related to the immune response and thus biologically meaningful. Notably, these biologically meaningful GO terms have more significant enrichment in the DEGs identified by Clipper than in those identified by edgeR and DESeq2. We further performed GO analysis on the DEGs uniquely identified by one method in pairwise comparisons of Clipper vs. DESeq2 and Clipper vs. edgeR. Fig. 4b and Supp. Fig. S13 show that multiple immune-related GO terms are enriched in Clipper-specific DEGs, while no GO terms are enriched in edgeR-specific or DESeq2-specific DEGs. Altogether, these results confirm the capacity of Clipper in real-data DEG analysis, and they are consistent with our simulation results that edgeR lacks power, while DESeq2 fails to control the FDR.

DIR analysis of Hi-C data (differential analysis II)

Hi-C experiments are widely used to investigate spatial organizations of chromosomes and map chromatin interactions across the genome. A Hi-C dataset is often processed and summarized into an interaction matrix, whose rows and columns represent manually binned chromosomal regions and whose (i, j)-th entry represents the measured contact intensity between the i-th and j-th binned regions. The DIR analysis aims to identify pairs of genomic regions whose interaction intensities differ between conditions. Same as DEG analysis, DIR analysis also uses the FDR as a decision criterion [11–13].

We compared Clipper with three popular DIR identification methods: diffHic [13], FIND [12], and multiHiCcompare [11] (Supp. Section S5.4). Specifically, we applied Clipper to DIR identification by treating pairs of genomic regions as features and interaction intensities as measurements. To verify the FDR control of Clipper (Supp. Section S7.4), diffHiC, FIND, and multiHiCcompare, we generated realistic synthetic data from real interaction matrices of ENCODE cell line GM12878 [44] with true spiked-in DIRs to evaluate the FDR and power (Supp. Section S6.4). We examined the actual FDR and power in a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3d shows that Clipper and diffHic are the only two methods that consistently control the FDR, while multiHiCcompare and FIND fail by a large margin. In terms of power, Clipper outperforms diffHic except for $q = 0.01$ and 0.02, even though Clipper has not been optimized for Hi-C data analysis. This result demonstrates Clipper's general applicability and strong potential for DIR analysis.

Discussion

In this paper, we proposed a new statistical framework, Clipper, for identifying interesting features with FDR control from high-throughput data. Clipper avoids the use of p-values and makes FDR control more reliable and flexible. We used comprehensive simulation studies to verify the FDR control by Clipper under various settings. We demonstrate that Clipper outperforms existing generic FDR control methods by having higher power and greater robustness to model misspecification. We further applied Clipper to four popular bioinformatics analyses: peak calling from ChIP-seq data, peptide identification from MS data, DEG identification from RNA-seq data, and DIR identification from Hi-C data. Our results indicate that Clipper can provide a powerful add-on to existing bioinformatics tools to improve the reliability of FDR control and thus the reproducibility of scientific discoveries.

We validated the FDR control by Clipper using extensive and concrete simulations, including both model-based and real-data-based data generation with ground truths. In contrast, in most bioinformatics method papers, the FDR control was merely mentioned but rarely validated. Many of them assumed that using the BH procedure on p-values would lead to valid FDR control; however, the reality is often otherwise because p-values would be invalid when model assumptions were violated or the p-value calculation was problematic. Here we voice the importance of validating the FDR control in bioinformatics method development, and we use this work as a demonstration. We believe that Clipper provides a powerful booster to this movement. As a p-value-free alternative to the classic p-value-based BH procedure, Clipper relies less on model assumptions and is thus more robust to model misspecifications, making it an appealing choice for FDR control in diverse high-throughput biomedical data analyses.

Clipper is a flexible framework that is easily generalizable to identify a variety of interesting features. The core component of Clipper summarizes each feature's measurements under each condition into an informative statistic (e.g., the sample mean); then Clipper combines each feature's informative statistics under two conditions into a contrast score to enable FDR control. The current implementation of Clipper only uses the sample mean as the informative statistic to identify the interesting features that have distinct expected values under two conditions. However, by modifying the informative statistic, we can

generalize Clipper to identify the features that are interesting in other aspects, e.g., having different variances between two conditions. Regarding the contrast score, Clipper makes careful choices between two contrast scores, difference and maximum, based on the number of replicates and the analysis task (enrichment vs. differential). Future studies are needed to explore other contrast scores and their power with respect to data characteristics and analysis tasks.

We have demonstrated the broad application potential of Clipper in various bioinformatics data analyses. Specifically, when used as an add-on to established, popular bioinformatics methods such as MACS2 for peak calling and Mascot for peptide identification, Clipper guaranteed the desired FDR control and in some cases boosted the power. However, many more careful thoughts are needed to escalate Clipper into standalone bioinformatics methods for specific data analyses, for which data processing and characteristics (e.g., peak lengths, GC contents, proportions of zeros, and batch effects) must be appropriately accounted for before Clipper is used for the FDR control [52, 53]. We expect that the Clipper framework will propel future development of bioinformatics methods by providing a flexible p-value-free approach to control the FDR, thus improving the reliability of scientific discoveries.

After finishing this manuscript, we were informed of the work by He et al. [54], which is highly similar to the part of Clipper for differential analysis, as both work use permutation for generating negative controls and the GZ procedure for thresholding (test statistics in He et al. and contrast scores in Clipper). However, the test statistics used in He et al. are the two-sample t statistic and the two-sample Wilcoxon statistic, both of which are different from the difference and maximum contrast scores used in Clipper. While we leave the optimization of contrast scores to future work, we note that the two-sample Wilcoxon statistic, though being a valid contrast score for differential analysis, requires a large sample size to achieve good power. For this reason, we did not consider it as a contrast score in the current Clipper implementation, whose focus is on sample-sample-size high-throughout biological data.

Online Methods

Notations and assumptions

We first introduce notations and assumptions used in Clipper. While the differential analysis treats the two conditions symmetric, the enrichment analysis requires one condition to be the experimental condition (i.e., the condition of interest) and the other condition to be the background condition (i.e., the negative control). For simplicity, we use the same set of notations for both analyses. For two random vectors $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, we write $\boldsymbol{X} \perp \boldsymbol{Y}$ if X_i is independent of Y_j for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. To avoid confusion, we use $\mathrm{card}(A)$ to denote the cardinality of a set A and $|c|$ to denote the absolute value of a scalar c. We define $a \vee b := \max(a, b)$.

Clipper only requires two inputs: the target FDR threshold $q \in (0, 1)$ and the input data. Regarding the input data, we use d to denote the number of features with measurements under two conditions, and we use m and n to denote the numbers of replicates under the two conditions. For each feature $j = 1, \ldots, d$, we use $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top \in \mathbb{R}^m_{\geq 0}$ and $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top \in \mathbb{R}^n_{\geq 0}$ to denote its measurements under the two conditions, where $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers. We assume that all measurements are non-negative, as in the case of most high-throughput experiments. (If this assumption does not hold, transformations can be applied to make data satisfy this assumption.)

Clipper has the following assumptions on the joint distribution of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$. For $j = 1, \ldots, d$, Clipper assumes that $X_{j1}, \ldots, X_{jm}$ are identically distributed, so are $Y_{j1}, \ldots, Y_{jn}$. Let $\mu_{Xj} = \mathbb{E}[X_{j1}]$ and $\mu_{Yj} = \mathbb{E}[Y_{j1}]$ denote the expected measurement of feature j under the two conditions,

respectively. Then conditioning on $\{\mu_{Xj}\}_{j=1}^d$ and $\{\mu_{Yj}\}_{j=1}^d$,

$$X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are mutually independent}; \quad (1)$$
$$\boldsymbol{X}_j \perp \boldsymbol{X}_k, \boldsymbol{Y}_j \perp \boldsymbol{Y}_k \text{ and } \boldsymbol{X}_j \perp \boldsymbol{Y}_k, \ \forall j, k = 1, \ldots, d.$$

An enrichment analysis aims to identify interesting features with $\mu_{Xj} > \mu_{Yj}$ (with $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ defined as the measurements under the experimental and background conditions, respectively), while a differential analysis aims to call interesting features with $\mu_{Xj} \neq \mu_{Yj}$. We define $\mathcal{N} := \{j : \mu_{Xj} = \mu_{Yj}\}$ as the set of uninteresting features and denote $N := \mathrm{card}(\mathcal{N})$. In both analyses, Clipper further assumes that an uninteresting feature j satisfies

$$X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are identically distributed}, \forall j \in \mathcal{N}. \quad (2)$$

Clipper consists of two main steps: construction and thresholding of contrast scores. First, Clipper computes contrast scores, one per feature, as summary statistics that reflect the extent to which features are interesting. Second, Clipper establishes a contrast-score cutoff and calls as discoveries the features whose contrast scores exceed the cutoff.

To construct contrast scores, Clipper uses two summary statistics $t(\cdot, \cdot) : \mathbb{R}_{\geq 0}^m \times \mathbb{R}_{\geq 0}^n \to \mathbb{R}$ to extract data information regarding whether a feature is interesting or not:

$$t^{\mathrm{diff}}(\boldsymbol{x}, \boldsymbol{y}) := \bar{x} - \bar{y}; \quad (3)$$
$$t^{\max}(\boldsymbol{x}, \boldsymbol{y}) := \max(\bar{x}, \bar{y}) \cdot \mathrm{sign}(\bar{x} - \bar{y}), \quad (4)$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^\top \in \mathbb{R}_{\geq 0}^m$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}_{\geq 0}^n$, $\bar{x} = \sum_{i=1}^m x_i/m$, $\bar{y} = \sum_{i=1}^n y_i/n$, and $\mathrm{sign}(\cdot) : \mathbb{R} \to \{-1, 0, 1\}$ with $\mathrm{sign}(x) = 1$ if $x > 0$, $\mathrm{sign}(x) = -1$ if $x < 0$, and $\mathrm{sign}(x) = 0$ otherwise.

Then we introduce how Clipper works in three analysis tasks: the enrichment analysis with equal numbers of replicates under two conditions ($m = n$), the enrichment analysis with different numbers of replicates under two conditions ($m \neq n$), and the differential analysis (when $m + n > 2$).

Enrichment analysis with equal numbers of replicates ($m = n$)

Under the enrichment analysis, we assume that $\boldsymbol{X}_j \in \mathbb{R}_{\geq 0}^m$ and $\boldsymbol{Y}_j \in \mathbb{R}_{\geq 0}^n$ are the measurements of feature j, $j = 1, \ldots, d$, under the experimental and background conditions with m and n replicates, respectively. We start with the simple case when $m = n$. Clipper defines a contrast score C_j of feature j in one of two ways:

$$C_j := t^{\mathrm{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{difference contrast score}, \quad (5)$$

or

$$C_j := t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{maximum contrast score}. \quad (6)$$

Fig. 5a shows a cartoon illustration of contrast scores when $m = n = 1$. Accordingly, a large positive value of C_j bears evidence that $\mu_{Xj} > \mu_{Yj}$. Motivated by Barber and Candès [35], Clipper uses the following procedure to control the FDR under the target level $q \in (0, 1)$.

Definition 1 (Barber-Candès (BC) procedure for thresholding contrast scores [35]) *Given contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0 \,;\, j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The BC procedure finds a contrast-score cutoff T^{BC} based on the target FDR threshold $q \in (0, 1)$*

as

$$T^{\mathrm{BC}} := \min\left\{t \in \mathcal{C} : \frac{\mathrm{card}(\{j : C_j \le -t\}) + 1}{\mathrm{card}(\{j : C_j \ge t\}) \vee 1} \le q\right\} \tag{7}$$

and outputs $\{j : C_j \ge T^{\mathrm{BC}}\}$ *as discoveries.*

Enrichment analysis with any numbers of replicates m and n

When $m \neq n$, Clipper constructs contrast scores via permutation of replicates across conditions. The idea is that, after permutation, every feature becomes uninteresting and can serve as its own negative control.

Definition 2 (Permutation) *We define σ as permutation, i.e., a bijection from the set $\{1, \cdots, m+n\}$ onto itself, and we rewrite the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$ into a matrix $\mathbf{W}$:*

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1m} & W_{1(m+1)} & \cdots & W_{1(m+n)} \\ & \vdots & & & \vdots & \\ W_{d1} & \cdots & W_{dm} & W_{d(m+1)} & \cdots & W_{d(m+n)} \end{bmatrix} := \begin{bmatrix} X_{11} & \cdots & X_{1m} & Y_{11} & \cdots & Y_{1n} \\ & \vdots & & & \vdots & \\ X_{d1} & \cdots & X_{dm} & Y_{d1} & \cdots & Y_{dn} \end{bmatrix}.$$

We then apply σ to permute the columns of $\mathbf{W}$ and obtain

$$\mathbf{W}_\sigma := \begin{bmatrix} W_{1\sigma(1)} & \cdots & W_{1\sigma(m)} & W_{1\sigma(m+1)} & \cdots & W_{1\sigma(m+n)} \\ & \vdots & & & \vdots & \\ W_{d\sigma(1)} & \cdots & W_{d\sigma(m)} & W_{d\sigma(m+1)} & \cdots & W_{d\sigma(m+n)} \end{bmatrix},$$

from which we obtain the permuted measurements $\{(\boldsymbol{X}_j^\sigma, \boldsymbol{Y}_j^\sigma)\}_{j=1}^d$, where

$$\begin{aligned} \boldsymbol{X}_j^\sigma &:= \left(W_{j\sigma(1)}, \ldots, W_{j\sigma(m)}\right)^\top, \\ \boldsymbol{Y}_j^\sigma &:= \left(W_{j\sigma(m+1)}, \ldots, W_{j\sigma(m+n)}\right)^\top. \end{aligned} \tag{8}$$

In the enrichment analysis, if two permutations σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\},$$

then we define σ and σ' to be in one equivalence class. That is, permutations in the same equivalence class lead to the same division of $m+n$ replicates (from the two conditions) into two groups with sizes m and n. In total, there are $\binom{m+n}{m}$ equivalence classes of permutations.

We define σ_0 as the identity permutation such that $\sigma_0(i) = i$ for all $i \in \{1, \cdots, m+n\}$. In addition, Clipper randomly samples h equivalence classes $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the other $h_{\max} := \binom{m+n}{m} - 1$ equivalence classes (after excluding the equivalence class containing σ_0). Note that $h_{\max}$ is the maximum value h can take.

Clipper then obtains $\{(\boldsymbol{X}_j^{\sigma_0}, \boldsymbol{Y}_j^{\sigma_0}), (\boldsymbol{X}_j^{\sigma_1}, \boldsymbol{Y}_j^{\sigma_1}), \cdots, (\boldsymbol{X}_j^{\sigma_h}, \boldsymbol{Y}_j^{\sigma_h})\}_{j=1}^d$, where $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ are the permuted measurements based on σ_ℓ, $\ell = 0, 1, \ldots, h$. Then Clipper computes $T_j^{\sigma_\ell} := t^{\mathrm{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ to indicate the degree of "interestingness" of feature j reflected by $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$. Note that Clipper chooses t^{diff} instead of $t^{\max}$ because empirical evidence shows that t^{diff} leads to better power. Sorting $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ gives

$$T_j^{(0)} \ge T_j^{(1)} \ge \cdots \ge T_j^{(h)}.$$

Then Clipper defines the contrast score of feature j, $j = 1, \ldots, d$, in one of two ways:

$$C_j := \begin{cases} T_j^{(0)} - T_j^{(1)} & \text{if } T_j^{(0)} = T_j^{\sigma_0} \\ T_j^{(1)} - T_j^{(0)} & \text{otherwise} \end{cases} \qquad \textbf{difference contrast score}\,, \tag{9}$$

or

$$C_j := \begin{cases} \left|T_j^{(0)}\right| & \text{if } T_j^{(0)} = T_j^{\sigma_0} > T_j^{(1)} \\ 0 & \text{if } T_j^{(0)} = T_j^{(1)} \\ -\left|T_j^{(0)}\right| & \text{otherwise} \end{cases} \qquad \textbf{maximum contrast score}\,. \tag{10}$$

The intuition behind the contrast scores is that, if $C_j < 0$, then $T_j^{(0)} \neq T_j^{\sigma_0}$, which means that at least one of $T_j^{\sigma_1}, \ldots, T_j^{\sigma_h}$ (after random permutation) is greater than $T_j^{\sigma_0}$ calculated from the original data (identity permutation), suggesting that feature j is likely an uninteresting feature in enrichment analysis. Fig. 5b (right) shows a cartoon illustration of contrast scores when $m = 2$ and $n = 1$. Motivated by Gimenez and Zou [55], we propose the following procedure for Clipper to control the FDR under the target level $q \in (0, 1)$.

Definition 3 (Gimenez-Zou (GZ) procedure for thresholding contrast scores [55]) *Given $h \in \{1, \cdots, h_{\max}\}$ and contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0\,;\ j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The GZ procedure finds a contrast-score cutoff T^{GZ} based on the target FDR threshold $q \in (0, 1)$ as:*

$$T^{\mathrm{GZ}} := \min\left\{t \in \mathcal{C} : \frac{\frac{1}{h} + \frac{1}{h}\mathrm{card}\left(\{j : C_j \leq -t\}\right)}{\mathrm{card}\left(\{j : C_j \geq t\}\right) \vee 1} \leq q\right\} \tag{11}$$

and outputs $\left\{j : C_j \geq T^{\mathrm{GZ}}\right\}$ as discoveries.

Differential analysis with $m + n > 2$

For differential analysis, Clipper also uses permutation to construct contrast scores. When $m \neq n$, the equivalence classes of permutations are defined the same as for the enrichment analysis with $m \neq n$. When $m = n$, there is a slight change in the definition of equivalence classes of permutations: if σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\} \text{ or } \{\sigma'(m+1), \cdots, \sigma'(2m)\}\,,$$

then we say that σ and σ' are in one equivalence class. In total, there are $h_{\mathrm{total}} := \binom{m+n}{m}$ (when $m \neq n$) or $\binom{2m}{m}/2$ (when $m = n$) equivalence classes of permutations. Hence, to have more than one equivalence class, we cannot perform differential analysis with $m = n = 1$; in other words, the total number of replicates $m + n$ must be at least 3.

Then Clipper randomly samples $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the $h_{\max} := h_{\mathrm{total}} - 1$ equivalence classes that exclude the class containing σ_0, i.e., the identity permutation. Note that $h_{\max}$ is the maximum value h can take. Next, Clipper computes $T_j^{\sigma_\ell} := \left|t^{\mathrm{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})\right|$, where $\boldsymbol{X}_j^{\sigma_\ell}$ and $\boldsymbol{Y}_j^{\sigma_\ell}$ are the permuted data defined in (8), and it defines C_j as the contrast score of feature j, $j = 1, \ldots, d$, in the same ways as in (9) or (10). Fig. 5b (left) shows a cartoon illustration of contrast scores when $m = 2$ and $n = 1$.

Same as in the enrichment analysis with $m \neq n$, Clipper also uses the GZ procedure [55] to set a cutoff on contrast scores to control the FDR under the target level $q \in (0, 1)$.

Granted, when we use permutations to construct contrast scores in the GZ procedure, we can convert contrast scores into permutation-based p-values (see Supp. S1.1.2). However, when the numbers of replicates are small, the number of possible permutations is small, so permutation-based p-values would have a low resolution (e.g., when $m = 2$ and $n = 1$, the number of non-identity permutations is only 2). Hence, applying the BH procedure to the permutation-based p-values would result in almost no power. Although Yekutieli and Benjamini proposed another thresholding procedure for permutation-based p-values [56], it still requires the number of permutations to be large to obtain a reliable FDR control. Furthermore, if we apply the SeqStep+ procedure by Barber and Candés [35] to permutation-based p-values, it would be equivalent to our application of the GZ procedure to contrast scores (Supp. Section S1.1.2).

Generic FDR control methods

In our simulation analysis, we compared Clipper against generic FDR control methods including p-value-based methods and local-fdr-based methods. Briefly, each p-value-based method is a combination of a p-value calculation approach and a p-value thresholding procedure. We use either the "paired" or "pooled" approach (see next paragraph) to calculate p-values of features and then threshold the p-values using the BH procedure (Supp. Definition S1) or Storey's qvalue procedure (Supp. Definition S2) to make discoveries (Supp. Section S1.1). As a result, we have four p-value-based methods: BH-pair, BH-pool, qvalue-pair, and qvalue-pool (Fig. 1b).

Regarding the existing p-value calculation approaches in bioinformatics tools, we categorize them as "paired" or "pooled." The paired approach has been widely used to detect DEGs and protein-binding sites [1, 2, 4, 5]. It examines one feature at a time and compares the feature's measurements between two conditions using a statistical test. In contrast, the pooled approach is popular in proteomics for identifying peptide sequences from MS data [57]. For every feature, it defines a test statistic and estimates a null distribution by pooling all features' observed test statistic values under the background condition. Finally, it calculates a p-value for every feature under the experimental condition based on the feature's observed test statistic and the null distribution.

In parallel to p-value-based methods, local-fdr-based methods estimate local fdrs of features and then threshold the local fdrs using the locfdr procedure (Supp. Definition S5) to make discoveries. The estimation of local fdrs takes one of two approaches: (1) empirical null, which is estimated parametrically from the test statistic values that are likely drawn from the null distribution, and (2) swapping null, which is constructed by swapping measurements between experimental and background conditions. The resulting two local-fdr-based-methods are referred to as locfdr-emp and locfdr-swap (Figs. 1b and 2). Supp. Section S1 provides a detailed explanation of these generic methods and how we implemented them in this work.

Specific to the p-value-based methods, for the paired approach, besides the ideal implementation that uses the correct model to calculate p-values (BH-pair-correct and qvalue-pair-correct), we also consider common mis-implementations. The first mis-implementations is misspecification of the distribution (BH-pair-mis and qvalue-pair-mis). An example is the detection of protein-binding sites from ChIP-seq data. A common assumption is that ChIP-seq read counts in a genomic region (i.e., a feature) follow the Poisson distribution [1, 2], which implies that the counts have the variance equal to the mean. However, if only two replicates are available, it is impossible to check whether this Poisson distribution is reasonably specified. The second mis-implementation is the misspecification of a two-sample test as a one-sample test (BH-pair-2as1 and qvalue-pair-2as1), which ignores the sampling randomness of replicates under one condition. This issue is implicit but widespread in bioinformatics methods [1, 58].

To summarize, we compared Clipper against the following implementations of generic FDR control

methods:

- **BH-pool** or **qvalue-pool**: p-values calculated by the pooled approach and thresholded by the BH or qvalue procedure.
- **BH-pair-correct** or **qvalue-pair-correct**: p-values calculated by the paired approach with the correct model specification and thresholded by the BH or qvalue procedure.
- **BH-pair-mis** or **qvalue-pair-mis**: p-values calculated by the paired approach with a misspecified model and thresholded by the BH or qvalue procedure.
- **BH-pair-2as1** or **qvalue-pair-2as1**: p-values calculated by the paired approach that misformulates a two-sample test as a one-sample test (2as1) and thresholded by the BH or qvalue procedure.
- **locfdr-emp**: local fdrs calculated by the empirical null approach and thresholded by the locfdr procedure.
- **locfdr-swap**: local fdrs calculated by the swapping approach and thresholded by the locfdr procedure.

Real datasets

- The H3K4me3 ChIP-seq dataset with one experimental sample (GEO accession number GSM733708) and two control samples (GEO accession number GSM733742) from the cell line GM12878 is available at `ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/`, with the experimental sample under the filename `wgEncodeBroadHistoneGm12878H3k4me3StdAlnRep1.bam` and the two control samples under the filenames `wgEncodeBroadHistoneGm12878ControlStdAlnRep1.bam` and `wgEncodeBroadHistoneGm12878ControlStdAlnRep2.bam`. The processed dataset is available at `https://zenodo.org/record/4404882`.
- The MS benchmark dataset will be published in a future manuscript. Interested readers should contact Dr. Leo Wang at lewang@coh.org.
- The human monocyte RNA-Seq dataset is available at `https://www.ncbi.nlm.nih.gov/Traces/study/?acc=srp082682`. The dataset includes 17 samples of classical monocytes and 17 samples of nonclassical monocytes, and it is converted to a sample-by-gene count matrix by `R` package `GenomicFeatures` (v 1.40.1). The processed count matrix is available at `https://zenodo.org/record/4404882`.
- The Hi-C dataset from the cell line GM12878 is available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525`. The count matrix is under the filename `GSE63525_GM12878_primary_intrachromosomal_contact_matrices.tar.gz`, and the matrix corresponding to Chromosome 1 and bin width 1MB is used. The processed dataset is available at `https://zenodo.org/record/4404882`.

Software packages used in this study

- **p.adjust** `R` function (in `R` package `stats` v 4.0.2 with default arguments) [14]: used for BH-pool, BH-pair-correct, BH-pair-mis, and BH-pair-2as1.

- **qvalue** R package (v 2.20.0 with default arguments) [59]: used for qvalue-pool, qvalue-pair-correct, qvalue-pair-mis, and qvalue-pair-2as1.
- **locfdr** R package (v 1.1-8 with default arguments) [60]: used for locfdr-emp.
- **MACS2** software package (v 2.2.6 with default settings) [1]: available at https://github.com/macs3-project/MACS/releases/tag/v2.2.6.
- **HOMER** software package (findPeaks v 3.1.9.2 with default settings) [2]: available at https://www.bcgsc.ca/platform/bioinfo/software/findpeaks/releases/3.1.9.2/findpeaks3-1-9-2-tar.gz.
- **Mascot** in Proteome Discoverer (v 2.3.0.523 with default settings) [3]: commercial software by ThermoScientific.
- **edgeR** R package (v 3.30.0 with default arguments) [4]: available at https://www.bioconductor.org/packages/release/bioc/html/edgeR.html.
- **DESeq2** R package (v 1.28.1 with default arguments) [5]: available at http://bioconductor.org/packages/release/bioc/html/DESeq2.html.
- **MultiHiCcompare** R package (v 1.6.0 with default arguments) [11]: available at https://bioconductor.org/packages/release/bioc/html/multiHiCcompare.html.
- **diffHic** R package (v 1.20.0 with default arguments) [13]: available at https://www.bioconductor.org/packages/release/bioc/html/diffHic.html.
- **FIND** R package (v 0.99 with default arguments) [12]: available at https://bitbucket.org/nadhir/find/src/master/.

Software, code, and video tutorial

- The **Clipper** R package is available at https://github.com/JSB-UCLA/Clipper/.
- The code and processed data for reproducing the figures are available at https://zenodo.org/record/4404882.
- A video introduction of Clipper is available at https://youtu.be/-GXyHiJMpLo.

Funding

This work was supported by the following grants: NIH-NCI T32LM012424 (to Y.E.C.); NCI K08 CA201591, the Gabrielle's Angel Foundation and Alex's Lemonade Stand Foundation (to L.D.W.); NIH R01HG007538, R01CA193466, and R01CA228140 (to W.L.); National Science Foundation DBI-1846216, NIH/NIGMS R01GM120507, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

Conflicts of interests

L.D.W. holds equity in Magenta Therapeutics.

Acknowledgements

The authors would like to thank Dr. Yu-Cheng Yang for his suggestions on the figures and R package. The authors would also like to thank Mr. Nikos Ignatiadis, Dr. Lihua Lei, and Dr. Rina Barber for their insightful comments after we presented this work at the International Seminar on Selective Inference (https://www.selectiveinferenceseminar.com/past-talks). The authors also appreciate the comments and feedback from Mr. Tianyi Sun, Ms. Kexin Li, and other members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

References

[1] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9 (2008), pp. 1–9.

[2] Sven Heinz et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Molecular cell* 38.4 (2010), pp. 576–589.

[3] David N Perkins et al. "Probability-based protein identification by searching sequence databases using mass spectrometry data". In: *ELECTROPHORESIS: An International Journal* 20.18 (1999), pp. 3551–3567.

[4] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[5] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[6] Cole Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature biotechnology* 31.1 (2013), pp. 46–53.

[7] Jun Li et al. "Normalization, testing, and false discovery rate estimation for RNA-sequencing data". In: *Biostatistics* 13.3 (2012), pp. 523–538.

[8] Thomas J Hardcastle and Krystyna A Kelly. "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data". In: *BMC bioinformatics* 11.1 (2010), pp. 1–14.

[9] Gordon K Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Statistical applications in genetics and molecular biology* 3.1 (2004).

[10] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[11] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. "multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments". In: *Bioinformatics* 35.17 (2019), pp. 2916–2923.

[12] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. "FIND: difFerential chromatin INteractions Detection using a spatial Poisson process". In: *Genome research* 28.3 (2018), pp. 412–422.

[13] Aaron TL Lun and Gordon K Smyth. "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data". In: *BMC bioinformatics* 16.1 (2015), pp. 1–11.

[14] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[15] Bradley Efron and Robert Tibshirani. “Empirical Bayes methods and false discovery rates for microarrays”. In: *Genetic epidemiology* 23.1 (2002), pp. 70–86.

[16] Bradley Efron et al. “Empirical Bayes analysis of a microarray experiment”. In: *Journal of the American statistical association* 96.456 (2001), pp. 1151–1160.

[17] Matthew Stephens. “False discovery rates: a new deal”. In: *Biostatistics* 18.2 (2017), pp. 275–294.

[18] John D Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[19] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. “Identifying differentially expressed genes using false discovery rate controlling procedures”. In: *Bioinformatics* 19.3 (2003), pp. 368–375.

[20] Bing Yang et al. “Identification of cross-linked peptides from complex samples”. In: *Nature methods* 9.9 (2012), pp. 904–906.

[21] James Robert White, Niranjan Nagarajan, and Mihai Pop. “Statistical methods for detecting differentially abundant features in clinical metagenomic samples”. In: *PLoS Comput Biol* 5.4 (2009), e1000352.

[22] Andrey A Shabalin. “Matrix eQTL: ultra fast eQTL analysis via large matrix operations”. In: *Bioinformatics* 28.10 (2012), pp. 1353–1358.

[23] Stijn Hawinkel et al. “A broken promise: microbiome differential abundance methods do not control the false discovery rate”. In: *Briefings in bioinformatics* 20.1 (2019), pp. 210–221.

[24] Ye Zheng and Sündüz Keleş. “FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation”. In: *Nature Methods* 17.1 (2020), pp. 37–40.

[25] John D Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[26] Yoav Benjamini and Yosef Hochberg. “Multiple hypotheses testing with weights”. In: *Scandinavian Journal of Statistics* 24.3 (1997), pp. 407–418.

[27] Nikolaos Ignatiadis et al. “Data-driven hypothesis weighting increases detection power in genome-scale multiple testing”. In: *Nature methods* 13.7 (2016), pp. 577–580.

[28] Lihua Lei and William Fithian. “Adapt: an interactive procedure for multiple testing with side information”. In: *arXiv preprint arXiv:1609.06035* (2016).

[29] Simina M Boca and Jeffrey T Leek. “A direct approach to estimating false discovery rates conditional on covariates”. In: *PeerJ* 6 (2018), e6035.

[30] Joses Ho et al. “Moving beyond P values: data analysis with estimation graphics”. In: *Nature methods* 16.7 (2019), pp. 565–566.

[31] Dongyuan Song and Jingyi Jessica Li. “PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data”. In: *bioRxiv* (2020).

[32] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. “Significance analysis of microarrays applied to the ionizing radiation response”. In: *Proceedings of the National Academy of Sciences* 98.9 (2001), pp. 5116–5121.

[33] Jesse Hemerik and Jelle J Goeman. “False discovery proportion estimation by permutations: confidence for significance analysis of microarrays”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1 (2018), pp. 137–155.

[34] Jesse Hemerik, Aldo Solari, and Jelle J Goeman. “Permutation-based simultaneous confidence bounds for the false discovery proportion”. In: *Biometrika* 106.3 (2019), pp. 635–649.

[35] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[36] Ery Arias-Castro and Shiyun Chen. “Distribution-free multiple testing”. In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1983–2001.

[37] Yoav Benjamini. “Selective inference: The silent killer of replicability”. In: *Issue 2.4* 2.4 (2020).

[38] Kristen Emery et al. “Multiple Competition-Based FDR Control and Its Application to Peptide Detection”. In: *International Conference on Research in Computational Molecular Biology*. Springer. 2020, pp. 54–71.

[39] Abhishek K Sarkar and Matthew Stephens. “Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis”. In: *BioRxiv* (2020).

[40] Peter J Park. “ChIP–seq: advantages and challenges of a maturing technology”. In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[41] Pamela J Mitchell and Robert Tjian. “Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins”. In: *Science* 245.4916 (1989), pp. 371–378.

[42] Mark Ptashne and Alexander Gann. “Transcriptional activation by recruitment”. In: *Nature* 386.6625 (1997), pp. 569–577.

[43] Timothy Bailey et al. “Practical guidelines for the comprehensive analysis of ChIP-seq data”. In: *PLoS Comput Biol* 9.11 (2013), e1003326.

[44] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[45] Aaron Diaz et al. “Normalization, bias correction, and peak calling for ChIP-seq”. In: *Statistical applications in genetics and molecular biology* 11.3 (2012).

[46] Boris Bogdanow, Henrik Zauber, and Matthias Selbach. “Systematic errors in peptide and protein identification and quantification by modified peptides”. In: *Molecular & Cellular Proteomics* 15.8 (2016), pp. 2791–2801.

[47] Michael P Washburn, Dirk Wolters, and John R Yates. “Large-scale analysis of the yeast proteome by multidimensional protein identification technology”. In: *Nature biotechnology* 19.3 (2001), pp. 242–247.

[48] Marshall Bern, Yong J Kil, and Christopher Becker. “Byonic: advanced peptide and protein identification software”. In: *Current protocols in bioinformatics* 40.1 (2012), pp. 13–20.

[49] Claire R Williams et al. “Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq”. In: *BMC bioinformatics* 18.1 (2017), p. 38.

[50] Keegan Korthauer et al. “A practical guide to methods controlling false discoveries in computational biology”. In: *Genome biology* 20.1 (2019), pp. 1–21.

[51] Guangchuang Yu et al. “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287.

[52] Jonathan Thorsen et al. “Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies”. In: *Microbiome* 4.1 (2016), p. 62.

[53] Charlotte Soneson and Mark D Robinson. “Bias, robustness and scalability in single-cell differential expression analysis”. In: *Nature methods* 15.4 (2018), p. 255.

[54] Kun He et al. “Null-free False Discovery Rate Control Using Decoy Permutations for Multiple Testing”. In: *arXiv preprint arXiv:1804.08222* (2018).

[55] Jaime Roquero Gimenez and James Zou. “Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization”. In: *arXiv preprint arXiv:1810.11378* (2018).

[56] Daniel Yekutieli and Yoav Benjamini. “Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics”. In: *Journal of Statistical Planning and Inference* 82.1-2 (1999), pp. 171–196.

[57] Alexey I Nesvizhskii. “A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics”. In: *Journal of proteomics* 73.11 (2010), pp. 2092–2123.

[58] Wei Li et al. “MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens”. In: *Genome biology* 15.12 (2014), p. 554.

[59] John D. Storey et al. *qvalue: Q-value estimation for false discovery rate control*. R package version 2.20.0. 2020. URL: http://github.com/jdstorey/qvalue.

[60] Bradley Efron. “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis”. In: *Journal of the American Statistical Association* 99.465 (2004), pp. 96–104.

Figures

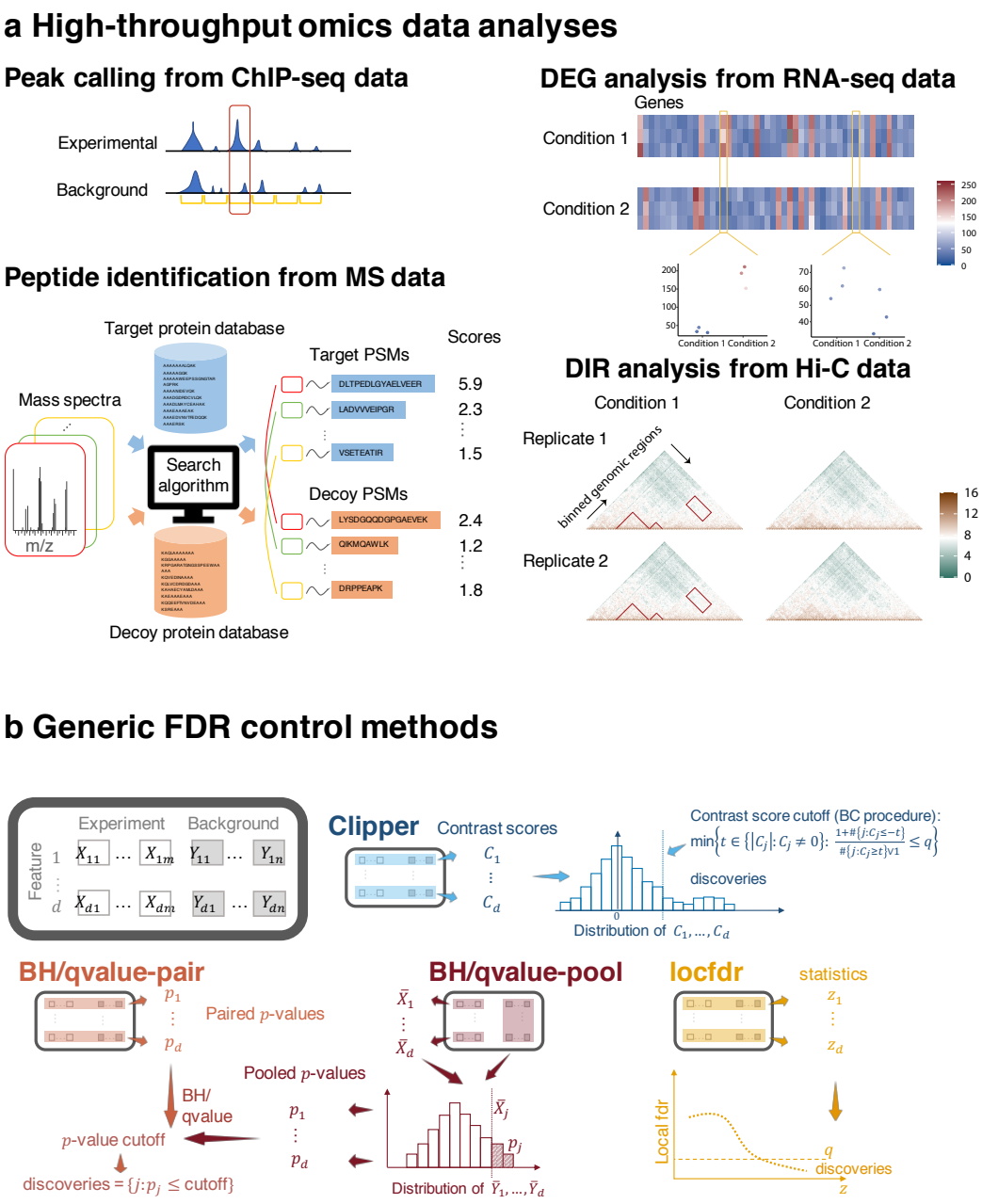


Figure 1: High-throughput omics data analyses and generic FDR control methods. (a) Illustration of four common high-throughput omics data analyses: peak calling from ChIP-seq data, peptide identification from MS data, DEG analysis from RNA-seq data, and DIR analysis from Hi-C data. In these four analyses, the corresponding features are genomic regions (yellow intervals), peptide-spectrum matches (PSMs; a pair of a mass spectrum and a peptide sequence), genes (columns in the heatmaps), and chromatin interacting regions (entries in the heatmaps). **(b)** Illustration of Clipper and five generic FDR control methods: BH-pair (and qvalue-pair), BH-pool (and qvalue-pool), and locfdr. The input data are d features with m and n repeated measurements under the experimental and background conditions, respectively. Clipper computes a contrast score for each feature based on the feature's m and n measurements, decides a contrast-score cutoff, and calls the features with contrast scores above the cutoff as discoveries. (This illustration is Clipper for enrichment analysis with $m = n$.) BH-pair or qvalue-pair computes a p-value for each feature based on the feature's m and n measurements, sets a p-value cutoff, and calls the features with p-values below the cutoff as discoveries. BH-pool or qvalue-pool constructs a null distribution from the d features' average (across the n replicates) measurements under the background condition, calculates a p-value for each feature based on the null distribution and the feature's average (across the m replicates) measurements under the experimental condition, sets a p-value cutoff, and calls the features with p-values below the cutoff as discoveries. The locfdr method computes a summary statistic for each feature based on the feature's m and n measurements, estimates the empirical null distribution and the empirical distribution of the statistic across features, computes a local fdr for each feature, sets a local fdr cutoff, and calls the features with local fdr below the cutoff as discoveries.

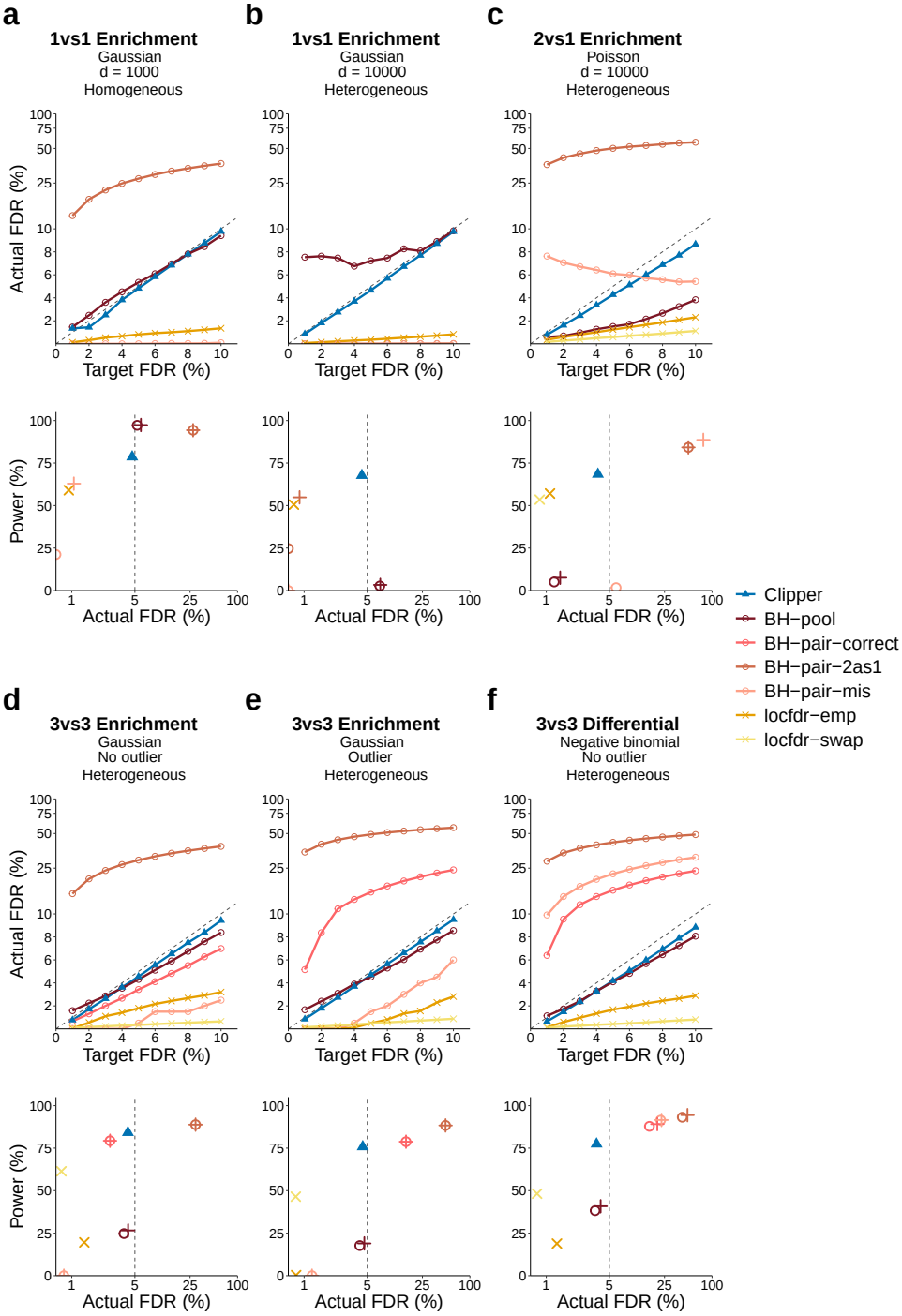


Figure 2: Comparison of Clipper with generic FDR control methods in terms of their FDR control and power in six example simulation studies. (a) 1vs1 enrichment analysis with 1000 features generated from the Gaussian distribution with a homogeneous background; **(b)** 1vs1 enrichment analysis with $10{,}000$ features generated from the Gaussian distribution with a heterogeneous background; **(c)** 2vs1 enrichment analysis with $10{,}000$ features generated from the Poisson distribution with a heterogeneous background; **(d)** 3vs3 enrichment analysis with $10{,}000$ features generated from the Gaussian distribution without outliers and with a heterogeneous background; **(e)** 3vs3 enrichment analysis with $10{,}000$ features generated from the Gaussian distribution without outliers and with a heterogeneous background; **(f)** 3vs3 differential analysis with $10{,}000$ features generated from the negative binomial distribution with a heterogeneous background. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are approximated by the averages of false discovery proportions (see Eq. (S14) in the Supplementary) and power evaluated on 200 simulated datasets. In each panel, the top row shows each method's actual FDRs at target FDR thresholds: whenever the actual FDR is larger than the target FDR (the solid line is higher than the dashed line), FDR control is failed; the bottom row shows each method's actual FDRs and power at the target FDR threshold $q = 5\%$: whenever the actual FDR is greater than q (on the right of the vertical dashed line), FDR control is failed. Under the FDR control, the larger the power, the better. Note that BH-pair-correct is not included in (a)–(c) because it is impossible to correctly specify the model with only one replicate per condition; locfdr-swap is not included in (a)–(b) because it is inapplicable to the 1vs1 design.

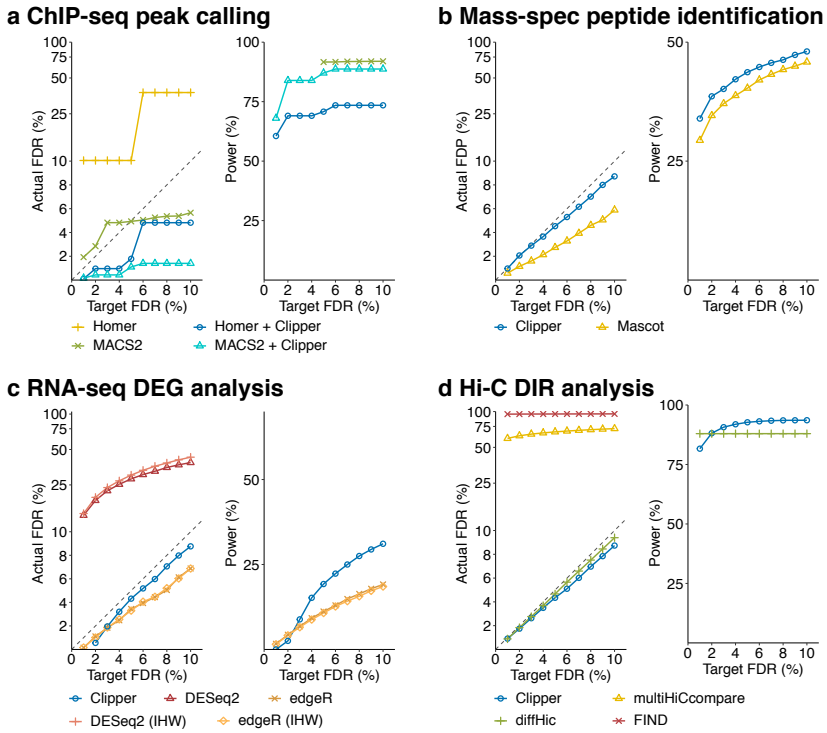
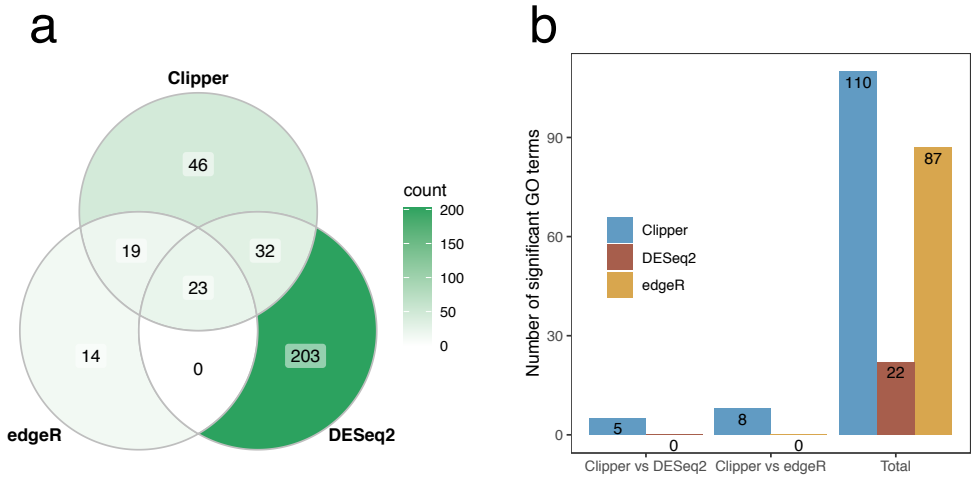


Figure 3: Comparison of Clipper and popular bioinformatics methods in terms of FDR control and power. (a) peaking calling analysis on synthetic ChIP-seq data; **(b)** peptide identification on real proteomics data; **(c)** DEG analysis on synthetic bulk RNA-seq data; **(d)** DIR analysis on synthetic Hi-C data. In all four panels, the target FDR level q ranges from 1% to 10%. Points above the dashed line indicate failed FDR control; when this happens, the power of the corresponding methods is not shown, including HOMER in (a), MACS2 for target FDR less than 5% in (a), DESeq2 and DESeq2 (IHW) in (c), and multiHICcompare and FIND in (d). In all four applications, Clipper controls the FDR while maintaining high power, demonstrating Clipper's broad applicability in high-throughput data analyses.



c

Top GO terms enriched in all three sets of identified DEGs

GO term (ID)	qvalue (DESeq2)	qvalue (edgeR)	qvalue (Clipper)
leukocyte chemotaxis (GO:0030595)	9.930044e−06	9.594885e−09	3.104557e−10
myeloid leukocyte migration (GO:0097529)	1.107612e−05	2.921486e−08	5.740217e−10
granulocyte chemotaxis (GO:0071621)	2.698853e−05	1.008808e−08	1.167108e−09
neutrophil chemotaxis (GO:0030593)	2.698853e−05	2.921486e−08	2.691033e−09

Figure 4: Application of Clipper, DESeq2, and edgeR to identifying DEGs from the classical and non-classical human monocyte dataset. (a) A Venn diagram showing the overlaps of the identified DEGs (at the FDR threshold $q = 5\%$) by the three DE methods. **(b)** Numbers of GO terms enriched (with enrichment q-values < 0.01) in the DEGs found by Clipper, DESeq2 and edgeR (column 3), or in the DEGs specifically identified by Clipper or DESeq2/edgeR in the pairwise comparison between Clipper and DESeq2 (column 1) or between Clipper and edgeR (column 2). More GO terms are enriched in the DEGs identified by Clipper than in those identified by edgeR or DESeq2. **(c)** Enrichment q-values of four GO terms that are found enriched (with enrichment q-values < 0.01) in all three sets of identified DEGs, one set per method. All the four terms are most enriched in the DEGs identified by Clipper.

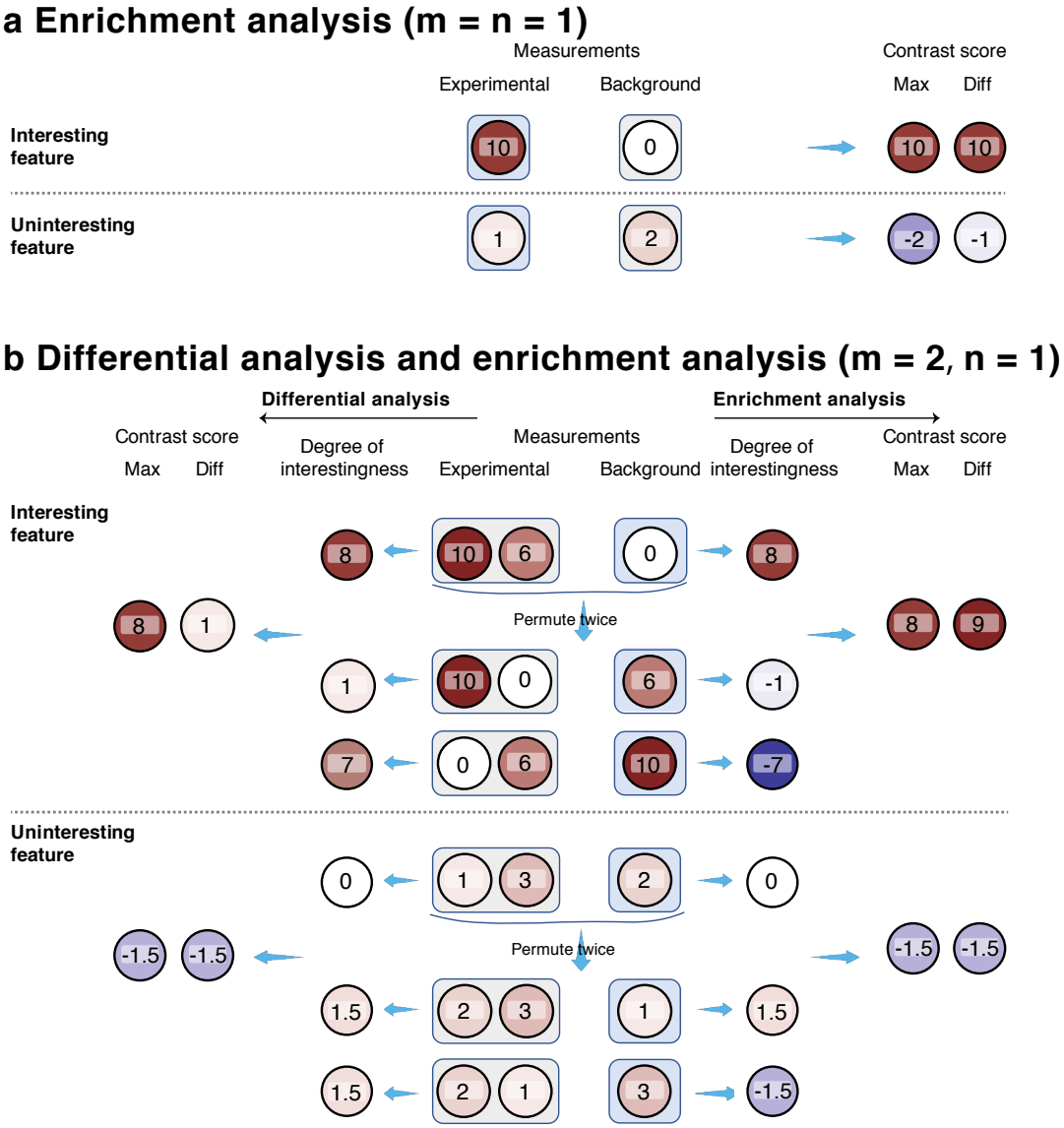


Figure 5: Illustration of the construction of contrast scores. (a) 1vs1 enrichment analysis; **(b)** 2vs1 differential analysis (left) or enrichment analysis (right). In each panel, an interesting feature (top) and an uninteresting feature (bottom) are plotted for contrast; both features have measurements under the experimental and background conditions. In (a), each feature's measurements are summarized into a maximum (max) contrast score or a difference (diff) contrast score. In (b), each feature's measurements are permuted across the two conditions, resulting in two sets of permuted measurements. Then for each feature, we calculate its degrees of interestingness (as the difference that equals the average of experimental measurements minus the average of background measurements (in enrichment analysis; right), or the absolute value of the difference (in differential analysis; left)) from its original measurements and permuted measurements, respectively. Finally, we summarize each feature's degrees of interestingness into a maximum (max) contrast score or a difference (diff) contrast score.

Supplementary Material

S1 Review of generic FDR control methods

To facilitate our discussion, we introduce the notations for data. For feature $j = 1, \ldots, d$, we use $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top \in \mathbb{R}^m_{\geq 0}$ and $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top \in \mathbb{R}^n_{\geq 0}$ to denote its measurements under the experimental and background conditions, respectively. We assume that $X_{j1}, \ldots, X_{jm}$ are identically distributed, so are $Y_{j1}, \ldots, Y_{jn}$. Let $\mu_{Xj} = \mathbb{E}[X_{j1}]$ and $\mu_{Yj} = \mathbb{E}[Y_{j1}]$ denote the expected measurement of feature j under the two conditions, respectively. Then we denote by $\bar{X}_j$ the sample average of $X_{j1}, \cdots, X_{jm}$ and by $\bar{Y}_j$ the sample average of $Y_{j1}, \cdots, Y_{jn}$.

S1.1 P-value-based methods

Here we describe the details of p-value-based FDR control methods, including BH-pair, BH-pool, qvalue-pair, and qvalue-pool. Each of these four methods first computes p-values using either the pooled approach or the paired approach, and it then relies on the BH procedure [1] or Storey's qvalue procedure [2] for FDR control. In short, every p-value-based method is a combination of a p-value calculation approach and a p-value thresholding procedure. Below we introduce two p-value calculation approaches (paired and pooled) and two p-value thresholding procedures (BH and Storey's qvalue).

S1.1.1 P-value calculation approaches

The paired approach. The paired approach examines one feature at a time and compares its measurements between two conditions. Besides the ideal implementation, i.e., the *correct paired approach* that uses the correct model to calculate p-values, we also include commonly-used flawed implementations that either misspecify the distribution, i.e., the *misspecified paired approach*, or misformulate the two-sample test as a one-sample test, i.e., the *2as1 paired approach*.

Here we use the negative binomial distribution as an example to demonstrate the ideas of the correct, misspecified, and 2as1 paried approaches. Suppose that for each feature j, its measurements under each condition follow a negative binomial distribution, and the two distributions under the two conditions have the same dispersion; that is, $X_{j1}, \cdots, X_{jm} \overset{\text{i.i.d.}}{\sim} \mathrm{NB}\left(\mu_{Xj}, \theta_j\right)$; $Y_{j1}, \cdots, Y_{jn} \overset{\text{i.i.d.}}{\sim} \mathrm{NB}\left(\mu_{Yj}, \theta_j\right)$, where θ_j is the dispersion parameter such that the variance $\mathrm{Var}(X_{ji}) = \mu_{Xj} + \theta_j \mu_{Xj}^2$.

- The correct paired approach assumes that the two negative binomial distributions have the same dispersion parameter θ_j, and it uses the two-sample test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \mu_{Yj}$ (differential analysis).

- The misspecified paired approach misspecifies the negative binomial distribution as Poisson, and it uses the two-sample test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \mu_{Yj}$ (differential analysis).

- The 2as1 paired approach bluntly assumes $\mu_{Yj} = \bar{Y}_j$, and it performs the one-sample test based on $X_{j1}, \ldots, X_{jm}$ for the null hypotheses $H_0 : \mu_{Xj} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{Xj} > \bar{Y}_j$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \bar{Y}_j$ (differential analysis).

The pooled approach. The pooled approach pools all features' average measurements under the background condition $\left\{\bar{Y}_j\right\}_{j=1}^d$ to form a null distribution, and it calculates a p-value for each feature j

by comparing $\bar{X}_j$ to the null distribution. Specifically, in enrichment analysis, the p-value of feature j is computed as:

$$p_j = \frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \geq \bar{X}_j\right\}\right)}{d}.$$

In differential analysis, the p-value of feature j is computed as:

$$p_j = 2 \cdot \min\left(\frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \geq \bar{X}_j\right\}\right)}{d}, \frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \leq \bar{X}_j\right\}\right)}{d}\right).$$

S1.1.2 P-value thresholding procedures for FDR control

Definition S1 (BH procedure for thresholding p-values [1]) *The features' p-values $p_1, \ldots, p_d$ are sorted in an ascending order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$. Given the target FDR threshold q, the Benjamini–Hochberg (BH) procedure finds a p-value cutoff T^{BH} as*

$$T^{\mathrm{BH}} := p_{(k)}, \text{ where } k = \max\left\{j = 1, \ldots, d : p_{(j)} \leq \frac{j}{d} q\right\}. \tag{S1}$$

Then BH outputs $\left\{j : p_j \leq T^{\mathrm{BH}}\right\}$ as discoveries.

Definition S2 (Storey's qvalue procedure for thresholding p-values [2]) *The features' p-values $p_1, \ldots, p_d$ are sorted in an ascending order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$. Let $\hat{\pi}_0$ denote an estimate of the probability $P(\text{the } i\text{-th feature is uninteresting})$ (see Storey [2] for details). Storey's qvalue procedure defines the q-value for $p_{(d)}$ as*

$$\hat{q}(p_{(d)}) := \frac{\hat{\pi}_0 \cdot d \cdot p_{(d)}}{\operatorname{card}\left(\left\{k : p_k \leq p_{(d)}\right\}\right)} = \hat{\pi}_0 \cdot p_{(d)}.$$

Then for $j = d-1, d-2, \ldots, 1$, the q-value for $p_{(j)}$ is defined as:

$$\hat{q}(p_{(j)}) := \min\left(\hat{q}(p_{(j+1)}), \frac{\hat{\pi}_0 \cdot d \cdot p_{(j)}}{\operatorname{card}\left(\left\{k : p_k \leq p_{(j)}\right\}\right)}\right).$$

Then Storey's qvalue procedure outputs $\{j : \hat{q}(p_j) \leq q\}$ as discoveries.

We use function `qvalue` from `R` package `qvalue` (v 2.20.0; with default estimate $\hat{\pi}_0$) to calculate q-values.

Definition S3 (SeqStep+ procedure for thresholding p-values [3]) *Define H_0^j as the null hypothesis for feature j and p_j as the p-value for H_0^j, $j = 1, \ldots, d$. Order the null hypotheses $H_0^1, \ldots, H_0^d$ from the most to the least promising (here more promising means more likely to be interesting) and denote the resulting null hypotheses and p-values as $H_0^{(1)}, \ldots, H_0^{(d)}$ and $p_{(1)}, \ldots, p_{(d)}$. Given any target FDR threshold q, a pre-specified constant $s \in (0, 1)$, and subset $\mathcal{K} \subseteq \{1, \ldots, d\}$, the SeqStep+ procedure finds a cutoff $\hat{j}$ as*

$$\hat{j} := \max\left\{j \in \mathcal{K} : \frac{1 + \operatorname{card}\left(\left\{k \in \mathcal{K}, k \leq j : p_{(k)} > s\right\}\right)}{\operatorname{card}\left(\left\{k \in \mathcal{K}, k \leq j : p_{(k)} \leq s\right\}\right) \vee 1} \leq \frac{1-s}{s} q\right\} \tag{S2}$$

Then SeqStep+ rejects $\left\{H_0^{(j)} : p_{(j)} \leq s,\ j \leq \hat{j},\ j \in \mathcal{K}\right\}$. If the orders of the null hypotheses are independent of the p-values, the SeqStep+ procedure ensures FDR control.

The GZ procedure (Definition 3) used in Clipper is a special case of the SeqStep+ procedure with $s = 1/(h+1)$. Recall that given the number of non-identical permutations $h \in \{1, \cdots, h_{\max}\}$ and contrast

scores $\{C_j\}_{j=1}^d$, the GZ procedure sorts $\{|C_j|\}_{j=1}^d$ in a decreasing order:

$$|C_{(1)}| \geq |C_{(2)}| \geq \cdots \geq |C_{(d)}| \,. \tag{S3}$$

To see the connection between the GZ procedure and SeqStep+, we consider the null hypothesis for the j-th ordered feature, $j = 1, \ldots, d$, as $H_0^{(j)} : \mu_{X(j)} = \mu_{Y(j)}$ and define the corresponding p-value $p_{(j)} := \frac{r(T_{(j)}^{\sigma_0})}{h+1}$, where $r(T_{(j)}^{\sigma_0})$ is the rank of $T_{(j)}^{\sigma_0}$ in $\{T_{(j)}^{\sigma_0}, \cdots, T_{(j)}^{\sigma_h}\}$ in a descending order. We also define $\mathcal{K} := \{j = 1, \ldots, d : C_j \neq 0\}$ as the subset of features with non-zero C_j's. Finally, we input the p-values, null hypothesis orders in (S3), $s = 1/(h+1)$, q and $\mathcal{K}$ into the SeqStep+ procedure, and we obtain the GZ procedure.

The BC procedure (Definition 1) is a further special case with $h = 1$, $p_{(j)} := \left(\mathbb{1}(C_{(j)} > 0) + 1\right)/2$, and $\mathcal{K} := \{j = 1, \ldots, d : C_j \neq 0\}$.

S1.2 Local-fdr-based methods

The FDR is statistical criterion that ensures the reliability of discoveries as a whole. In contrast, the local fdr focuses on the reliability of each discovery. The definition of the local fdr relies on some pre-computed summary statistics z_j for feature j, $j = 1, \ldots, d$. In the calculation of local fdr, $\{z_1, \ldots, z_d\}$ are assumed to be realizations of an abstract random variable Z that represents any feature. Let p_0 or p_1 denote the prior probability that any feature is uninteresting or interesting, with $p_0 + p_1 = 1$. Let $f_0(z) := \mathbb{P}(Z = z \,|\, \text{uninteresting})$ or $f_1(z) := \mathbb{P}(Z = z \,|\, \text{interesting})$ denote the conditional probability density of Z at z given that Z represents an uninteresting or interesting feature. Thus by Bayes' theorem, the posterior probability of any feature being uninteresting given its summary statistic $Z = z$ is

$$\mathbb{P}(\text{uninteresting} \mid Z = z) = p_0 f_0(z)/f(z) \,, \tag{S4}$$

where $f(z) := p_0 f_0(z) + p_1 f_1(z)$ is the marginal probability density of Z. Accordingly, the local fdr of feature j is defined as follows.

Definition S4 (Local fdr [4]) *Given notations defined above, the local fdr of feature j is defined as*

$$\textit{local-fdr}_j := f_0(z_j)/f(z_j) \,.$$

Because $p_0 \leq 1$, local-fdr$_j$ is an upper bound of the posterior probability of feature j being uninteresting given its summary statistic z_j, defined in (S4).

Note that another definition of the local fdr is the posterior probability $\mathbb{P}(\text{uninteresting} \mid z)$ in (S4) [5]. Although this other definition is more reasonable, we do not use it but choose Definition S4 because the estimation of p_0 is ususally difficult. Another reason is that uninteresting features are the dominant majority in high-throughput biological data, so p_0 is often close to 1.

We define local-fdr-based methods as a type of FDR control methods by thresholding local fdrs of features under the target FDR threshold q. Although the local fdr is different from FDR, it has been shown that thresholding the local fdrs at q will approximately control the FDR under q [4]. This makes local-fdr-based methods competitors against Clipper and p-value-based methods.

Every local-fdr-based method is a combination of a local fdr calculation approach and a local fdr thresholding procedure. Below we introduce two local fdr calculation approaches (empirical null and swapping) and one local fdr thresholding procedure. After the combination, we have two local-fdr-based methods: locfdr-emp and locfdr-swap.

S1.2.1 Local fdr calculation approaches

With $z_1, \ldots, z_d$, the calculation of local fdr defined in Definition S4 requires the estimation of f_0 and f, two probability densities. f is estimated by nonparametric density estimation, and f_0 is estimated by either the empirical null approach [4] or the swapping approach, which shuffles replicates between conditions [5]. With the estimated $\hat{f}$ and $\hat{f}_0$, the estimated local fdr of feature j is

$$\widehat{\text{local-fdr}}_j := \hat{f}_0(z_j)/\hat{f}(z_j)\,. \tag{S5}$$

The empirical null approach. This approach assumes a parametric distribution, typically the Gaussian distribution, to estimate f_0. Then with the density estimate $\hat{f}$, the local fdr is estimated for each feature j. The implementation of this approach depends on the numbers of replicates.

- In 1vs1 enrichment and differential analyses, we define z_j as

$$z_j := \frac{D_j}{\sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(D_j - \bar{D}\right)^2}}\,,$$

  where $D_j = X_{j1} - Y_{j1}$ and $\bar{D} = \sum_{j=1}^{d} D_j/d$.

- In 2vs1 enrichment and differential analyses, we define z_j as

$$z_j := \frac{\bar{X}_j - Y_{j1}}{\sqrt{\frac{s_{Xj}^2}{2}}}\,,$$

  where $s_{Xj}^2 = \sum_{i=1}^{2}(X_{ji} - \bar{X}_j)^2$.

- In $m$vs$n$ enrichment and differential analyses with $m, n \geq 2$, we define z_j as the two-sample t-statistic with unequal variances:

$$z_j := \frac{\bar{X}_j - \bar{Y}_j}{\sqrt{\frac{s_{Xj}^2}{m} + \frac{s_{Yj}^2}{n}}}\,,$$

  where $s_{Xj}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{ji} - \bar{X}_j)^2$ and $s_{Yj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ji} - \bar{Y}_j)^2$ are the sample variances of feature j under the experimental and background conditions.

Then $\{\widehat{\text{locfdr}}_j\}_{j=1}^{d}$ are estimated from $\{z_j\}_{j=1}^{d}$ by function `locfdr` in `R` package `locfdr` (v 1.1-8; with default arguments).

The swapping approach. This approach swaps $\lceil m/2 \rceil$ replicates under the experimental condition with $\lceil n/2 \rceil$ replicates under the background condition. Then it calculates the summary statistic for each feature on the swapped data, obtaining $z'_1, \ldots, z'_d$. Finally, it estimates f_0 and f by applying kernel density estimation to $z'_1, \ldots, z'_d$ and $z_1, \ldots, z_d$, respectively (by function `kde` in `R` package `ks`). With $\hat{f}_0$ and $\hat{f}$, $\{\widehat{\text{locfdr}}_j\}_{j=1}^{d}$ are calculated by Definition S4.

The implementation of this approach depends on the numbers of replicates. Below are three special cases included in this work.

- In 1vs1 enrichment and differential analyses, the swapping approach is inapplicable because interesting features would not become uninteresting after the swapping.

- In 2vs1 enrichment and differential analyses, we define z_j and z'_j as

$$z_j = \frac{X_{j1} + X_{j2}}{2} - Y_{j1}\,,$$
$$z'_j = \frac{X_{j1} + Y_{j1}}{2} - X_{j2}\,.$$

- In 3vs3 enrichment and differential analyses with, we define z_j and z'_j as

$$z_j = \frac{X_{j1} + X_{j2}}{2} - \frac{Y_{j1} + Y_{j2}}{2}\,,$$
$$z'_j = \frac{X_{j1} + Y_{j1}}{2} - \frac{X_{j2} + Y_{j2}}{2}\,.$$

Then we apply kernel density estimation to $\{z_j\}_{j=1}^d$ and $\{z'_j\}_{j=1}^d$ to obtain $\hat{f}$ and $\hat{f}_0$, respectively. By (S5), we calculate $\{\widehat{\text{locfdr}}_j\}_{j=1}^d$.

S1.2.2 The local fdr thresholding procedure

Definition S5 (locfdr procedure) *Given the local fdr estimates* $\{\widehat{\textit{local-fdr}}_j\}_{j=1}^d$ *and the target FDR threshold* q*, the locfdr procedure outputs* $\{j = 1, \ldots, d : \widehat{\textit{local-fdr}}_j \leq q\}$ *as discoveries.*

S2 The Clipper methodology

Clipper is a flexible framework that reliably controls the FDR without using p-values in high-throughput data analysis with two conditions. Clipper has two functionalities: (I) enrichment analysis, which identifies the "interesting" features that have higher expected measurements (i.e., true signals) under the experimental condition than the background, a.k.a. negative control condition (if the goal is to identify the interesting features with smaller expected measurements under the experimental condition, enrichment analysis can be applied after the values are negated); (II) differential analysis, which identifies the interesting features that have different expected measurements between the two conditions. For both functionalities, uninteresting features are defined as those that have equal expected measurements under the two conditions.

Clipper only relies on two fundamental statistical assumptions of biological data analysis: (1) measurement errors (i.e., differences between measurements and their expectations, with the expectations including biological signals and batch effects) are independent across all features and experiments; (2) every uninteresting feature has measurement errors identically distributed across all experiments. These two assumptions are used in almost all bioinformatics tools and commonly referred to as the "measurement model" in statistical genomics [6].

In the following subsections, we will first introduce notations and assumptions used in Clipper. Then we will detail how Clipper works and discuss its theoretical guarantee in three analysis tasks: the enrichment analysis with equal numbers of replicates under two conditions ($m = n$), the enrichment analysis with different numbers of replicates under two conditions ($m \neq n$), and the differential analysis (when $m + n > 2$)..

S2.1 Notations and assumptions

To facilitate our discussion, we first introduce the following mathematical notations. For two random vectors $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, or two sets of random variables $\mathcal{X} = \{X_1, \ldots, X_m\}$

and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$, we write $\boldsymbol{X} \perp \boldsymbol{Y}$ or $\mathcal{X} \perp \mathcal{Y}$ if X_i is independent of Y_j for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. To avoid confusion, we use $\mathrm{card}(A)$ to denote the cardinality of a set A and $|c|$ to denote the absolute value of a scalar c. We define $a \vee b := \max(a, b)$.

Clipper only requires two inputs: the target FDR threshold $q \in (0, 1)$ and the input data. Regarding the input data, we use d to denote the number of features with measurements under two conditions, and we use m and n to denote the numbers of replicates under the two conditions. For each feature $j = 1, \ldots, d$, we use $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top \in \mathbb{R}^m_{\geq 0}$ and $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top \in \mathbb{R}^n_{\geq 0}$ to denote its measurements under the two conditions, where $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers. We assume that all measurements are non-negative, as in the case of most high-throughput experiments. (If this assumption does not hold, transformations can be applied to make data satisfy this assumption.)

Clipper has the following assumptions on the joint distribution of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$. For $j = 1, \ldots, d$, Clipper assumes that $X_{j1}, \ldots, X_{jm}$ are identically distributed, so are $Y_{j1}, \ldots, Y_{jn}$. Let $\mu_{Xj} = \mathbb{E}[X_{j1}]$ and $\mu_{Yj} = \mathbb{E}[Y_{j1}]$ denote the expected measurement of feature j under the two conditions, respectively. Then conditioning on $\{\mu_{Xj}\}_{j=1}^d$ and $\{\mu_{Yj}\}_{j=1}^d$,

$$\begin{aligned} &X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are mutually independent}\,; \\ &\boldsymbol{X}_j \perp \boldsymbol{X}_k, \boldsymbol{Y}_j \perp \boldsymbol{Y}_k \text{ and } \boldsymbol{X}_j \perp \boldsymbol{Y}_k\,,\ \forall j, k = 1, \ldots, d. \end{aligned} \tag{S6}$$

An enrichment analysis aims to identify interesting features with $\mu_{Xj} > \mu_{Yj}$ (with $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ defined as the measurements under the experimental and background conditions, respectively), while a differential analysis aims to call interesting features with $\mu_{Xj} \neq \mu_{Yj}$. We define $\mathcal{N} := \{j : \mu_{Xj} = \mu_{Yj}\}$ as the set of uninteresting features and denote $N := \mathrm{card}(\mathcal{N})$. In both analyses, Clipper further assumes that an uninteresting feature j satisfies

$$X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are identically distributed}\,, \forall j \in \mathcal{N}\,. \tag{S7}$$

Clipper consists of two main steps: construction and thresholding of contrast scores. First, Clipper computes contrast scores, one per feature, as summary statistics that reflect the extent to which features are interesting. Second, Clipper establishes a contrast-score cutoff and calls as discoveries the features whose contrast scores exceed the cutoff.

To construct contrast scores, Clipper uses two summary statistics $t(\cdot, \cdot) : \mathbb{R}^m_{\geq 0} \times \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ to extract data information regarding whether a feature is interesting or not:

$$t^{\mathrm{diff}}(\boldsymbol{x}, \boldsymbol{y}) := \bar{x} - \bar{y}\,; \tag{S8}$$

$$t^{\max}(\boldsymbol{x}, \boldsymbol{y}) := \max(\bar{x}, \bar{y}) \cdot \mathrm{sign}(\bar{x} - \bar{y})\,, \tag{S9}$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^\top \in \mathbb{R}^m_{\geq 0}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n_{\geq 0}$, $\bar{x} = \sum_{i=1}^m x_i/m$, $\bar{y} = \sum_{i=1}^n y_i/n$, and $\mathrm{sign}(\cdot) : \mathbb{R} \to \{-1, 0, 1\}$ with $\mathrm{sign}(x) = 1$ if $x > 0$, $\mathrm{sign}(x) = -1$ if $x < 0$, and $\mathrm{sign}(x) = 0$ otherwise.

S2.2 Enrichment analysis with equal numbers of replicates ($m = n$)

Under the enrichment analysis, we assume that $\boldsymbol{X}_j \in \mathbb{R}^m_{\geq 0}$ and $\boldsymbol{Y}_j \in \mathbb{R}^n_{\geq 0}$ are the measurements of feature j, $j = 1, \ldots, d$, under the experimental and background conditions with m and n replicates, respectively. We start with the simple case when $m = n$. Clipper defines a contrast score C_j of feature

j in one of two ways:

$$C_j := t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{difference contrast score}\,, \tag{S10}$$

or

$$C_j := t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{maximum contrast score}\,. \tag{S11}$$

Accordingly, a large positive value of C_j bears evidence that $\mu_{Xj} > \mu_{Yj}$. Motivated by Barber and Candès [3] and Arias-Castro and Chen [7], Clipper proposes the following BC procedure to control the FDR under the target level $q \in (0, 1)$.

Definition S6 (Barber-Candès (BC) procedure for thresholding contrast scores [3]) *Given contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0\,;\ j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The BC procedure finds a contrast-score cutoff T^{BC} based on the target FDR threshold $q \in (0, 1)$ as*

$$T^{\text{BC}} := \min\left\{t \in \mathcal{C} : \frac{\text{card}(\{j : C_j \leq -t\}) + 1}{\text{card}(\{j : C_j \geq t\}) \vee 1} \leq q\right\} \tag{S12}$$

and outputs $\{j : C_j \geq T^{\text{BC}}\}$ as discoveries.

Theorem 1 *Suppose that the input data satisfy the Clipper assumptions (S6)–(S7) and $m = n$. Then for any $q \in (0, 1)$ and either definition of constrast scores in (S10) or (S11), the contrast-score cutoff T^{BC} found by the BC procedure guarantees that the discoveries have the FDR under q:*

$$\text{FDR} = \mathbb{E}\left[\frac{\text{card}\left(\{j \in \mathcal{N} : C_j \geq T^{\text{BC}}\}\right)}{\text{card}(\{j : C_j \geq T^{\text{BC}}\}) \vee 1}\right] \leq q\,,$$

where $\mathcal{N} = \{j : \mu_{Xj} = \mu_{Yj}\}$ denotes the set of uninteresting features.

The proof of Theorem 1 (Supp. Section S8) requires two key ingredients: Lemma 1, which states important properties of contrast scores, and Lemma 2 from [8], which states a property of a Bernoulli process with independent but not necessarily identically distributed random variables. The cutoff T^{BC} can be viewed as a stopping time of a Bernoulli process.

Lemma 1 *Suppose that the input data that satisfy the Clipper assumptions (S6)–(S7) and $m = n$, and that Clipper constructs contrast scores $\{C_j\}_{j=1}^d$ based on (S10) or (S11). Denote $S_j = \text{sign}\,(C_j) \in \{-1, 0, 1\}$. Then $\{S_j\}_{j=1}^d$ satisfy the following properties:*

(a) $S_1, \ldots, S_d$ are mutually independent ;

(b) $\mathbb{P}(S_j = 1) = \mathbb{P}(S_j = -1)$ for all $j \in \mathcal{N}$;

(c) $\{S_j\}_{j \in \mathcal{N}} \perp \mathcal{C}$.

Lemma 2 *Suppose that $Z_1, \ldots, Z_d$ are independent with $Z_j \sim \text{Bernoulli}(\rho_j)$, and $\min_j \rho_j \geq \rho > 0$. Let J be a stopping time in reverse time with respect to the filtration $\{\mathcal{F}_j\}$, where*

$$\mathcal{F}_j = \sigma\left(\{(Z_1 + \cdots + Z_j),\, Z_{j+1},\, \cdots,\, Z_d\}\right)\,, \tag{S13}$$

with $\sigma(\cdot)$ denoting a σ-algebra. Then

$$\mathbb{E}\left[\frac{1 + J}{1 + Z_1 + \cdots + Z_J}\right] \leq \rho^{-1}.$$

Here we give a brief intuition about how Lemma 2 bridges Lemma 1 and Theorem 1 for FDR control. First we note that the false discovery proportion (FDP), whose expectation is the FDR, satisfies

$$\begin{aligned}
\mathrm{FDP} &:= \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j : C_j \geq T^{\mathrm{BC}}\}\right) \vee 1} && \text{(S14)}\\
&= \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}\right) + 1} \cdot \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\right\}\right) + 1}{\mathrm{card}\left(\{j : C_j \geq T^{\mathrm{BC}}\}\right) \vee 1} && \text{(S15)}\\
&\leq \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}\right) + 1} \cdot \frac{\mathrm{card}\left(\left\{j : C_j \leq -T^{\mathrm{BC}}\right\}\right) + 1}{\mathrm{card}\left(\{j : C_j \geq T^{\mathrm{BC}}\}\right) \vee 1} && \text{(S16)}\\
&\leq \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}\right) + 1} \cdot q\,, && \text{(S17)}
\end{aligned}$$

where the last inequality follows from the definition of T^{BC} (S12).

By its definition, if T^{BC} exists, it is positive. This implies that Clipper would never call the features with $C_j = 0$ as discoveries. Here we sketch the idea of proving Theorem 1 by considering a simplified case where $\mathcal{C}$ is fixed instead of being random; that is, we assume the features with non-zero contrast scores to be known. Then, without loss of generality, we assume $\mathcal{C} = \{1, \ldots, d\}$. Then we order the absolute values of uninteresting features' contrast scores, i.e., elements in $\{|C_j| : j \in \mathcal{N}\}$, from the largest to the smallest, denoted by $|C_{(1)}| \geq |C_{(2)}| \geq \cdots \geq |C_{(N)}|$. Let $J = \sum_{j \in \mathcal{N}} \mathbb{1}\left(|C_j| \geq T^{\mathrm{BC}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{BC}. When $J > 0$, $|C_{(1)}| \geq \cdots \geq |C_{(J)}| \geq T^{\mathrm{BC}}$. Define $Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, N$. Then for each order k, the following holds

$$\begin{aligned}
C_{(k)} \geq T^{\mathrm{BC}} &\iff \left|C_{(k)}\right| \geq T^{\mathrm{BC}} \text{ and } C_{(k)} > 0 \iff k \leq J \text{ and } Z_k = 0\,;\\
C_{(k)} \leq -T^{\mathrm{BC}} &\iff \left|C_{(k)}\right| \geq T^{\mathrm{BC}} \text{ and } C_{(k)} < 0 \iff k \leq J \text{ and } Z_k = 1\,.
\end{aligned}$$

Then the upper bound of FDP becomes

$$\begin{aligned}
\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}) + 1} \cdot q &= \frac{\sum_{k=1}^{N} \mathbb{1}\left(C_{(k)} \geq T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{N} \mathbb{1}\left(C_{(k)} \leq -T^{\mathrm{BC}}\right)} \cdot q\\
&= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geq T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leq -T^{\mathrm{BC}}\right)} \cdot q\\
&= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \cdot q\\
&= \left(\frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\right) \cdot q\,.
\end{aligned}$$

By Lemma 1(a)–(b), $Z_k \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(0.5)$, which together with Lemma 1(c) satisfy the condition of Lemma 2 and make $\rho = 0.5$. Then by Lemma 2, we have

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] \leq \mathbb{E}\left[\frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\right] \cdot q \leq (\rho^{-1} - 1) \cdot q = q\,,$$

which is the statement of Theorem 1. The complete proof of Theorem 1 is in Supp. Section S8.

S2.2.1 An optional, heuristic fix if the BC procedure makes no discoveries

Although the BC procedure has theoretical guarantee of FDR control, it lacks power when the number of replicates $m = n$, the target FDR threshold q, and the number of features d are all small (e.g., $m = n = 1$, $q = 0.01$ and $d = 1000$ in Fig. S14). As a result, the BC procedure may lead to no discoveries. In that case, Clipper implements a heuristic fix—an approximate p-value Benjamini-Hochberg (aBH) procedure—to increase the power. The aBH procedure constructs an empirical null distribution of contrast scores by additionally assuming that uninteresting features' contrast scores follow a symmetric distribution around zero; it then computes approximate p-values of features based on the empirical null distribution, and finally it uses the BH procedure [1] to threshold the approximate p-values.

Definition S7 (The aBH procedure) *Given contrast scores $\{C_j\}_{j=1}^d$, an empirical null distribution is defined on $\mathcal{E} := \{C_j : C_j < 0; j = 1, \ldots, d\} \cup \{-C_j : C_j < 0; j = 1, \ldots, d\}$. The aBH procedure defines the approximate p-value of feature j as*

$$p_j := \frac{\sum_{c \in \mathcal{E}} \mathbb{1}(c \geq C_j)}{\mathrm{card}(\mathcal{E}) \vee 1}.$$

Then it applies the BH procedure with the target FDR threshold q to $\{p_j\}_{j=1}^d$ to call discoveries.

S2.3 Enrichment analysis with any numbers of replicates m and n

When $m \neq n$, the BC procedure cannot guarantee FDR control because Lemma 1 no longer holds. To control the FDR in a more general setting ($m = n$ or $m \neq n$), Clipper constructs contrast scores via permutation of replicates across conditions. The idea is that, after permutation, every feature becomes uninteresting and can serve as its own negative control.

Definition S8 (Permutation) *We define σ as permutation, i.e., a bijection from the set $\{1, \cdots, m+n\}$ onto itself, and we rewrite the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$ into a matrix $\mathbf{W}$:*

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1m} & W_{1(m+1)} & \cdots & W_{1(m+n)} \\ & \vdots & & & \vdots & \\ W_{d1} & \cdots & W_{dm} & W_{d(m+1)} & \cdots & W_{d(m+n)} \end{bmatrix} := \begin{bmatrix} X_{11} & \cdots & X_{1m} & Y_{11} & \cdots & Y_{1n} \\ & \vdots & & & \vdots & \\ X_{d1} & \cdots & X_{dm} & Y_{d1} & \cdots & Y_{dn} \end{bmatrix}.$$

We then apply σ to permute the columns of $\mathbf{W}$ and obtain

$$\mathbf{W}_\sigma := \begin{bmatrix} W_{1\sigma(1)} & \cdots & W_{1\sigma(m)} & W_{1\sigma(m+1)} & \cdots & W_{1\sigma(m+n)} \\ & \vdots & & & \vdots & \\ W_{d\sigma(1)} & \cdots & W_{d\sigma(m)} & W_{d\sigma(m+1)} & \cdots & W_{d\sigma(m+n)} \end{bmatrix},$$

from which we obtain the permuted measurements $\{(\boldsymbol{X}_j^\sigma, \boldsymbol{Y}_j^\sigma)\}_{j=1}^d$, where

$$\begin{aligned} \boldsymbol{X}_j^\sigma &:= \left(W_{j\sigma(1)}, \ldots, W_{j\sigma(m)}\right)^\top, \\ \boldsymbol{Y}_j^\sigma &:= \left(W_{j\sigma(m+1)}, \ldots, W_{j\sigma(m+n)}\right)^\top. \end{aligned} \tag{S18}$$

In the enrichment analysis, if two permutations σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\},$$

then we define σ and σ' to be in one equivalence class. That is, permutations in the same equivalence class lead to the same division of $m+n$ replicates (from the two conditions) into two groups with sizes m and n. In total, there are $\binom{m+n}{m}$ equivalence classes of permutations.

We define σ_0 as the identity permutation such that $\sigma_0(i) = i$ for all $i \in \{1, \cdots, m+n\}$. In addition, Clipper randomly samples h equivalence classes $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the other $h_{\max} := \binom{m+n}{m} - 1$ equivalence classes (after excluding the equivalence class containing σ_0). Note that $h_{\max}$ is the maximum value h can take.

Clipper then obtains $\left\{(\boldsymbol{X}_j^{\sigma_0}, \boldsymbol{Y}_j^{\sigma_0}), (\boldsymbol{X}_j^{\sigma_1}, \boldsymbol{Y}_j^{\sigma_1}), \cdots, (\boldsymbol{X}_j^{\sigma_h}, \boldsymbol{Y}_j^{\sigma_h})\right\}_{j=1}^d$, where $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ are the permuted measurements based on σ_ℓ, $\ell = 0, \ldots, h$. Then Clipper computes $T_j^{\sigma_\ell} := t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ to indicate the degree of "interestingness" of feature j reflected by $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$. Note that Clipper chooses t^{diff} instead of $t^{\max}$ because empirical evidence shows that t^{diff} leads to better power. Sorting $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ gives

$$T_j^{(0)} \geq T_j^{(1)} \geq \cdots \geq T_j^{(h)}.$$

Then Clipper defines the contrast score of feature j, $j = 1, \ldots, d$, in one of two ways:

$$C_j := \begin{cases} T_j^{(0)} - T_j^{(1)} & \text{if } T_j^{(0)} = T_j^{\sigma_0} \\ T_j^{(1)} - T_j^{(0)} & \text{otherwise} \end{cases} \qquad \textbf{difference contrast score}, \tag{S19}$$

or

$$C_j := \begin{cases} \left|T_j^{(0)}\right| & \text{if } T_j^{(0)} = T_j^{\sigma_0} > T_j^{(1)} \\ 0 & \text{if } T_j^{(0)} = T_j^{(1)} \\ -\left|T_j^{(0)}\right| & \text{otherwise} \end{cases} \qquad \textbf{maximum contrast score}. \tag{S20}$$

The intuition behind the contrast scores is that, if $C_j < 0$, then $\mathbb{1}(T_j^{(0)} = T_j^{\sigma_0}) = 0$, which means that at least one of $T_j^{\sigma_1}, \ldots, T_j^{\sigma_h}$ (after random permutation) is greater than $T_j^{\sigma_0}$ calculated from the original data (identity permutation), suggesting that feature j is likely an uninteresting feature in enrichment analysis. Motivated by Gimenez and Zou [9], we propose the following procedure for Clipper to control the FDR under the target level $q \in (0, 1)$.

Definition S9 (Gimenez-Zou (GZ) procedure for thresholding contrast scores [9]) *Given $h \in \{1, \cdots, h_{\max}\}$ and contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0\,;\ j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The GZ procedure finds a contrast-score cutoff T^{GZ} based on the target FDR threshold $q \in (0, 1)$ as:*

$$T^{\text{GZ}} := \min\left\{t \in \mathcal{C} : \frac{\frac{1}{h} + \frac{1}{h}\text{card}\left(\{j : C_j \leq -t\}\right)}{\text{card}\left(\{j : C_j \geq t\}\right) \vee 1} \leq q\right\} \tag{S21}$$

and outputs $\left\{j : C_j \geq T^{\text{GZ}}\right\}$ as discoveries.

Theorem 2 *Suppose that the input data that satisfy the Clipper assumptions (S6)–(S7). Then for any $q \in (0, 1)$ and either definition of contrast scores in (S19) or (S20), the contrast-score cutoff T^{GZ} found by the GZ procedure (S21) guarantees that the discoveries have the FDR under q:*

$$\text{FDR} = \mathbb{E}\left[\frac{\text{card}\left(\left\{j \in \mathcal{N} : C_j \geq T^{\text{GZ}}\right\}\right)}{\text{card}\left(\{j : C_j \geq T^{\text{GZ}}\}\right) \vee 1}\right] \leq q,$$

where $\mathcal{N}$ denotes the set of uninteresting features.

The proof of Theorem 2 (Supp. Section S8) is similar to that of Theorem 1 and requires two key ingredients: Lemma 2, which is also used in the proof of Theorem 1, and Lemma 3, which is similar

to Lemma 1 and is about the properties of signs of $\{C_j\}_{j=1}^d$. The cutoff T^{GZ} can also be viewed as a stopping time of a Bernoulli process.

Lemma 3 *For input data that satisfy the Clipper assumptions* (S6) *and* (S7)*, Clipper constructs contrast scores* $\{C_j\}_{j=1}^d$ *based on* (S20) *or* (S19)*. Denote* $S_j = \mathrm{sign}(C_j) \in \{-1, 0, 1\}$*. Then* $\{S_j\}_{j=1}^d$ *and* $\{C_j\}_{j=1}^d$ *satisfy the following properties:*

(a) $S_1, \ldots, S_d$ *are mutually independent ;*

(b) $\mathbb{P}(S_j = 1) \leq \frac{1}{h+1}$ *for all* $j \in \mathcal{N}$*;*

(c) $\{S_j\}_{j \in \mathcal{N}} \perp \mathcal{C}$*.*

We note that the GZ procedure is also applicable to the enrichment analysis with equal numbers of replicates, i.e., $m = n$ (Section S2.2). We will compare the GZ procedure against the BC procedure in our results.

S2.4 Differential analysis with $m + n > 2$

For differential analysis, Clipper also uses permutation to construct contrast scores. When $m \neq n$, the equivalence classes of permutations are defined the same as for the enrichment analysis with $m \neq n$. When $m = n$, there is a slight change in the definition of equivalence classes of permutations: if σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\} \text{ or } \{\sigma'(m+1), \cdots, \sigma'(2m)\},$$

then we say that σ and σ' are in one equivalence class. In total, there are $h_{\mathrm{total}} := \binom{m+n}{m}$ (when $m \neq n$) or $\binom{2m}{m}/2$ (when $m = n$) equivalence classes of permutations. Hence, to have more than one equivalence class, we cannot perform differential analysis with $m = n = 1$; in other words, the total number of replicates $m + n$ must be at least 3.

Then Clipper randomly samples $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the $h_{\max} := h_{\mathrm{total}} - 1$ equivalence classes that exclude the class containing σ_0, i.e., the identity permutation. Note that $h_{\max}$ is the maximum value h can take. Next, Clipper computes $T_j^{\sigma_\ell} := \left|t^{\mathrm{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})\right|$, where $\boldsymbol{X}_j^{\sigma_\ell}$ and $\boldsymbol{Y}_j^{\sigma_\ell}$ are the permuted data defined in (S18), and it defines C_j as the contrast score of feature j, $j = 1, \ldots, d$, in the same ways as in (S19) or (S20).

Same as in the enrichment analysis with $m \neq n$, Clipper also uses the GZ procedure [9] to set a cutoff on contrast scores to control the FDR under the target level $q \in (0, 1)$, following Theorem 2.

S2.5 Clipper variant algorithms

For nomenclature, we assign the following names to Clipper variant algorithms, each of which combines a contrast score definition with a thresholding procedure.

- **Clipper-diff-BC**: difference contrast score $C_j = t^{\mathrm{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ (S10) and BC procedure (Definition S6);
- **Clipper-diff-aBH**: difference contrast score $C_j = t^{\mathrm{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and aBH procedure (Definition S7);
- **Clipper-diff-GZ**: difference contrast score $\tau_j = T_j^{(0)} - T_j^{(1)}$ (S19) and GZ procedure (Definition S9);

- **Clipper-max-BC**: maximum contrast score $C_j = t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ (S11) and BC procedure;
- **Clipper-max-aBH**: maximum contrast score $C_j = t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and aBH procedure;
- **Clipper-max-GZ**: maximum contrast score $\tau_j = T_j^{(0)}$ (S20) and GZ procedure.

S2.6 `R` package "Clipper"

In the `R` package `Clipper`, the default implementation is as follows. Based on the power comparison results in Section S3 and Figs. S14, S15, S16, and S17, Clipper uses Clipper-diff-BC as the default algorithm for the enrichment analysis with equal numbers of replicates; when there are no discoveries, Clipper suggests users to increase the target FDR threshold q or to use the Clipper-diff-aBH algorithm with the current q. For the enrichment analysis with different numbers of replicates under two conditions or the differential analysis, Clipper uses the Clipper-max-GZ algorithm by default.

S3 Comparison of Clipper variant algorithms

We compared Clipper variant algorithms applicable to each experimental design. Based on the comparison results, we selected a variant algorithm as the default Clipper implementation for each design.

- **1vs1 enrichment analysis.** Under each of the 12 settings, we compared Clipper-diff-BC, Clipper-diff-aBH, Clipper-max-BC, and Clipper-max-aBH (Section S2.5), the only four Clipper variant algorithms applicable to 1vs1 enrichment analysis. The results in Fig. S14 show that, regardless of the contrast scores being difference (diff) or maximum (max), the BC procedure always guarantees the FDR control under a range of target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$. Notably, in terms of power, the two contrast scores consistently have different advantages under the two background scenarios: Clipper-max-BC has higher power under the homogeneous background, while Clipper-diff-BC is more powerful under the heterogeneous background. Considering that the heterogeneous scenario is prevalent in high-throughput biological data, the difference contrast score is preferred. As the power of Clipper-diff-BC drops when q is too small ($q \leq 3\%$) and d is not too large ($d = 1000$), we consider the aBH procedure as an alternative to control the FDR. The results show that Clipper-diff-aBH is indeed more powerful when Clipper-diff-BC lacks power; however, Clipper-diff-aBH cannot guarantee the exact FDR control as Clipper-diff-BC does. Therefore, Clipper uses **Clipper-diff-BC** by default in 1vs1 enrichment analysis, and it recommends users to increase q when too few discoveries are made; if users reject this option, then Clipper would use Clipper-diff-aBH to increase power for the current q.
- **2vs1 enrichment analysis.** Under each of the 6 settings, we compared Clipper-diff-GZ and Clipper-max-GZ (Section S2.5), the only two Clipper variant algorithms applicable to 2vs1 enrichment analysis. For either algorithm, we further compared two numbers of permutation equivalence classes: $h = 1$ or 2, where the latter is $h_{\max} = \binom{3}{1} - 1$—the maximum number of equivalence classes that do not include the identity permutation. Note that h is a required input parameter for the GZ procedure. The results in Fig. S15 show that, regardless of h and the contrast score definition—maximum (max) or difference (diff), the GZ procedure always guarantees the FDR control under all target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$. In terms of power, Clipper-max-GZ($h = 1$) is consistently more powerful than the other three Clipper variants under all settings. Therefore, Clipper uses **Clipper-max-GZ($h = 1$)** by default in enrichment analysis with unequal numbers of replicates under two conditions.

- **3vs3 enrichment analysis.** Under each of the 12 settings, we compared five Clipper variant algorithms: Clipper-diff-BC, Clipper-diff-aBH, Clipper-max-BC, Clipper-max-aBH, and Clipper-max-GZ (Section S2.5). Fig. S16 shows the comparison of the first four variants: regardless of the contrast scores being difference (diff) or maximum (max), the BC procedure simultaneously guarantees the FDR control and achieves good power under a range of target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$. Similar to the results in the 1vs1 enrichment analysis, Clipper-max-BC has higher power under the homogeneous background, while Clipper-diff-BC is more powerful under the heterogeneous background. By the same reasoning—the prevalent heterogeneous scenarios in high-throughput biological data—we prefer the difference contrast score. Unlike the 1vs1 enrichment analysis, here Clipper-diff-BC is consistently as powerful as Clipper-diff-aBH, even when q is small, but Clipper-diff-aBH cannot guarantee the exact FDR control. Therefore, Clipper-diff-BC achieves the overall best performance among the first four Clipper variants. Given that the GZ procedure is also applicable to this setting, we further compared Clipper-diff-BC with Clipper-max-GZ($h = 1$), the most powerful Clipper variant with the GZ procedure and the default Clipper implementation in the 2vs1 enrichment and differential analyses and the 3vs3 differential analysis. The results in Fig. S18 show that while both **Clipper-diff-BC** and Clipper-max-GZ($h = 1$) control the FDR, the former is more powerful. Hence, we will use Clipper-diff-BC as the default when both conditions have more than one and the same number of replicates.

- **2vs1 differential analysis.** Similar to 2vs1 enrichment analysis, under each of the 6 settings, we compared Clipper-diff-GZ and Clipper-max-GZ (Section S2.5) with $h = 1$ or 2. The results in Fig. S15 show that, regardless of h and the contrast score definition—maximum (max) or difference (diff), the GZ procedure always guarantees the FDR control under a range of target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$. Notably, in terms of power, Clipper-diff-GZ($h = 2$) is the most powerful when when q is very small ($q \leq 2\%$) under Poisson and negative binomial settings, while Clipper-max-GZ($h = 1$) is the most powerful otherwise. Considering that Clipper-max-GZ($h = 1$) outperforms the other three Clipper variants in most cases, Clipper uses **Clipper-max-GZ($h = 1$)** by default in 2vs1 differential analysis, and it recommends users to use Clipper-diff-GZ($h = 2$) when too few discoveries are made.

- **3vs3 differential analysis.** Under each of the 12 settings, we compared Clipper-diff-GZ, and Clipper-max-GZ (Section S2.5) with $h = 1, 3$ or 9, where $h = 9$ is $h_{\max} = \binom{6}{3}/2 - 1$—the maximum number of equivalence classes that do not include the identity permutation. The results in Fig. S17 show that, regardless of h and the contrast score definition—maximum (max) or difference (diff), the GZ procedure always guarantees the FDR control under a range of target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$. In terms of power, Clipper-max-GZ($h = 1$) is consistently more powerful than the other Clipper variant algorithms under all settings. Therefore, Clipper uses **Clipper-max-GZ($h = 1$)** by default in 3vs3 differential analysis.

In summary, whenever Clipper-diff-BC is applicable (enrichment analysis with equal number of replicates under two conditions), it is chosen as the default Clipper implementation; otherwise, Clipper-max-GZ($h = 1$) is the default.

S4 Data generation and detailed implementation of the paired approach (a p-value calculation approach) in simulation studies

We describe how we simulated data and how we implemented the paired approach in different simulation settings: 1vs1 enrichment analysis, 2vs1 enrichment analysis, 3vs3 enrichment analysis, 2vs1 differential analysis, and 3vs3 differential analysis, combined with three distribution families (Gaussian, Poisson, and negative binomial) and two background scenarios (homogeneous and heterogeneous). Under some settings, we considered different numbers of features and the existence of outliers.

In each simulation setting, we generated 200 simulated datasets, computed an FDP and an empirical power on each dataset, and averaged the 200 FDPs and 200 empirical powers to approximate the FDR and power, repsectively. For notation simplicity, we use $N(\mu, \sigma^2)$ to denote the Gaussian distribution with mean μ and variance σ^2, $\mathrm{Pois}(\lambda)$ to denote the Poisson distribution with mean λ, and $\mathrm{NB}(\mu, \theta)$ to denote the negative binomial distribution with mean μ and dispersion θ (such that its variance equals $\mu + \theta\mu^2$).

For each design and analysis, we compared the default Clipper implementation with other generic FDR control methods. Specifically, seven generic methods (BH-pool, qvalue-pool, BH-pair-mis, qvalue-pair-mis, BH-pair-2as1, qvalue-pair-2as1, and locfdr-emp) are included in all designs and analyses. The two methods relying on correct model specification, BH-pair-correct and qvalue-pair-correct, are only included in the 3vs3 enrichment and differential analyses, because it is almost impossible to correctly specify a model with fewer than three replicates per condition. The permutation-based method, locfdr-swap, is excluded from the 1vs1 enrichment analysis because it requires at least one condition to have more than one replicate.

In addition to the above designs and analyses, we also compared the default Clipper implementation with BH-pair methods that use parametric or non-parametric tests to calculate p-values when the numbers of replicates are 10 under both conditions for enrichment analysis, i.e., 10vs10 enrichment analysis.

S4.1 1vs1 enrichment analysis

We simulated data with $d = 1000$ and $10{,}000$ features under two background scenarios and three distributional families—a total of 12 settings. In each setting, 10% of the features are interesting ($\mu_{Xj} > \mu_{Yj}$), and the rest are uninteresting (with $\mu_{Xj} = \mu_{Yj}$). Recall that $\mathcal{N}$ denotes the set of uninteresting features.

Gaussian distribution

We simulated data from Gaussian using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 0$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 0$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $N(5, 1)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $N(0, 2)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $N(5, 1)$.
- We independently generated X_{j1} from $N(\mu_{Xj}, 1)$ and Y_{j1} from $N(\mu_{Yj}, 1)$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that the null distribution of $X_{j1} - Y_{j1}$, $j = 1, \ldots, d$ is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{1}{d-1} \sum_{j=1}^{d} \left(X_{j1} - \frac{1}{d} \sum_{j=1}^{d} X_{j1} \right)^2 + \frac{1}{d-1} \sum_{j=1}^{d} \left(Y_{j1} - \frac{1}{d} \sum_{j=1}^{d} Y_{j1} \right)^2 .$$

This is a misspecified model that assumes that μ_{Xj}'s are all equal and so are μ_{Yj}'s. Then we computed the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $X_{j1} - Y_{j1}$ in $N(0, \hat{\sigma}^2)$, i.e., $1 - \Phi\left(\frac{X_{j1} - Y_{j1}}{\hat{\sigma}}\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $N(Y_{j1}, 1)$ conditioning on the observed Y_{j1} as the null distribution of X_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of X_{j1} in $N(Y_{j1}, 1)$, i.e., $1 - \Phi\left(X_{j1} - Y_{j1}\right)$.

Poisson distribution

We simulated data from Poisson using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{Pois}(20)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- We independently generated X_{j1} from $\mathrm{Pois}(\mu_{Xj})$ and Y_{j1} from $\mathrm{Pois}(\mu_{Yj})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we first defined a log-transformation $f(x) = \log(x + 0.01)$, which we applied to X_{j1} and Y_{j1}, $j = 1, \ldots, d$. We assumed that the null distribution of $f(X_{j1}) - f(Y_{j1})$, $j = 1, \ldots, d$ is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{1}{d-1} \sum_{j=1}^{d} \left(f(X_{j1}) - \frac{1}{d} \sum_{j=1}^{d} f(X_{j1}) \right)^2 + \frac{1}{d-1} \sum_{j=1}^{d} \left(f(Y_{j1}) - \frac{1}{d} \sum_{j=1}^{d} f(Y_{j1}) \right)^2 .$$

This model misspecifies the Poisson distribution as the log-normal distribution.

Then we computed the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $f(X_{j1}) - f(Y_{j1})$ in $N(0, \hat{\sigma}^2)$, i.e., $1 - \Phi\left(\frac{f(X_{j1}) - f(Y_{j1})}{\hat{\sigma}}\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $\mathrm{Pois}(Y_{j1})$ conditioning on the observed Y_{j1} as the null distribution of X_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of X_{j1} in $\mathrm{Pois}(Y_{j1})$, i.e., $\mathbb{P}(Z \geq X_{j1})$ where $Z \sim \mathrm{Pois}(Y_{j1})$.

Negative binomial distribution

We simulated data from negative binomial using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.

- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^d$ i.i.d. from $\mathrm{NB}(20, 20^{-1})$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j\notin\mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.

- We independently generated X_{j1} from $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ and Y_{j1} from $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that for each uninteresting feature j, Y_{j1} and X_{j1} follow the same Poisson distribution. We calculated the p-value of feature j from a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ using function `poisson.test` in R package `stats`.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $\mathrm{NB}(Y_{j1}, Y_{j1}^{-1})$ conditioning on the observed Y_{j1} as the null distribution of X_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of X_{j1} in $\mathrm{NB}(Y_{j1}, Y_{j1}^{-1})$.

S4.2 2vs1 enrichment analysis

We simulated data with $d = 10{,}000$ features under two background scenarios and three distributional families—a total of 6 settings. In each setting, 10% of the features are interesting ($\mu_{Xj} > \mu_{Yj}$) and the rest are uninteresting (with $\mu_{Xj} = \mu_{Yj}$). Recall that $\mathcal{N}$ denotes the set of uninteresting features.

Gaussian distribution

We simulated data from Gaussian using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 0$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 0$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j\notin\mathcal{N}}$ i.i.d. from $N(5, 1)$.

- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j\in\mathcal{N}}$ i.i.d. from $N(0, 2)$ and set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. We next generated $\{\mu_{Yj}\}_{j\notin\mathcal{N}}$ i.i.d. from $N(0, 2)$ and $\{\mu_{Xj}\}_{j\notin\mathcal{N}}$ i.i.d. from $N(5, 1)$.

- We independently generated X_{j1} and X_{j2} from $N(\mu_{Xj}, 1)$ and Y_{j1} from $N(\mu_{Yj}, 1)$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that the null distribution of $\frac{1}{2}(X_{j1} + X_{j2}) - Y_{j1}$, $j = 1, \ldots, d$, is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{1}{2(2d-1)} \sum_{j=1}^{d} \sum_{i=1}^{2} \left(X_{ji} - \frac{1}{2d} \sum_{j=1}^{d} \sum_{i=1}^{2} X_{ji} \right)^2 + \frac{1}{d-1} \sum_{j=1}^{d} \left(Y_{j1} - \frac{1}{d} \sum_{j=1}^{d} Y_{j1} \right)^2 .$$

This is a misspecified model that assumes μ_{Xj}'s are all equal and so are μ_{Yj}'s. Then we computed the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $\frac{1}{2}(X_{j1} + X_{j2}) - Y_{j1}$ in $N(0, \hat{\sigma}^2)$, i.e., $1 - \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{\hat{\sigma}}\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $N(Y_{j1}, 1/2)$ conditioning on the observed Y_{j1} as the null distribution of $\frac{1}{2}(X_{j1} + X_{j2})$. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $\frac{1}{2}(X_{j1} + X_{j2})$ in $N(Y_{j1}, 1/2)$, i.e., $1 - \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{1/\sqrt{2}}\right)$.

Poisson distribution

We simulated data from Poisson using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{Pois}(20)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- We independently generated X_{j1} and X_{j2} from $\mathrm{Pois}(\mu_{Xj})$ and Y_{j1} from $\mathrm{Pois}(\mu_{Yj})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we first defined a log-transformation $f(x) = \log(x + 0.01)$, which we applied to X_{j1} and Y_{j1}, $j = 1, \ldots, d$. We assumed that the null distribution of $f(X_{j1}) + f(X_{j2}) - 2f(Y_{j1})$, $j = 1, \ldots, d$ is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{6}{d-1} \sum_{j=1}^{d} \left(f(Y_{j1}) - \frac{1}{d} \sum_{j=1}^{d} f(Y_{j1}) \right)^2 .$$

This model misspecifies the Poisson distribution as the log-normal distribution.

Then we computed the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $f(X_{j1}) + f(X_{j2}) - 2f(Y_{j1})$ in $N(0, \hat{\sigma}^2)$, i.e., $1 - \Phi\left(\frac{f(X_{j1}) + f(X_{j2}) - 2f(Y_{j1})}{\hat{\sigma}}\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we assumed that for each uninteresting feature j, X_{j1} and X_{j2} independently follow $\mathrm{Pois}(Y_{j1})$ conditioning on the observed Y_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ by performing a one-sample Poisson test using the R function `poisson.test` for the null hypothesis $H_0 : \mu_{X_j} = Y_{j1}$ against the alternative hypothesis $H_1 : \mu_{X_j} > Y_{j1}$.

Negative binomial distribution

We simulated data from negative binomial using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{NB}(20, 20^{-1})$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.
- We independently generated X_{j1} and X_{j2} from $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ and Y_{j1} from $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that for each uninteresting feature j, X_{ji}, $i = 1, 2$ and Y_{j1} follow the same Poisson distribution. We calculated the p-value of feature j from a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ using the function `poisson.test` in R package `stats`.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $\mathrm{NB}(2Y_{j1}, (2Y_{j1})^{-1})$ conditioning on the observed Y_{j1} as the null distribution of $X_{j1} + X_{j2}$. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $X_{j1} + X_{j2}$ in $\mathrm{NB}(2Y_{j1}, (2Y_{j1})^{-1})$.

S4.3 3vs3 enrichment analysis

We simulated data with and without outliers under two background scenarios and three distributional families—a total of 12 settings. In each setting, we generated $d = 10{,}000$ features, among which 10% are interesting (with $\mu_{Xj} > \mu_{Yj}$) and the rest are uninteresting (with $\mu_{Xj} = \mu_{Yj}$).

Under the settings with outliers, we generated $\{O^X_{ji} : j = 1, \ldots, d; i = 1, \ldots, 3\}$ and $\{O^Y_{ji} : j = 1, \ldots, d; i = 1, \ldots, 3\}$ i.i.d. from $\mathrm{Bernoulli}(0.1)$, where $O^X_{ji} = 1$ or $O^Y_{ji} = 1$ indicates X_{ji} or Y_{ji} is an outlier, respectively. Under settings without outliers, $O^X_{ji} = O^Y_{ij} = 0$ for all $j = 1, \ldots, d; i = 1, \ldots, 3$.

Gaussian distribution

- Under the homogeneous background scenario, we set $\mu_{Yj} = 0$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 0$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $N(5, 1)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^d$ i.i.d. from $N(0, 2)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $N(5, 1)$.
- We independently generated X_{ji} from $N(\mu_{Xj}, 1)$ if $O^X_{ji} = 0$ or from the top 1% percentile of $N(\mu_{Xj}, 1)$ if $O^X_{ji} = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $N(\mu_{Yj}, 1)$ if $O^Y_{ji} = 0$ or from the top 1% percentile of $N(\mu_{Yj}, 1)$ if $O^Y_{ji} = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$.
- For the results in Figure S11, we generated correlated features. We first selected 10 groups of features (2 groups of interesting features and 8 groups of uninteresting features), with each group containing 200 features. For each group k, we used $k_1, \ldots, k_{200}$ to denote the indices of the 200 features within that group and generated $\{X_{k_l i}\}_{l=1}^{200}$ from a multivariate Gaussian distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k = (\mu_{Xk_1}, \ldots, \mu_{Xk_{200}})$ and $\boldsymbol{\Sigma}_k$ is a matrix with diagonal entries as 1 and other entries as a fixed correlation. In our simulation, the fixed correlation took two values: 0.2 and 0.4.

To implement the correct paired approach (as in BH-pair-correct and qvalue-pair-correct), we calculated the p-value of feature j from a two-sample t-test with equal variance for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we calculated the p-value of feature j from a two-sample t-test with unequal variance for the null hypothesis $H_0 : \mu_{X_j} = \mu_{Y_j}$ against the alternative hypothesis $H_1 : \mu_{X_j} > \mu_{Y_j}$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we assumed that for each uninteresting feature j, X_{ji}, $i = 1, \ldots, 3$ are i.i.d. Gaussian with mean $\bar{Y}_j$ conditioning on the observed $\bar{Y}_j$ and unknown variance. We calculated the p-value of feature j using a one-sample t-test for the null hypothesis $H_0 : \mu_{X_j} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{X_j} > \bar{Y}_j$.

Poisson distribution

We simulated data from Poisson using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{Pois}(20)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{Pois}(40)$.
- We independently generated X_{ji} from $\mathrm{Pois}(\mu_{Xj})$ if $O_{ji}^{X} = 0$ or from the top 1% percentile of $\mathrm{Pois}(\mu_{Xj})$ if $O_{ji}^{X} = 1$, $j = 1, \ldots, d$, $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $\mathrm{Pois}(\mu_{Yj})$ if $O_{ji}^{Y} = 0$ or from the top 1% percentile of $\mathrm{Pois}(\mu_{Yj})$ if $O_{ji}^{Y} = 1$, $j = 1, \ldots, d$; $1, \ldots, 3$.

To implement the correct paired approach (as in BH-pair-correct and qvalue-pair-correct), we calculated the p-value of feature j by performing a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ using the function `poisson.test` in R package `stats`.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we first defined a log-transformation $f(x) = \log(x + 0.01)$, which we applied to X_{ji} and Y_{ji}, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. We assumed that for each uninteresting feature j, $\{f(X_{ji})\}_{i=1}^{3}$ and $\{f(Y_{ji})\}_{i=1}^{3}$ follow Gaussian distributions with mean $\mu_{f(Xj)}$ and $\mu_{f(Yj)}$, respectively. Then we computed the p-value of feature j using a two-sample equal variance t-test for the null hypothesis $H_0 : \mu_{f(Xj)} = \mu_{f(Yj)}$ against the alternative hypothesis $H_1 : \mu_{f(Xj)} > \mu_{f(Yj)}$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we assumed that for each uninteresting feature j, $\{X_{ji}\}_{i=1}^{3}$ follow $\mathrm{Pois}(\bar{Y}_j)$ conditioning on the observed $\bar{Y}_j$. We calculated the p-value of feature j by performing a one-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{Xj} > \bar{Y}_j$ using R function `poisson.test` from package `stats`.

Negative binomial distribution

We simulated data from negative binomial using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{NB}(20, 20^{-1})$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For interesting features, we generated $\{\mu_{Xj}\}_{j \notin \mathcal{N}}$ i.i.d. from $\mathrm{NB}(45, 45^{-1})$.
- We independently generated X_{ji} from $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ if $O_{ji}^{X} = 0$ or from the top 1% percentile of $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ if $O_{ji}^{X} = 1$, $j = 1, \ldots, d$, $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$ if $O_{ji}^{Y} = 0$ or from the top 1% percentile of $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$ if $O_{ji}^{Y} = 1$, $j = 1, \ldots, d$, $i = 1, \ldots, 3$.

To implement the correct paired approach (as in BH-pair-correct and qvalue-pair-correct), we performed a two-sample negative binomial test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative $H_1 : \mu_{Xj} > \mu_{Yj}$ using $T_j := \sum_{i=1}^{3} X_{ji} - \sum_{i=1}^{3} Y_{ji}$ as the test statistic. We computed the p-value of feature j as the right tail probability

$$\mathbb{P}(T_j \geq t_j) = \sum_{k_1=0}^{\infty} \sum_{k_2=k_1+t_j}^{\infty} \mathbb{P}\left(\sum_{i=1}^{3} X_{ji} \geq k_2\right) \mathbb{P}\left(\sum_{i=1}^{3} Y_{ji} = k_1\right),$$

where t_j is the realization of T_j, $\mathbb{P}(\sum_{i=1}^3 X_{ji} \geq k_2)$ and $\mathbb{P}(\sum_{i=1}^3 Y_{ji} = k_1)$ can be estimated from the null distribution of X_{ji} and Y_{ji}, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. As $\sum_{i=1}^3 X_{ji}$ and $\sum_{i=1}^3 Y_{ji}$ follow the same distribution under null, we estimated μ_{Xj} and μ_{Yj} as $\hat{\mu}_{Xj} = \hat{\mu}_{Yj} := (\sum_{i=1}^3 X_{ji} + \sum_{i=1}^3 Y_{ji})/6$. Then, we calculated $\mathbb{P}(\sum_{i=1}^3 X_{ji} \geq k_2)$ and $\mathbb{P}(\sum_{i=1}^3 Y_{ji} = k_1)$ using the estimated distribution of X_{ji} and Y_{ji} as $\mathrm{NB}(\hat{\mu}_{Xj}, (\hat{\mu}_{Xj})^{-1})$ and $\mathrm{NB}(\hat{\mu}_{Yj}, (\hat{\mu}_{Yj})^{-1})$, respectively, $j = 1, \ldots, d$; $i = 1, \ldots, 3$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that for each uninteresting feature j, $\{X_{ji}\}_{j=1}^3$ and $\{Y_{ji}\}_{j=1}^3$ follow the same Poisson distribution. We calculated the p-value of feature j from a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ using function `poisson.test` in R package `stats`.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $\mathrm{NB}(\sum_{i=1}^3 Y_{ji}, (\sum_{i=1}^3 Y_{ji})^{-1})$ conditioning on the observed $\sum_{i=1}^3 Y_{ji}$ as the null distribution of $\sum_{i=1}^3 X_{ji}$. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the right tail probability of $\sum_{i=1}^3 X_{ji}$ in $\mathrm{NB}(\sum_{i=1}^3 Y_{ji}, (\sum_{i=1}^3 Y_{ji})^{-1})$.

S4.4 10vs10 enrichment analysis

We simulated data without outliers under heterogeneous background scenario and three distributional families—a total of 3 settings. In each setting, we generated $d = 10{,}000$ features, among which 10% are interesting (with $\mu_{Xj} > \mu_{Yj}$) and the rest are uninteresting (with $\mu_{Xj} = \mu_{Yj}$).

The data generation under the Gaussian, Poisson, and negative binomial distributions is the same as in the 3vs3 enrichment analysis (Section S4.3) except that we set the number of replicates to 10 under each condition, and we did not generate outliers.

The correct paired approaches in BH-pair-parametric are the same as the corresponding BH-pair-correct in the 3vs3 enrichment analysis (Section S4.3) except that, under the negative binomial distribution, the test statistic T_j and its null distribution should have the number of replicates changed from 3 to 10.

To implement the non-parametric paired approaches, we calculated the p-value of feature j from the one-sided two-sample Wilcoxon rank-sum test (using `R` function `wilcox.test` in package `stats`) in BH-pair-Wilcoxon and from the one-sided two-sample permutation test (using `R` function `oneway_test` in package `coin`) in BH-pair-permutation.

S4.5 2vs1 differential analysis

We simulated data with $d = 10{,}000$ features under two background scenarios and three distributional families—a total of 6 settings. In each setting, we set 10% features as "up-regulated" with $\mu_{Xj} > \mu_{Yj}$ and another 10% features as "down-regulated" with $\mu_{Xj} < \mu_{Yj}$.

Gaussian distribution

We simulated data from Gaussian using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 0$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 0$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $N(5, 1)$. For down-regulated features, generated μ_{Xj} i.i.d. from $N(-5, 1)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^d$ i.i.d. from $N(0, 2)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $N(5, 1)$. For down-regulated features, generated μ_{Xj} i.i.d. from $N(-5, 1)$.

- We independently generated X_{j1} and X_{j2} from $N(\mu_{Xj}, 1)$ and Y_{j1} from $N(\mu_{Yj}, 1)$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that the null distribution of $\frac{1}{2}(X_{j1} + X_{j2}) - Y_{j1}$, $j = 1, \ldots, d$, is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{1}{2(2d-1)} \sum_{j=1}^{d} \sum_{i=1}^{2} \left(X_{ji} - \frac{1}{2d} \sum_{j=1}^{d} \sum_{i=1}^{2} X_{ji} \right)^2 + \frac{1}{d-1} \sum_{j=1}^{d} \left(Y_{j1} - \frac{1}{d} \sum_{j=1}^{d} Y_{j1} \right)^2 .$$

This is a misspecified model assuming that μ_{Xj}'s are all equal and so are μ_{Yj}'s. Then we computed the p-value of feature $j = 1, \ldots, d$ as the two-sided tail probability of $\frac{1}{2}(X_{j1} + X_{j2}) - Y_{j1}$ in $N(0, \hat{\sigma}^2)$, i.e., $2 \cdot \min\left(1 - \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{\hat{\sigma}}\right), \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{\hat{\sigma}}\right)\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $N(Y_{j1}, 1)$ conditioning on the observed Y_{j1} as the null distribution of X_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the two-sided tail probability of $\frac{1}{2}(X_{j1} + X_{j2})$ in $N(Y_{j1}, 1/2)$, i.e., $2 \cdot \min\left(1 - \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{1/\sqrt{2}}\right), \Phi\left(\frac{\frac{1}{2}(X_{j1}+X_{j2})-Y_{j1}}{1/\sqrt{2}}\right)\right)$.

Poisson distribution

We simulated data from Poisson using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(60)$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(5)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{Pois}(20)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(60)$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(5)$.
- We independently generated X_{j1} and X_{j2} from $\mathrm{Pois}(\mu_{Xj})$ and Y_{j1} from $\mathrm{Pois}(\mu_{Yj})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we first defined a log-transformation $f(x) = \log(x + 0.01)$, which we applied to X_{j1} and Y_{j1}, $j = 1, \ldots, d$. We assumed that the null distribution of $f(X_{j1}) + f(X_{j2}) - 2f(Y_{j1})$, $j = 1, \ldots, d$ is $N(0, \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{6}{d-1} \sum_{j=1}^{d} \left(f(Y_{j1}) - \frac{1}{d} \sum_{j=1}^{d} f(Y_{j1}) \right)^2 .$$

This model misspecifies the Poisson distribution as the log-normal distribution. Then we computed the p-value of feature $j = 1, \ldots, d$ as the two-sided tail probability of $f(X_{j1}) + f(X_{j2}) - 2f(Y_{j1})$ in $N(0, \hat{\sigma}^2)$, i.e., $2 \cdot \min\left(1 - \Phi\left(\frac{f(X_{j1})+f(X_{j2})-2f(Y_{j1})}{\hat{\sigma}}\right), \Phi\left(\frac{f(X_{j1})+f(X_{j2})-2f(Y_{j1})}{\hat{\sigma}}\right)\right)$, where Φ is the cumulative distribution function of $N(0, 1)$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we assumed that for each uninteresting feature j, X_{j1} and X_{j2} independently follow $\mathrm{Pois}(Y_{j1})$ conditioning on the observed Y_{j1}. Then we calculated the p-value of feature $j = 1, \ldots, d$ by performing a one-sample Poisson test using the R function `poisson.test` for the null hypothesis $H_0 : \mu_{X_j} = Y_{j1}$ against the alternative hypothesis $H_1 : \mu_{X_j} \neq Y_{j1}$.

Negative binomial distribution

We simulated data from negative binomial using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 30$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 30$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(70, 70^{-1})$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(7, 7^{-1})$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $\mathrm{NB}(30, 30^{-1})$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(70, 70^{-1})$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(7, 7^{-1})$.
- We independently generated X_{j1} and X_{j2} from $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ and Y_{j1} from $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$, $j = 1, \ldots, d$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that for each uninteresting feature j, X_{j1}, X_{j2}, and Y_{j1} follow the same Poisson distribution. We calculated the p-value of feature j from a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \mu_{Yj}$ using the function `poisson.test` in R package `stats`.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $\mathrm{NB}(2Y_{j1}, (2Y_{j1})^{-1})$ conditioning on the observed Y_{j1} as the null distribution of $X_{j1}+X_{j2}$. Then we calculated the p-value of feature $j = 1, \ldots, d$ as the two-sided tail probability of $X_{j1}+X_{j2}$ in $\mathrm{NB}(2Y_{j1}, (2Y_{j1})^{-1})$, i.e., twice the smaller of the left-tail and right-tail probabilities.

S4.6 3vs3 differential analysis

We simulated data with or without outliers under two background scenarios and three distributional families—a total of 12 settings. In each setting, we generated $d = 10{,}000$ features, among which 10% features were "up-regulated features" with $\mu_{Xj} > \mu_{Yj}$ and another 10% were "down-regulated features" with $\mu_{Xj} < \mu_{Yj}$.

Under the settings with outliers, we generated $\{O_{ji}^X : j = 1, \ldots, d; i = 1, \ldots, 3\}$ and $\{O_{ji}^Y : j = 1, \ldots, d; i = 1, \ldots, 3\}$ i.i.d. from $\mathrm{Bernoulli}(0.1)$, where $O_{ji}^X = 1$ or $O_{ji}^Y = 1$ indicates X_{ji} or Y_{ji} is an outlier, respectively. Under settings without outliers, $O_{ji}^X = O_{ij}^Y = 0$ for all $j = 1, \ldots, d; i = 1, \ldots, 3$.

Gaussian distribution

- Under the homogeneous background scenario, we set $\mu_{Yj} = 0$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 0$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $N(5, 1)$. For down-regulated features, generated μ_{Xj} i.i.d. from $N(-5, 1)$.
- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^{d}$ i.i.d. from $N(0, 2)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $N(5, 1)$. For down-regulated features, generated μ_{Xj} i.i.d. from $N(-5, 1)$.
- We independently generated X_{ji} from $N(\mu_{Xj}, 1)$ if $O_{ji}^X = 0$ or from the top 1% percentile of $N(\mu_{Xj}, 1)$ if $O_{ji}^X = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $N(\mu_{Yj}, 1)$ if $O_{ji}^Y = 0$ or from the top 1% percentile of $N(\mu_{Yj}, 1)$ if $O_{ji}^Y = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$.

To implement the correct paired approach (as in BH-pair-correct and qvalue-pair-correct), we calculated the p-value of feature j from a two-sample t-test with equal variance for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \mu_{Yj}$.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we calculated the p-value of feature j from a two-sample t-test with unequal variance for the null hypothesis $H_0 : \mu_{X_j} = \mu_{Y_j}$ against the alternative hypothesis $H_1 : \mu_{X_j} \neq \mu_{Y_j}$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we treated $N(\bar{Y}_j, 1)$ conditioning on observed $\bar{Y}_j$ as the null distribution of X_{ji}, $i = 1, \ldots, 3$. We calculated the p-value of feature j using a one-sample t-test for the null hypothesis $H_0 : \mu_{X_j} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{X_j} \neq \bar{Y}_j$.

Poisson distribution

We simulated data from Poisson using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 20$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 20$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(40)$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(5)$.

- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^d$ i.i.d. from $\mathrm{Pois}(20)$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(40)$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{Pois}(5)$.

- We independently generated X_{ji} from $\mathrm{Pois}(\mu_{Xj})$ if $O_{ji}^X = 0$ or from the top 1% percentile of $\mathrm{Pois}(\mu_{Xj})$ if $O_{ji}^X = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $\mathrm{Pois}(\mu_{Yj})$ if $O_{ji}^Y = 0$ or from the top 1% percentile of $\mathrm{Pois}(\mu_{Yj})$ if $O_{ji}^Y = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$.

To implement the correct paired approach (as in BH-pair-correct and qvalue-pair-correct), we calculated the p-value of feature j by performing a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \mu_{Yj}$ using function `poisson.test` in R package `stats`.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we first defined a log-transformation $f(x) = \log(x + 0.01)$, which we applied to X_{ji} and Y_{ji}, $j = 1, \ldots, d$; $i = 1, \ldots, 3$. We assumed that for each uninteresting feature j, $\{f(X_{ji})\}_{i=1}^3$ and $\{f(Y_{ji})\}_{i=1}^3$ follow Gaussian distributions with mean $\mu_{f(Xj)}$ and $\mu_{f(Yj)}$, respectively. Then we computed the p-value of feature j using a two-sample equal variance t-test for the null hypothesis $H_0 : \mu_{f(Xj)} = \mu_{f(Yj)}$ against the alternative hypothesis $H_1 : \mu_{f(Xj)} \neq \mu_{f(Yj)}$.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we assumed that for each uninteresting feature j, $\{X_{ji}\}_{i=1}^3$ follow $\mathrm{Pois}(\bar{Y}_j)$ conditioning on the observed $\bar{Y}_j$. We calculated the p-value of feature j by performing a one-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \bar{Y}_j$ using the function `poisson.test` in R package `stats`.

Negative binomial distribution

We simulated data from negative binomial using the following procedure:

- Under the homogeneous background scenario, we set $\mu_{Yj} = 30$ for all d features. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj} = 30$ for $j \in \mathcal{N}$. For up-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(70, 70^{-1})$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(7, 7^{-1})$.

- Under the heterogeneous background scenario, we generated $\{\mu_{Yj}\}_{j=1}^d$ i.i.d. from $\mathrm{NB}(30, 30^{-1})$. For uninteresting features, we set $\mu_{Xj} = \mu_{Yj}$ for $j \in \mathcal{N}$. For up-regulated features, we gener-

ated μ_{Xj} i.i.d. from $\mathrm{NB}(70, 70^{-1})$. For down-regulated features, we generated μ_{Xj} i.i.d. from $\mathrm{NB}(7, 7^{-1})$.

- We independently generated X_{ji} from $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ if $O_{ji}^X = 0$ or from the top 1% percentile of $\mathrm{NB}(\mu_{Xj}, \mu_{Xj}^{-1})$ if $O_{ji}^X = 1$; $j = 1, \ldots, d$, $i = 1, \ldots, 3$. Similarly, we independently generated Y_{ji} from $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$ if $O_{ji}^Y = 0$ or from the top 1% percentile of $\mathrm{NB}(\mu_{Yj}, \mu_{Yj}^{-1})$ if $O_{ji}^Y = 1$, $j = 1, \ldots, d$; $i = 1, \ldots, 3$.

To implement the correct paired approach with unknown dispersion (as in BH-pair-correct and qvalue-pair-correct), we performed a two-sample negative binomial test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \mu_{Yj}$ using the coefficient from the negative binomial regression as the test statistic. Specifically, for each feature j we performed a negative binomial regression by treating the condition labels as a categorical covariate and feature j's measurements as the response. We implemented this regression analysis using function `glm.nb` in R package `MASS` and extracted the p-value of the coefficient as the p-value of feature j. The dispersion parameter was not pre-specified but estimated by `glm.nb`.

To implement the correct paired approach with known dispersion, we performed a similar negative binomial regression but with the pre-specified dispersion parameter 30^{-1} for each feature j. Then we computed the feature's p-value as the p-value of the coefficient of the condition covariate. We implemented this regression analysis using function `glm` in R package `stats`.

To implement the misspecified paired approach (as in BH-pair-mis and qvalue-pair-mis), we assumed that for each uninteresting feature j, $\{X_{ji}\}_{j=1}^3$ and $\{Y_{ji}\}_{j=1}^3$ follow the same Poisson distribution. We calculated the p-value of feature j from a two-sample Poisson test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} \neq \mu_{Yj}$ using function `poisson.test` in R package `stats`.

To implement the 2as1 paired approach (as in BH-pair-2as1 and qvalue-pair-2as1), we first used function `glm.nb` in R package `MASS` to estimate $\hat{\mu}_{Yj}$ and $\hat{\theta}_{Yj}$ from $\{Y_{ji}\}_{j=1}^3$. Then we computed the p-value of feature j by treating $\mathrm{NB}(3\hat{\mu}_{Yj}, (3\hat{\theta}_{Yj})^{-1})$ as the null distribution of $\sum_{i=1}^3 X_{ji}$ and calculated its two-sided tail probability, i.e., twice the smaller of the left-tail and right-tail probabilities.

S5 Bioinformatic methods with FDR control functionality

S5.1 Peak calling methods for ChIP-seq data

MACS2 MACS2 [10] uses sliding windows with a fixed length across the genome and identifies peaks by using a Poisson distribution to model the read counts within each window, which has one read count per replicate. Specifically, for each region (which is combined from sliding windows), MACS2 performs a one-sample Poisson test to calculate a p-value, where the null distribution is set to be Poisson with its parameter estimated from the background. By thresholding p-values, MACS2 identifies a set of candidate peaks. It also estimates for each candidate peak a q-value by swapping the experimental sample with the background (negative control) sample, and the q-values are used for FDR control. We used MACS2 software (version 2.2.6) with its default settings.

HOMER We used findPeaks, a program in HOMER [11], to perform peak calling on ChIP-seq data. The p-value calculation in findPeaks is similar to that in MACS2; that is, findPeaks also uses the Poisson distribution as the null distribution of read counts in a genomic region, and it also estimates the Poisson mean from the background sample. Then findPeaks identifies peaks by setting thresholds on

p-values and fold-changes (the folder change of a region is defined as the observed read count under the experimental sample divided by the estimated Poisson mean from the the background sample). We used findPeaks version 3.1.9.2.

S5.2 Mascot for peptide identification from MS data

Mascot Mascot uses probability-based scoring to identify PSMs from mass-spectrometry data. We ran Mascot in Proteome Discoverer 2.3.0.523 (ThermoScientific) with the following settings: 10 ppm precursor tolerance; 0.6 Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M). We ran Percolator [12] in conjunction with Mascot with the target/decoy selection mode set to "separate." For Mascot, for a range of target FDR thresholds ($q \in \{1\%, 2\%, \ldots, 10\%\}$), we identified the target PSMs with Mascot q-values no greater than q as discoveries. To prepare the input for Clipper, we set peptide and protein FDRs to 100% to obtain the entire lists of target PSMs and decoy PSMs with their Mascot q-values.

S5.3 Differentially expressed gene (DEG) methods for bulk RNA-seq data

edgeR edgeR models each gene's read counts by using a negative binomial regression, where the condition is incorporated as an indicator covariate, and the condition's coefficient represents the gene-wise differential expression effect [13]. We used `R` package `edgeR` version 3.30.0.

DESeq2 DESeq2 uses a similar negative binomial regression as edgeR to model each gene's read counts under two conditions. DESeq2 differs from edgeR mainly in their estimation of the dispersion parameter in the negative binomial distribution [14]. We used the `R` package `DESeq2` version 1.28.1.

S5.4 Differentially interacting chromatin regions (DIR) methods for Hi-C data

MultiHiCcompare MultiHiCcompare relies on a non-parametric method to jointly normalize multiple Hi-C interaction matrices [15]. It uses a generalized linear model to detect DIRs. MultiHiCcompare is an extension of the `HiCcompare` package [16]. We used `R` package `multiHiCcompare` version 1.6.0.

diffHic diffHic uses the statistical framework of the `edgeR` package to model biological variability and to test for significant differences between conditions [17]. We used `R` package `diffHic` version 1.20.0.

FIND FIND uses a spatial Poisson process to detect chromosomal regions that display a significant difference between two regions' interaction intensity and their neighbouring interaction intensities [18]. We used `R` package `FIND` version 0.99.

S6 Benchmark data generation in omics data applications

S6.1 ChIP-seq data with synthetic spike-in peaks

We used one experimental sample and two control samples (which we refer to as Control 1 and Control 2) from H3K4me3 ChIP-seq data in Chromosome 1 of the cell line GM12878 [19].

(i) We input the experimental sample and Control 1 into MACS2 [10] to call peaks, and we refer to the identified peaks as "tentative peaks."

(ii) Among tentative peaks, we further identified 25 "true peaks" as follows. We validated these peaks by comparing their read coverages in the experimental sample and Control 1 (Fig. S19).

(iii) We created a synthetic experimental sample by adding synthetic true peaks to Control 1. To mimic real H3K4me3 ChIP-seq data, where peaks are located predominantly in promoter regions, we added synthetic true peaks to promoter regions only. Specifically, we randomly sampled 585 genes' promoter regions from Chromosome 1. For each promoter region, we selected one of the 25 true peaks whose length was the closet to the length of the promoter region, and we down-sampled this peak's reads with probability p (that is, every read is sampled with probability p), which was drawn uniformly from 20% to 60%. Then we added the resulting down-sampled peak (i.e., its reads) as a synthetic true peak to that promoter region. After such addition for all the 587 promoter regions, we obtained a synthetic experimental sample.

(iv) We repeated Step (iii) for 20 times to generate 20 synthetic experimental samples. For each experimental sample, we paired it with Control 2 as the background sample. Hence, we obtained 20 synthetic ChIP-seq datasets, each containing 585 synthetic true peaks.

(v) After applying a peak calling method to these 20 synthetic datasets, we evaluated the 20 FDPs and 20 empirical power, which are then averaged as the approximate FDR and power of that peaking call method. In the evaluation, a called peak is a true positive if it overlapped with a synthetic true peak; otherwise, it is a false positive.

S6.2 Real MS standard data

The data generation information will be published in a future manuscript. Interested readers should contact Dr. Leo Wang at lewang@coh.org.

S6.3 Bulk RNA-seq data with synthetic spike-in DEGs

We used the human monocyte RNA-Seq dataset including 17 samples of classical monocytes and 17 samples of nonclassical monocytes [20]. Each sample contains expression levels of $d = 52{,}376$ genes.

(i) We first performed normalization on all 34 samples using the edgeR normalization method Trimmed Mean of M-values (TMM) [21]. We denote the resulting normalized read count matrix of classical and non-classical monocytes by $\mathbf{X}^{\mathrm{cl}}$ and $\mathbf{X}^{\mathrm{ncl}}$, respectively. Following the convention in bioinformatics, the columns and rows of $\mathbf{X}^{\mathrm{cl}}$ and $\mathbf{X}^{\mathrm{ncl}}$ represent biological samples and genes, respectively.

(ii) To define true DEGs, we first computed the log fold change of gene j by $\text{logfc}_j = \log_2\left[(\bar{\mathbf{X}}_j^{\mathrm{ncl}}+1)/(\bar{\mathbf{X}}_j^{\mathrm{cl}}+1)\right]$ for $j = 1, \ldots, d$, where $\mathbf{X}_i^{\mathrm{cl}}$ and $\mathbf{X}_i^{\mathrm{ncl}}$ denote the i-th row vector of $\mathbf{X}^{\mathrm{cl}}$ and $\mathbf{X}^{\mathrm{ncl}}$ respectively and $\bar{\cdot}$ denotes the average of elements in a vector. We added the pseudo-count of 1 to avoid division by 0. We defined true DEGs as those with $\text{logfc}_j \geq 4$ or $\text{logfc}_j \leq -4$.

(iii) We generated synthetic data with 3 samples under both the experimental and background conditions, a typical design in bulk RNA-seq experiments. Specifically, if gene j is a DEG, we randomly sampled without replacement 3 values from $\mathbf{X}_j^{\mathrm{cl}}$ as counts under the experimental condition, and another 3 values from $\mathbf{X}_j^{\mathrm{ncl}}$ as counts under the background condition. If gene j is not a DEG, we randomly sampled 6 values without replacement from $(\mathbf{X}_j^{\mathrm{cl}}, \mathbf{X}_j^{\mathrm{ncl}})$ and randomly split them into 3 and 3 counts under two conditions.

(iv) We repeated Step (iii) for 200 times to generate 200 synthetic datasets.

The human monocyte RNA-Seq data set is available in the NCBI Sequence Read Archive (SRA) under accession number SRP082682 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=srp082682).

S6.4 Hi-C data with synthetic spike-in DIRs

The real Hi-C interaction matrix contains the pairwise interaction intensities of 250 binned genomic regions in Chromosome 1. It is from the cell line GM12878 and available in the NCBI Gene Expression Omnibus(GEO) under accession number GSE63525. We denote the real interaction matrix as $\mathbf{X}^{\text{real}}$. Because $\mathbf{X}^{\text{real}}$ is symmetric, we only focus on its upper triangular part.

(i) Among the $(250 \times 250 - 250)/2 = 31{,}125$ upper triangular entries (i.e., region pairs), we selected 404 entries as true up-regulated DIRs, and 550 entries as true down-regulated DIRs (Fig. S20).

(ii) Next, for the (i, j)-th entry, we generated a log fold change, denoted by f_{ij}, between the two conditions as follows. We simulated f_{ij} from truncated Normal$(100/|i - j|, 0.5^2)$ with support $[0.05, \infty)$ if the (i, j)-th entry is up-regulated, or from truncated Normal$(-100/|i - j|, 0.5^2)$ with support $(-\infty, -0.05]$ if the (i, j)-th entry is down-regulated; if the (i, j)-th entry is not differential, we set $f_{ij} = 0$.

(iii) Then we specify the mean measurement of the (i, j)-th entry under the two conditions as $\mu_{Xij} = [\mathbf{X}^{\text{real}}]_{ij}$ and $\mu_{Yij} = [\mathbf{X}^{\text{real}}]_{ij} \cdot e^{f_{ij}}$, respectively.

(iv) We generated synthetic read counts of the (i, j)-th entry from $\text{NB}(\mu_{Xij}, 1000^{-1})$ and $\text{NB}(\mu_{Yij}, 1000^{-1})$ respectively under the two conditions.

(v) We repeated Step (iv) for 200 times to generate 200 synthetic datasets.

S7 Implementation of Clipper in omics data applications

Below we briefly introduce the implementation of Clipper in the four omics data applications. All the results were obtained by running using `R` package `Clipper` (see package vignette for details: https://github.com/JSB-UCLA/Clipper/blob/master/vignettes/Clipper.pdf).

S7.1 Peak calling from ChIP-seq data

(i) We consider each genomic location, i.e., a base pair, as a feature and each ChIP-seq sample as a replicate under the experimental or background condition. Then we consider the read count of each location in each sample as the corresponding feature's measurement. Doing so, we summarized ChIP-seq data into a $d \times (m + n)$ matrix, where d is the number of locations, and m and n are the numbers of experimental and control samples, respectively. We then applied Clipper to perform an enrichment analysis to obtain the contrast score C_j for each location j. In our study, $m = n = 1$, so the default Clipper implementation is Clipper-diff-BC.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) We then used existing peak calling methods, e.g., MACS2 and HOMER, to call candidate peaks with the least stringent q-value cutoff. For example, when we used MACS2, we set the q-value cutoff as 1.

(iv) We computed the contrast score of each candidate peak as the median of the contrast scores of all the locations within.

(v) The candidate peaks with contrast scores greater than or equal to T_q are called discoveries.

S7.2 Peptide identification from mass spectrometry data

(i) We consider each mass spectrum as a feature and its target/decoy PSM as a replicate under the experimental/background condition respectively. Then we consider $-\log_{10}(\text{q-value} + 0.01)$ as the measurement of each PSM, where the q-value is output by Mascot. Doing so, we summarized the Mascot output into a $d \times (m+n)$ matrix, where d is the number of mass spectra, and m and n are the numbers of experimental and control samples, respectively. We then applied Clipper to perform an enrichment analysis to obtain a contrast score C_j for each mass spectrum j. If the mass spectrum has no decoy or background measurement, we set $C_j = 0$. In our study, $m = n = 1$, so the default Clipper implementation is Clipper-diff-BC.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The target PSMs whose mass spectra have contrast scores greater than or equal to T_q are called discoveries.

S7.3 DEG identification from bulk RNA-seq data

(i) We consider each gene as a feature and the class label—classical and non-classical human monocytes—as the two conditions. Then we consider $\log_2$-transformed read counts with a pseudocount 1 as measurements. Doing so, we summarized the gene expression matrix into a $d \times (m+n)$ matrix, where d is the number of genes, and m and n are the numbers of samples under the two conditions, respectively. We then applied Clipper to perform a differential analysis to obtain a contrast score C_j for each gene. In our study, $m = n = 3$, so the default Clipper implementation is Clipper-max-GZ with $h = 1$.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The genes with contrast scores greater than or equal to T_q are called discoveries.

S7.4 DIR identification from Hi-C data

(i) We consider each pair of genomic regions as a feature and manually created two conditions. Then we consider $\log$-transformed read counts as measurements. Doing so, we summarized the gene expression matrix into a $d \times (m+n)$ matrix, where d is the total pairs of genomic regions, and m and n are the numbers of samples under the two conditions, respectively. We then applied Clipper to perform a differential analysis to obtain a contrast score C_j for each pair of genomic regions. In our study, $m = n = 2$, so the default Clipper implementation is Clipper-max-GZ with $h = 1$.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The pairs of genomic regions with contrast scores greater than or equal to T_q are called discoveries.

S8 Proofs

S8.1 Proof of Theorem 1

We first prove Theorem 1, which relies on Lemmas 1 and 2. Here we only include the proof of Lemma 1 and defer the proof of Lemma 2 to Section S8.3.

Proof 1 (Proof of Lemma 1) *Here we prove that Lemma 1 holds when C_j is constructed using (S10); the proof is similar when C_j is constructed using (S11).*

When input data satisfy (S6) and (S7) and $m = n$, properties (a) and (b) can be derived directly. To prove property (c), it suffices to prove that for any $j \in \mathcal{N}$ with $C_j \neq 0$, S_j is independent of $|C_j|$.

Note that $\bar{X}_j$ and $\bar{Y}_j$ are i.i.d for $j \in \mathcal{N}$ when $m = n$. Hence for any measurable set $\mathcal{A} \subset [0, +\infty)$,

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) = \mathbb{P}\left(t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \in \mathcal{A}\right) = \mathbb{P}\left(t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j) \in \mathcal{A}\right) = \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A}\right).$$

The first equality holds because $t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) = C_j = |C_j|$ when $S_j = 1$. The second equality holds because $t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and $t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j)$ are identically distributed when $j \in \mathcal{N}$. The third equality holds because $t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j) = -C_j$; if $-C_j \in \mathcal{A}$, then $S_j = -1$.

Because $\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) + \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A}\right) = \mathbb{P}\left(|C_j| \in \mathcal{A}\right)$, it follows that

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) = \frac{1}{2}\mathbb{P}\left(|C_j| \in \mathcal{A}\right) = \mathbb{P}(S_j = 1)\mathbb{P}\left(|C_j| \in \mathcal{A}\right),$$

where the last equality holds because $\mathbb{P}(S_j = 1) = 1/2$ by property (b).

Hence, S_j and $|C_j|$ are independent $\forall j \in \mathcal{N}$.

Proof 2 (Proof of Theorem 1) *Define a random subset of $\mathcal{N}$ as $\mathcal{M} := \mathcal{N} \backslash \{j \in \mathcal{N} : C_j = 0\} = \{j \in \mathcal{N} : S_j \neq 0\}$.*

First note that by Lemma 1(b), $\mathbb{P}(S_j = -1) = \mathbb{P}(C_j < 0) = 1/2$ for all $j \in \mathcal{M} \subset \mathcal{N}$. Assume without loss of generality that $\mathcal{M} = \{1, \ldots, d'\}$. We order $\{|C_j| : j \in \mathcal{M}\}$, from the largest to the smallest, denoted by $|C_{(1)}| \geq |C_{(2)}| \geq \cdots \geq |C_{(d')}|$. Let $J = \sum_{j \in \mathcal{N}} \mathbb{1}\left(|C_j| \geq T^{\text{BC}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{BC}. When $J > 0$, $|C_{(1)}| \geq \cdots \geq |C_{(J)}| \geq T^{\text{BC}}$. Define $Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, d'$. Then for each order k, the following holds

$$C_{(k)} \geq T^{\text{BC}} \Longleftrightarrow \left|C_{(k)}\right| \geq T^{\text{BC}} \text{ and } C_{(k)} > 0 \Longleftrightarrow k \leq J \text{ and } Z_k = 0\,;$$
$$C_{(k)} \leq -T^{\text{BC}} \Longleftrightarrow \left|C_{(k)}\right| \geq T^{\text{BC}} \text{ and } C_{(k)} < 0 \Longleftrightarrow k \leq J \text{ and } Z_k = 1\,.$$

Then

$$\begin{aligned}
\frac{\text{card}(\{j \in \mathcal{M} : C_j \geq T^{\text{BC}}\})}{\text{card}(\{j \in \mathcal{M} : C_j \leq -T^{\text{BC}}\}) + 1} &= \frac{\sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \geq T^{\text{BC}}\right)}{1 + \sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \leq -T^{\text{BC}}\right)} \\
&= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geq T^{\text{BC}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leq -T^{\text{BC}}\right)} \\
&= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \\
&= \frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\,.
\end{aligned}$$

Because $\{S_j\}_{j \in \mathcal{N}}$ is independent of $\mathcal{C}$ (Lemma 1(c)), Lemma 1(a)-(b) still holds after $C_1, \ldots, C_{d'}$ are

reordered as $C_{(1)}, \ldots, C_{(d')}$. *Thus* $Z_1, \ldots, Z_{d'}$ *are i.i.d. from* $\mathrm{Bernoulli}(1/2)$. *To summarize, it holds that*

$$\{Z_j\}_{j \in \mathcal{M}} \,\Big|\, \mathcal{M} \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(1/2)\,.$$

Then by applying Lemma 2 and making $\rho = 0.5$, *we have:*

$$\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leq -T^{\mathrm{BC}}\}) + 1} \,\Bigg|\, \mathcal{M}\right] \leq 1 \tag{S22}$$

Then

$$\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{BC}}\}) \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}) + 1} \cdot \frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}) + 1}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{BC}}\}) \vee 1}\right] \\
&\leq \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}) + 1} \cdot \frac{\mathrm{card}(\{j : C_j \leq -T^{\mathrm{BC}}\}) + 1}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{BC}}\}) \vee 1}\right] \\
&\leq q \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{BC}}\}) + 1}\right] \\
&\leq q \cdot \mathbb{E}\left[\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geq T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leq -T^{\mathrm{BC}}\}) + 1} \,\Bigg|\, \mathcal{M}\right]\right] \\
&\leq q,
\end{aligned}$$

where $\mathcal{M}$ *is random subset of* $\mathcal{N}$ *such that for each* $j \in \mathcal{M}$, $|C_j| > 0$. *The last inequality follows from* (S22).

S8.2 Proof of Theorem 2

We then prove Theorem 2, which relies on Lemmas 2 and 3. Here we introduce the proof of Lemma 3 and defer the proof of Lemma 2 to Section S8.3.

Proof 3 (Proof of Lemma 3) *With input data satisfying* (S6) *and* (S7), C_j *constructed from* (S19) *or* (S20), *property (a) can be derived directly.*

To show property (b), note that for each uninteresting feature $j \in \mathcal{N}$, $\boldsymbol{X}_j$ *and* $\boldsymbol{Y}_j$ *are from the same distribution; thus* $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ *are identically distributed. Define an event* $\mathcal{E}_j := \left\{\sum_{\ell=0}^h \mathbb{1}(T_j^{\sigma_\ell} = T_j^{(0)}) = 1\right\}$, *which indicates that* $T_j^{(0)}$, *the maximizer of* $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$, *is unique. Then conditional on* $\mathcal{E}_j$, *the maximizer is equally likely to be any of* $\{0, \ldots, h\}$, *and it follows that* $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j) = \mathbb{P}(T_j^{\sigma_0} = T_j^{(0)} \mid \mathcal{E}_j) = 1/(h+1)$. *Conditioning on that* $\mathcal{E}_j$ *does not happen,* $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j^c) = 0$. *Thus* $\mathbb{P}(S_j = 1) = \mathbb{P}(S_j = 1 \mid \mathcal{E}_j)\mathbb{P}(\mathcal{E}_j) + \mathbb{P}(S_j = 1 \mid \mathcal{E}_j^c)\mathbb{P}(\mathcal{E}_j^c) \leq 1/(h+1)$.

The proof of property (c) is similar to the Proof of Lemma 1(c). It suffices to show that for any $j \in \mathcal{N}$ *with* $C_j \neq 0$ *(that is,* $\mathcal{E}_j$ *occurs),* S_j *is independent of* $|C_j|$. *As* $\boldsymbol{X}_j$ *and* $\boldsymbol{Y}_j$ *are from the same distribution,* $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ *are identically distributed. Hence for any measurable set* $\mathcal{A} \subset [0, +\infty)$,

$$\begin{aligned}
\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) &= \mathbb{P}\left(T_j^{\sigma_0} = T_j^{(0)}, |C_j| \in \mathcal{A} \,\middle|\, \mathcal{E}_j\right) \\
&= \frac{1}{h}\mathbb{P}\left(T_j^{\sigma_0} \neq T_j^{(0)}, |C_j| \in \mathcal{A} \,\middle|\, \mathcal{E}_j\right) \\
&= \frac{1}{h}\mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right).
\end{aligned}$$

The first equality holds because $T_j^{\sigma_0} = T_j^{(0)}$ when $S_j = 1$. The second equality holds because $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ are identically distributed when $j \in \mathcal{N}$. The third equality holds because $T_j^{\sigma_0} \neq T_j^{(0)}$ when $S_j = -1$.

Because $\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) + \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right)$, it follows that

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \frac{1}{h+1}\mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \mathbb{P}(S_j = 1 \mid \mathcal{E}_j)\mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) ,$$

where the last equality holds because $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j) = 1/(h+1)$.

Hence, S_j and $|C_j|$ are independent $\forall j \in \mathcal{N}$ with $C_j \neq 0$.

Proof 4 (Proof of Theorem 2) *Define a random subset of $\mathcal{N}$ as $\mathcal{M} := \mathcal{N} \backslash \{j \in \mathcal{N} : C_j = 0\} = \{j \in \mathcal{N} : S_j \neq 0\}$. Assume without loss of generality that $\mathcal{M} = \{1, \ldots, d'\}$. We order $\{|C_j| : j \in \mathcal{M}\}$, from the largest to the smallest, denoted by $|C_{(1)}| \geq |C_{(2)}| \geq \cdots \geq |C_{(d')}|$. Let $J = \sum_{j \in \mathcal{N}} \mathbb{1}\left(|C_j| \geq T^{\mathrm{GZ}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{GZ}. When $J > 0$, $|C_{(1)}| \geq \cdots \geq |C_{(J)}| \geq T^{\mathrm{GZ}}$. Define $Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, d'$. Then for each order k, the following holds:*

$$C_{(k)} \geq T^{\mathrm{GZ}} \Longleftrightarrow \left|C_{(k)}\right| \geq T^{\mathrm{GZ}} \text{ and } C_{(k)} > 0 \Longleftrightarrow k \leq J \text{ and } Z_k = 0\,;$$
$$C_{(k)} \leq -T^{\mathrm{GZ}} \Longleftrightarrow \left|C_{(k)}\right| \geq T^{\mathrm{GZ}} \text{ and } C_{(k)} < 0 \Longleftrightarrow k \leq J \text{ and } Z_k = 1\,.$$

Then it follows that

$$\begin{aligned}
\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leq -T^{\mathrm{GZ}}\}) + 1} &= \frac{\sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \geq T^{\mathrm{GZ}}\right)}{1 + \sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \leq -T^{\mathrm{GZ}}\right)} \\
&= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geq T^{\mathrm{GZ}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leq -T^{\mathrm{GZ}}\right)} \\
&= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \\
&= \frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\,.
\end{aligned}$$

Because $\{S_j\}_{j \in \mathcal{N}}$ is independent of $\mathcal{C}$ (Lemma 1(c)), Lemma 1(a)-(b) still holds after $C_1, \ldots, C_{d'}$ are reordered as $C_{(1)}, \ldots, C_{(d')}$. Thus $Z_1, \ldots, Z_{d'}$ are i.i.d. from $\mathrm{Bernoulli}(\rho_k)$. To summarize, it holds that

$$\{Z_j\}_{j \in \mathcal{M}} \,\Big|\, \mathcal{M} \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\rho_k)\,.$$

Then by applying Lemma 2 and making $\rho = h/(h+1)$, we have:

$$\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leq -T^{\mathrm{GZ}}\}) + 1}\right] \leq 1/h\,. \tag{S23}$$

Then

$$
\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{GZ}}\}) + 1} \cdot \frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{GZ}}\}) + 1}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&\leq h \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{GZ}}\}) + 1} \cdot \frac{\frac{1}{h}\mathrm{card}(\{j : C_j \leq -T^{\mathrm{GZ}}\}) + \frac{1}{h}}{\mathrm{card}(\{j : C_j \geq T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&\leq hq \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leq -T^{\mathrm{GZ}}\}) + 1}\right] \\
&\leq hq \cdot \mathbb{E}\left[\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geq T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leq -T^{\mathrm{GZ}}\}) + 1} \,\middle|\, \mathcal{M}\right]\right] \\
&\leq q\,,
\end{aligned}
$$

where the second inequality follows from the definition of T_{GZ} (S21) *and the last inequality follows from* (S23).

S8.3 Proof of Lemma 2

Finally, we derive Lemma 2 by following the same proof same as in [8], which relies on Lemma 4 and Corollary 1.

Lemma 4 *Suppose that* $Z_1, \ldots, Z_d \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\rho)$. *Let* J *be a stopping time in reverse time with respect to the filtration* $\{\mathcal{F}_j\}$, *where* $\mathcal{F}_j = \sigma\left(\{(Z_1 + \cdots + Z_j),\, Z_{j+1},\, \cdots,\, Z_d\}\right)$ *with* $\sigma(\cdot)$ *denoting a* σ*-algebra, and the variables* $Z_1, \ldots, Z_j$ *are exchangeable with respect to* $\{\mathcal{F}_j\}$. *Then*

$$
\mathbb{E}\left[\frac{1 + J}{1 + Z_1 + \cdots + Z_J}\right] \leq \rho^{-1}.
$$

Proof 5 (Proof of Lemma 4) *Define*

$$
Y_j = Z_1 + \cdots + Z_j \in \mathcal{F}_j
$$

and define the process

$$
M_j = \frac{1 + j}{1 + Z_1 + \cdots + Z_j} = \frac{1 + j}{1 + Y_j} \in \mathcal{F}_j.
$$

In [3], it is shown that $\mathbb{E}[M_d] \leq \rho^{-1}$. *Therefore, by the optional stopping time theorem it suffices to show that* $\{M_j\}$ *is a supermartingale with respect to* $\{\mathcal{F}_j\}$. *As* $\{Z_1, \ldots, Z_{j+1}\}$ *are exchangeable with respect to* $\mathcal{F}_{j+1}$, *we have*

$$
\mathbb{P}\left(Z_{j+1} = 1 \,|\, \mathcal{F}_{j+1}\right) = \frac{Y_j + 1}{1 + j}.
$$

Therefore, if $Y_{j+1} > 0$,

$$\begin{aligned}
\mathbb{E}\left[M_j \mid \mathcal{F}_{j+1}\right] &= \frac{1+j}{1+Y_{j+1}} \cdot \mathbb{P}\left(Z_{j+1}=0 \mid \mathcal{F}_{j+1}\right) + \frac{1+j}{1+Y_{j+1}-1} \cdot \mathbb{P}\left(Z_{j+1}=1 \mid \mathcal{F}_{j+1}\right) \\
&= \frac{1+j}{1+Y_{j+1}} \cdot \frac{1+j-Y_{j+1}}{1+j} + \frac{1+j}{1+Y_{j+1}-1} \cdot \frac{Y_{j+1}}{1+j} \\
&= \frac{1+j-Y_{j+1}}{1+Y_{j+1}} + 1 \\
&= \frac{1+(j+1)}{1+Y_{j+1}} \\
&= M_{j+1}.
\end{aligned}$$

If instead $Y_{j+1} = 0$, *then trivially* $Y_j = 0$, *and* $M_j = 1 + j < 2 + j = M_{j+1}$. *This proves that* $\{M_j\}$ *is a supermartingale with respect to* $\{\mathcal{F}_j\}$ *as desired.*

Corollary 1 *Suppose that* $\mathcal{A} \subseteq \{1, \ldots, d\}$ *is fixed, while* $Z_1, \ldots, Z_d \overset{i.i.d.}{\sim} \text{Bernoulli}(\rho)$. *Let* J *be a stopping time in reverse time with respect to the filtration* $\{\mathcal{F}_j\}$, *where* $\mathcal{F}_j = \sigma\left(\{\sum_{k \le j, k \in \mathcal{A}} Z_k\} \cup \{Z_k : j < k < d, k \in \mathcal{A}\}\right)$ *with* $\sigma(\cdot)$ *denoting a* σ*-algebra, and the variables* $\{Z_k : k \le j, k \in \mathcal{A}\}$ *are exchangeable with respect to* $\mathcal{F}_j$. *Then*

$$\mathbb{E}\left[\frac{1+\text{card}\left(\{k : k \le J, k \in \mathcal{A}\}\right)}{1+\sum_{k \le J, k \in \mathcal{A}} Z_k}\right] \le \rho^{-1}.$$

Proof 6 (Proof of Corollary 1) *Let* $\mathcal{A} = \{j_1, \ldots, j_m\}$ *where* $1 \le j_1 < \cdots < j_m \le d$. *Then by considering the i.i.d. sequence*

$$Z_{j_1}, \ldots, Z_{j_m}$$

in place of $Z_1, \ldots, Z_d$, *we see that this result is equivalent to Lemma 4.*

Proof 7 (Proof of Lemma 2) *[From [3]] We may assume* $\rho < 1$ *to avoid the trivial case. We first introduce a different definition for* $\{Z_j\}_{j=1}^d$ *by defining a random set* $\mathcal{A} \subseteq \{1, \ldots, d\}$ *where for each* j, *independently,*

$$\mathbb{P}\left(j \in \mathcal{A}\right) = \frac{1-\rho_j}{1-\rho}.$$

We then define random variables $Q_1, \ldots, Q_d \overset{i.i.d.}{\sim} \text{Bernoulli}(\rho)$, *which are generated independently of the random set* $\mathcal{A}$. *Finally, we define*

$$Z_j = Q_j \cdot \mathbb{1}\left(j \in \mathcal{A}\right) + \mathbb{1}\left(j \notin \mathcal{A}\right). \tag{S24}$$

Then $\{Z_j\}_{k=1}^d$ *are mutually independent and* $\mathbb{P}\left(Z_j = 1\right) = 1 - \mathbb{P}(j \in \mathcal{A}) \cdot \mathbb{P}(Q_j = 0) = \rho_j$, *that is,* $Z_j \sim \text{Bernoulli}(\rho_j)$. *This new definition of* $\{Z_j\}_{j=1}^d$ *meet all the conditions required by Lemma 2, so that we can apply this new definition in the following proof.*

As $Z_j = Q_j \cdot \mathbb{1}\left(j \in \mathcal{A}\right) + \mathbb{1}\left(j \notin \mathcal{A}\right)$ *for all* j, *we have*

$$\frac{1+J}{1+Z_1+\cdots+Z_J} = \frac{1+\text{card}\left(\{j \le J : j \in \mathcal{A}\}\right)+\text{card}\left(\{j \le J : j \notin \mathcal{A}\}\right)}{1+\sum_{j \le J, j \in \mathcal{A}} Q_j + \text{card}\left(\{j \le J : j \notin \mathcal{A}\}\right)} \le \frac{1+\text{card}\left(\{j \le J : j \in \mathcal{A}\}\right)}{1+\sum_{j \le J, j \in \mathcal{A}} Q_j}, \tag{S25}$$

where the last step uses the identify $\frac{a+c}{b+c} \le \frac{a}{b}$ *whenever* $0 < b \le a$ *and* $c \ge 0$. *Therefore, it will be sufficient to prove that*

$$\mathbb{E}\left[\frac{1+\text{card}\left(\{j \le J : j \in \mathcal{A}\}\right)}{1+\sum_{j \le J, j \in \mathcal{A}} Q_j} \,\middle|\, \mathcal{A}\right] \le \rho^{-1}, \tag{S26}$$

To prove (S26), first let $\widetilde{Q}_j = Q_j \cdot \mathbb{1}(j \in \mathcal{A})$, and define a filtration $\{\mathcal{F}'_j\}$ where $\mathcal{F}'_j$ is the σ-algebra generated as

$$\mathcal{F}'_j = \sigma\left(\left\{\widetilde{Q}_1 + \cdots + \widetilde{Q}_j, \widetilde{Q}_{j+1}, \cdots, \widetilde{Q}_d, \mathcal{A}\right\}\right).$$

Next for any j, by (S24) we see that

$$Z_1 + \cdots + Z_j, Z_{j+1}, \ldots, Z_d \in \mathcal{F}'_j \Rightarrow \mathcal{F}_j \subseteq \mathcal{F}'_j,$$

so J is a stopping time (in reverse time) with respect to $\mathcal{F}'_j$. Finally, since the Q_j's are independent of $\mathcal{A}$, (S26) follows from Corollary 1 after conditioning on $\mathcal{A}$.

S9 Supplementary figures

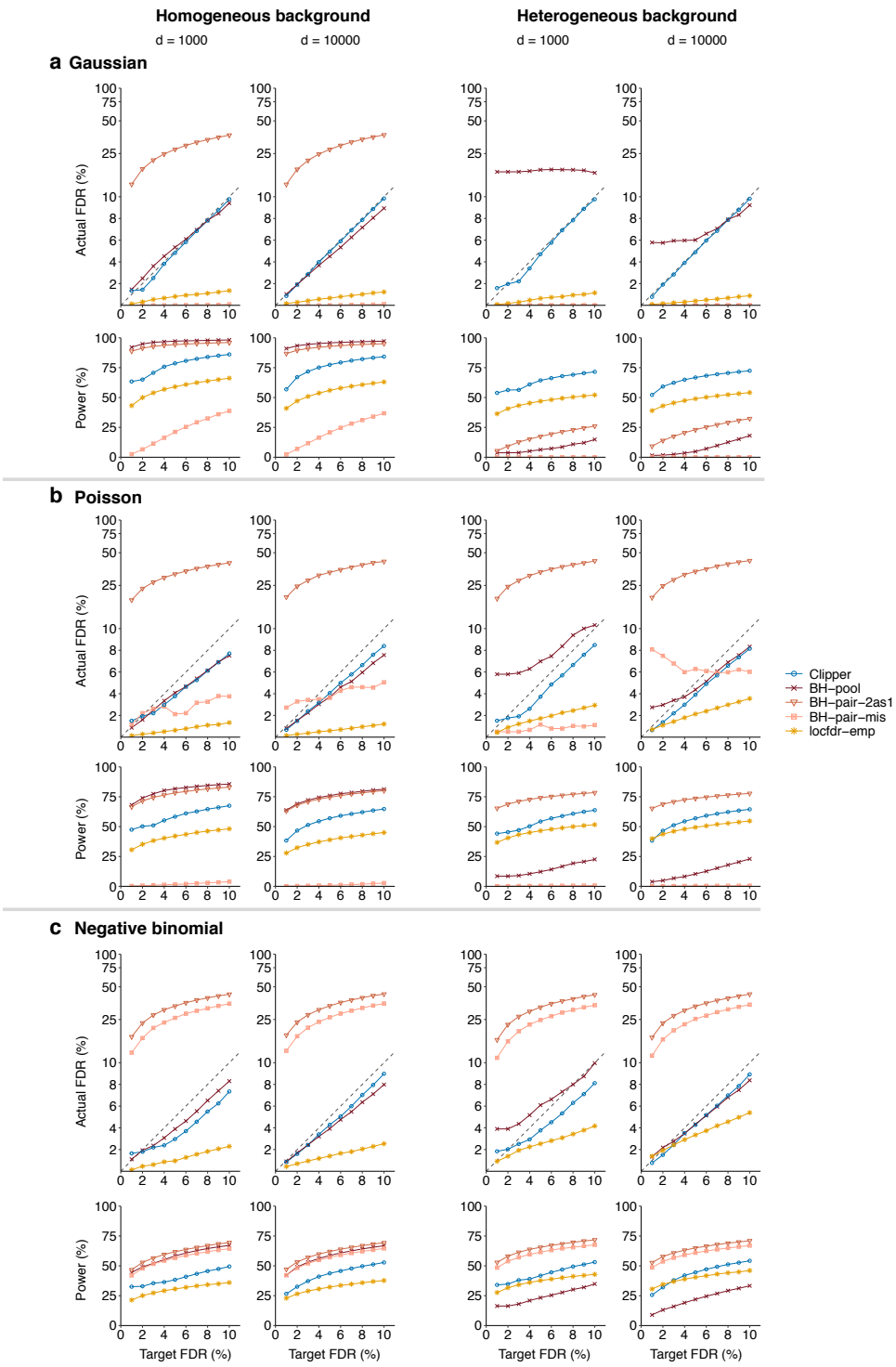


Figure S1: In 1vs1 enrichment analysis, comparison of Clipper and four other generic FDR control methods (BH-pool, BH-pair-2as1, BH-pair-mis, and locfdr-emp) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 1000$ or $10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

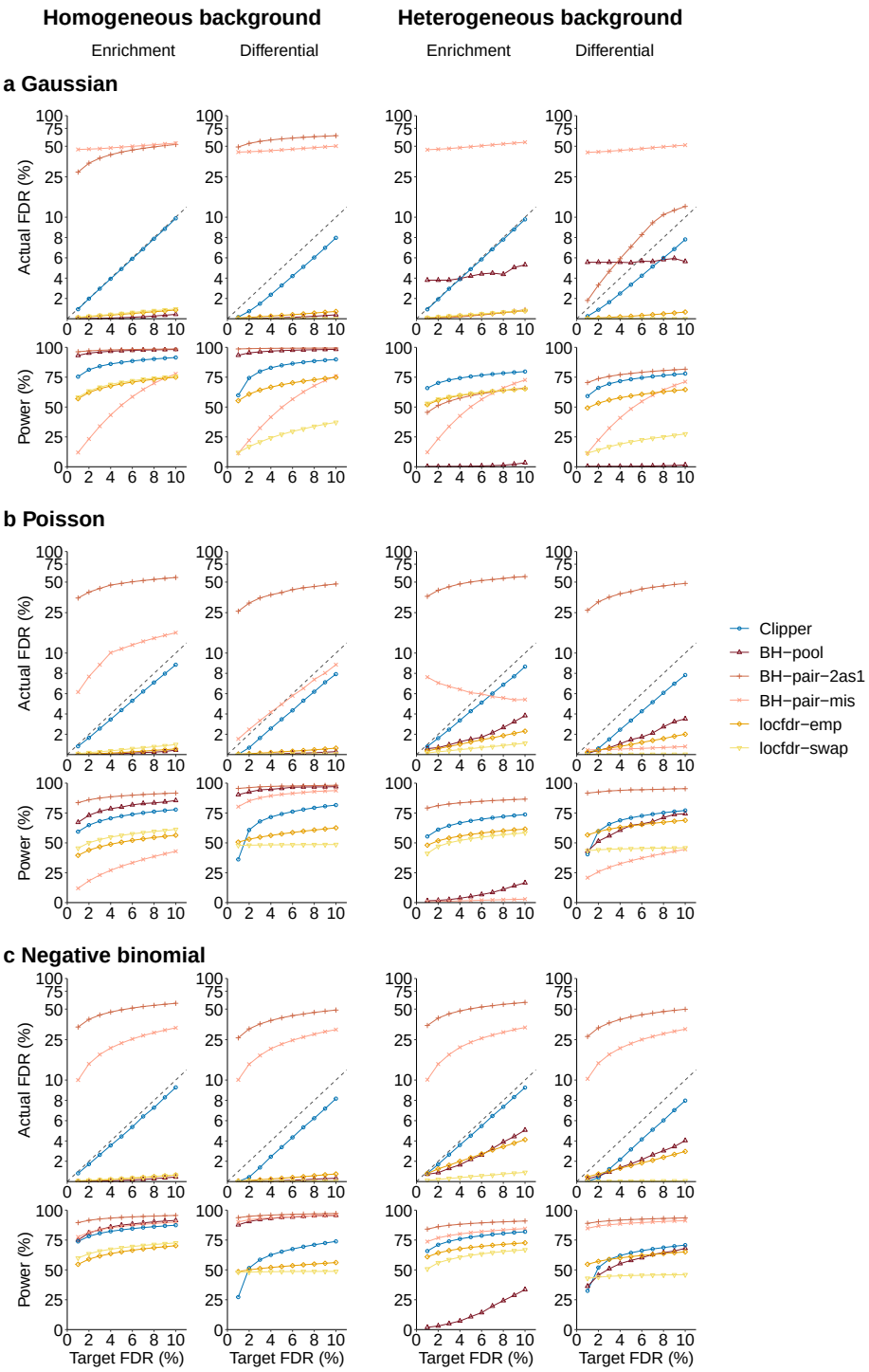


Figure S2: In 2vs1 enrichment analysis (columns 1 and 3) and differential analysis (columns 2 and 4), comparison of Clipper and five other generic FDR control methods (BH-pooled, BH-pair-2as1, BH-pair-mis, locfdr-emp, and locfdr-swap) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous(two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background (except for differential analysis with $q \leq 2\%$).

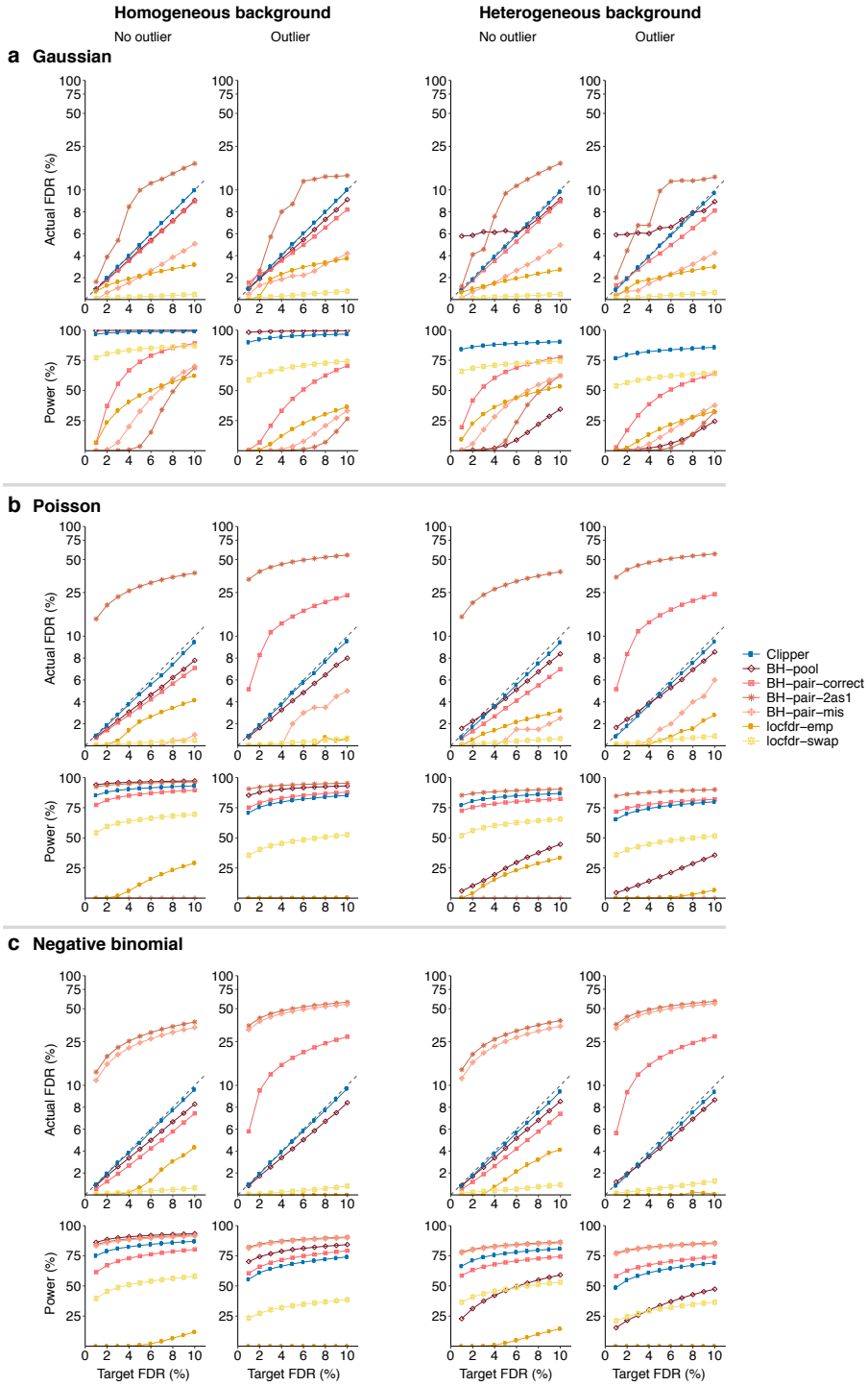


Figure S3: In 3vs3 enrichment analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of Clipper and six other generic FDR control methods (BH-pooled, BH-pair-correct, BH-pair-2as1, BH-pair-mis, locfdr-emp, and locfdr-swap)in terms of their FDR control and power in 3vs3 enrichment analysis with possible outliers. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

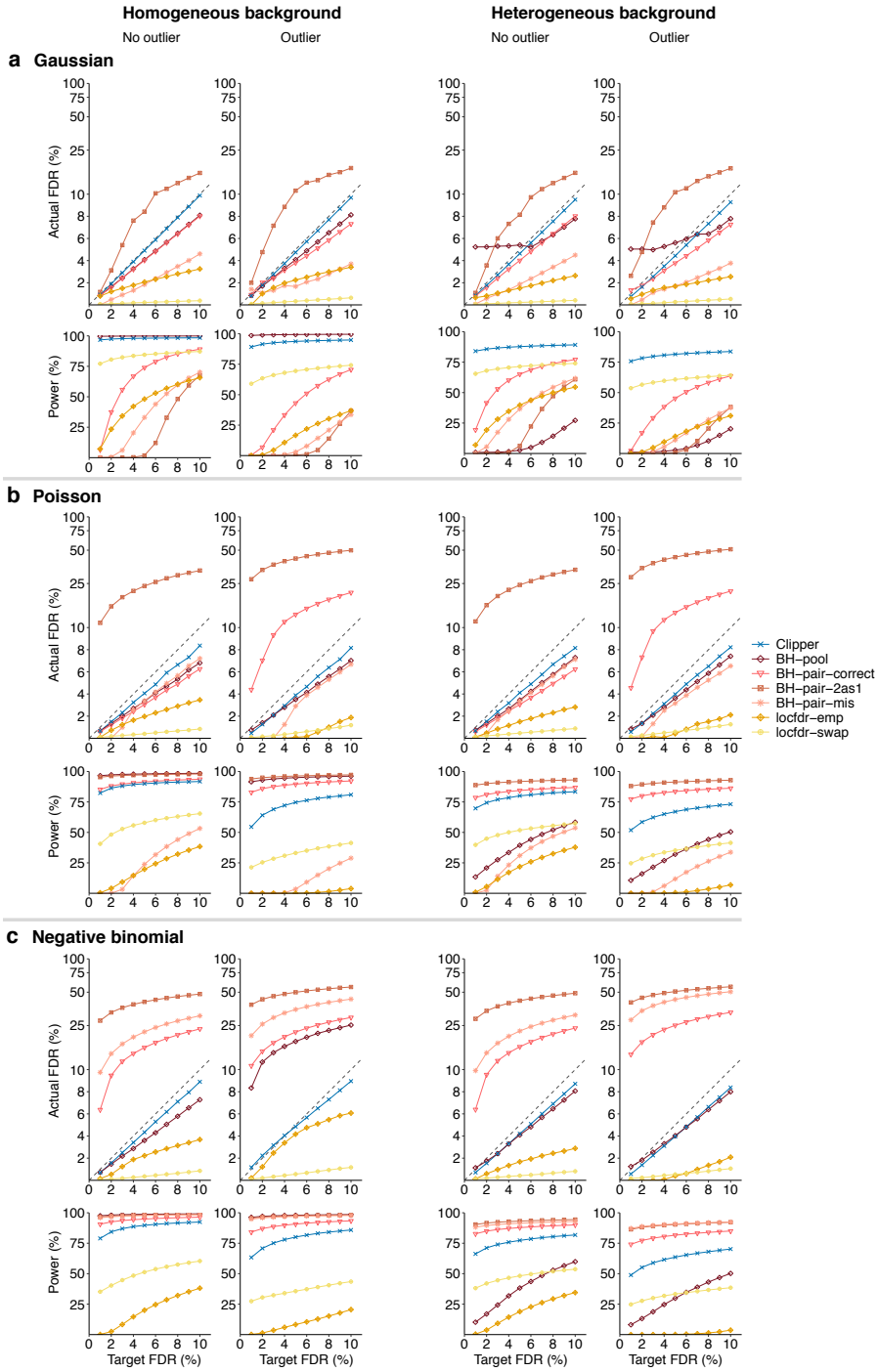


Figure S4: In 3vs3 differential analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of Clipper and six other generic FDR control methods (BH-pooled, BH-pair-correct, BH-pair-2as1, BH-pair-mis, locfdr-emp, and locfdr-swap) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background (except for Poisson distribution where Clipper is second to BH-pair-correct, an idealistic method).

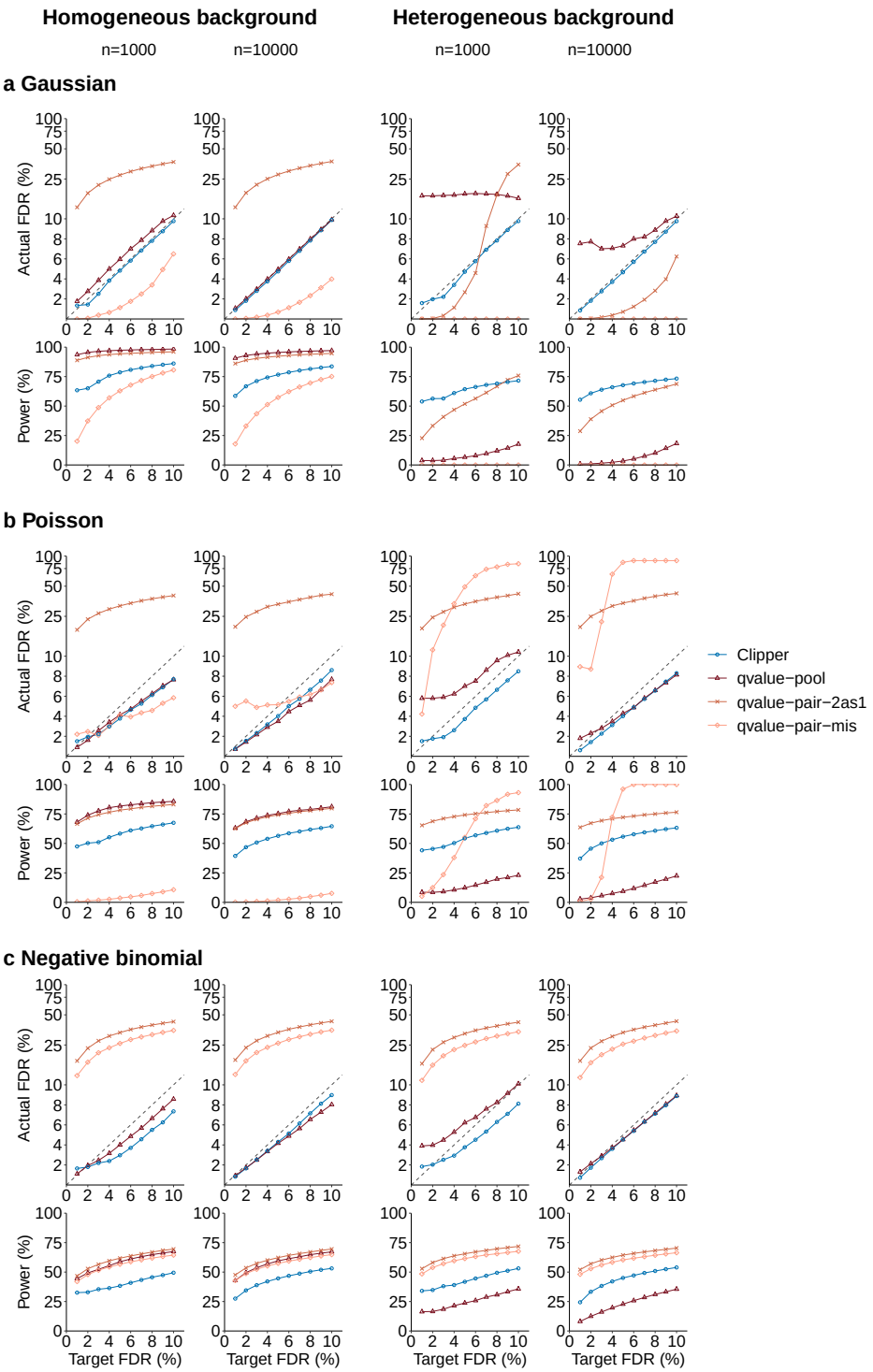


Figure S5: In 1vs1 enrichment analysis, comparison of Clipper and three other generic FDR control methods using Storey's q-value (qvalue-pool, qvalie-pair-2as1, and qvalue-pair-mis) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 1000$ or $10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

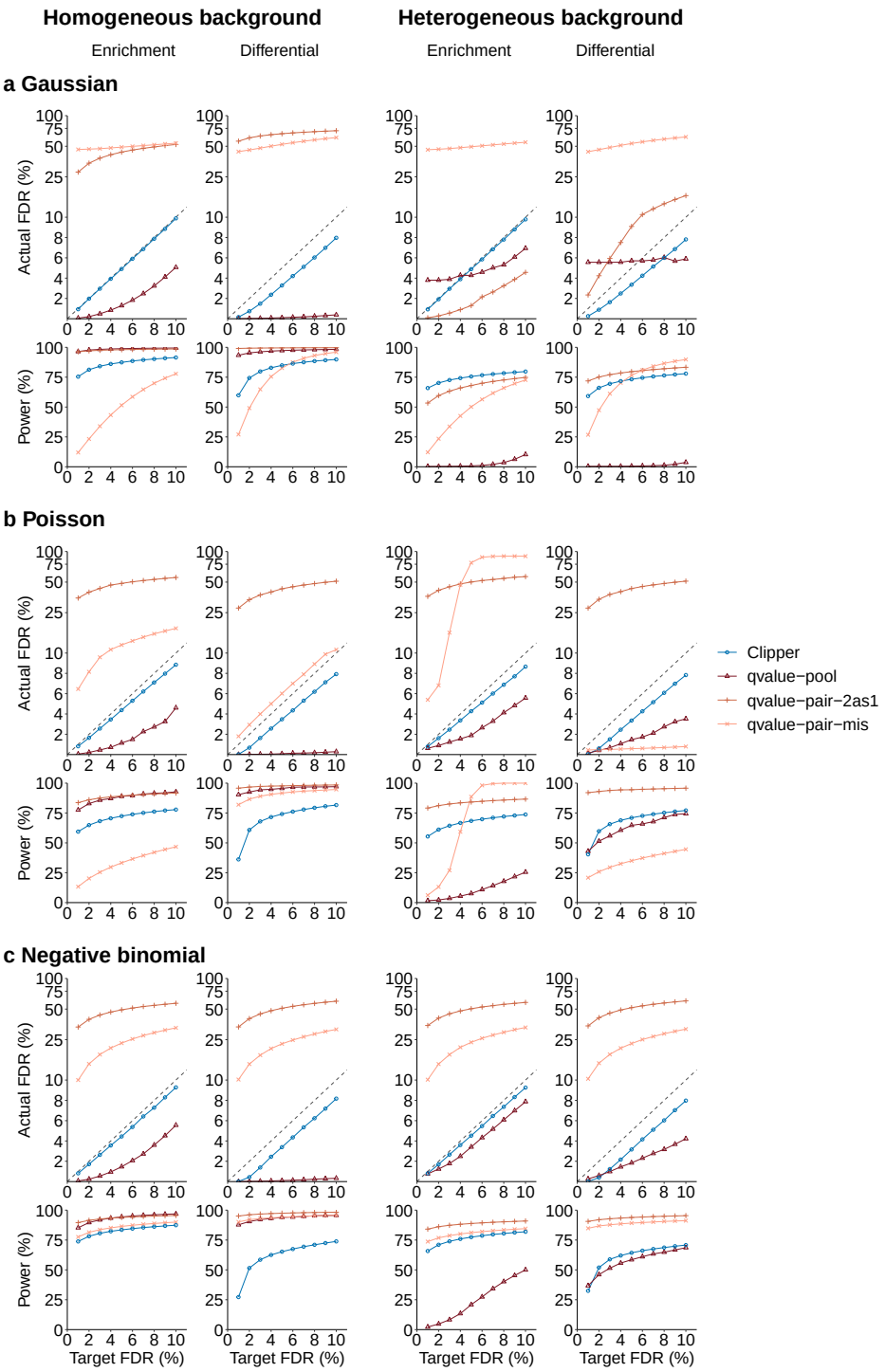


Figure S6: In 2vs1 enrichment analysis (columns 1 and 3) and differential analysis (columns 2 and 4), comparison of Clipper and three other generic FDR control methods using Storey's q-value (qvalue-pool, qvalie-pair-2as1, and qvalue-pair-mis) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

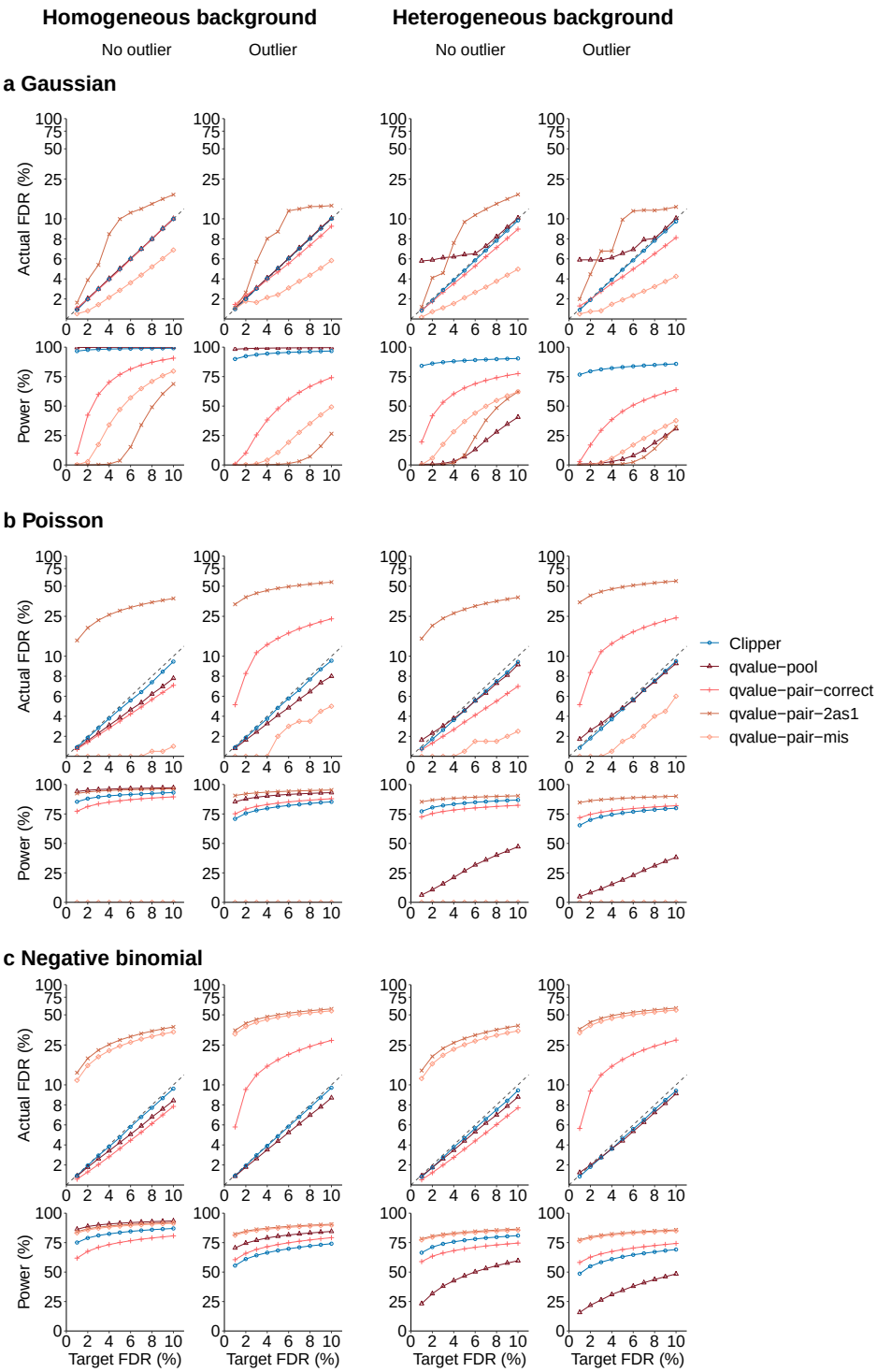


Figure S7: In 3vs3 enrichment analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of Clipper and four other generic FDR control methods using Storey's q-value (qvalue-pooled, qvalue-pair-correct, qvalue-pair-2as1, and qvalue-pair-mis) in terms of their FDR control and power in 3vs3 enrichment analysis with possible outliers. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

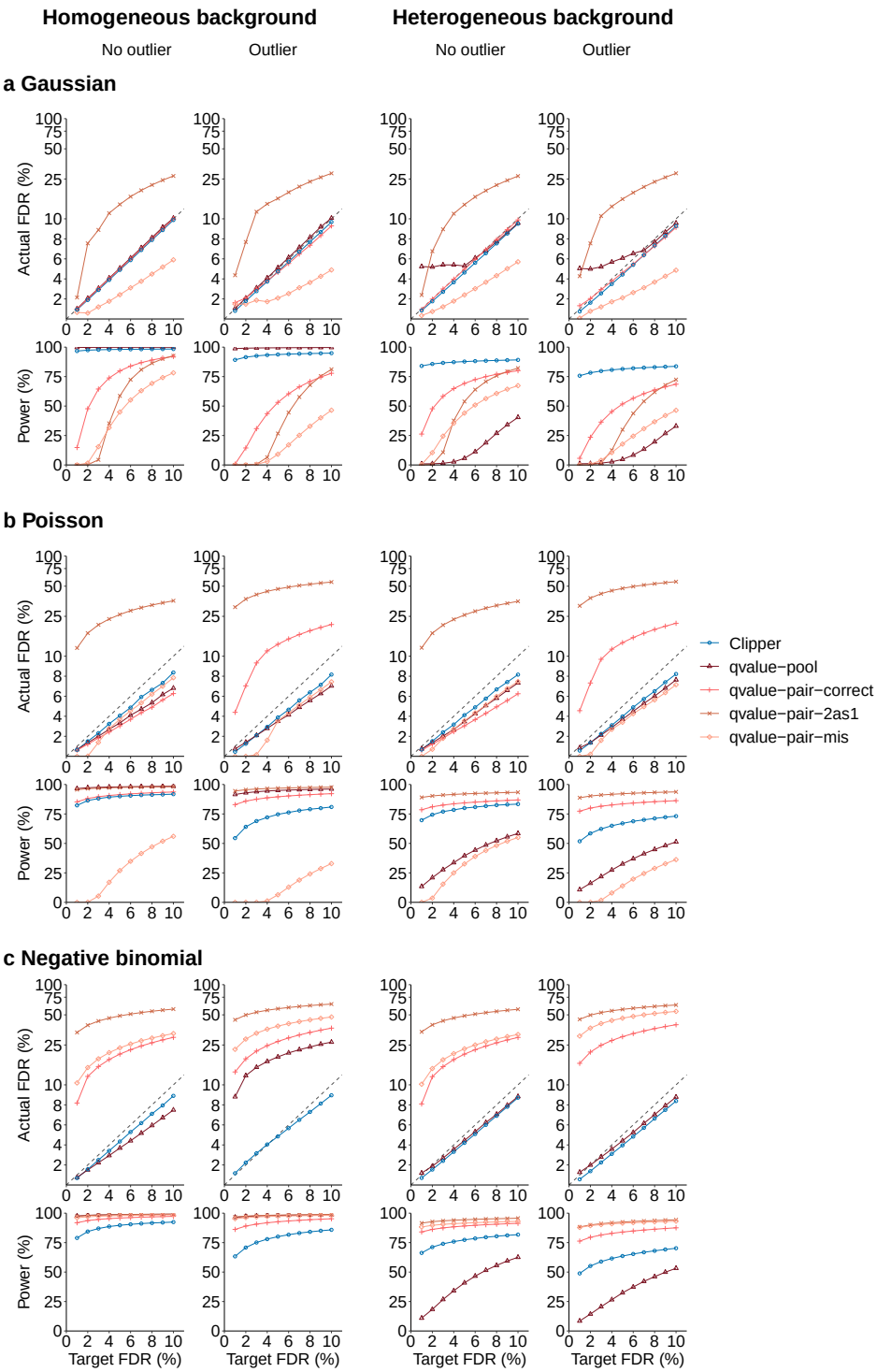


Figure S8: In 3vs3 differential analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of Clipper and four other generic FDR control methods using Storey's q-value (qvalue-pooled, qvalue-pair-correct, qvalue-pair-2as1, and qvalue-pair-mis) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background (except for Poisson distribution where Clipper is second to qvalue-pair-correct, an idealistic method).

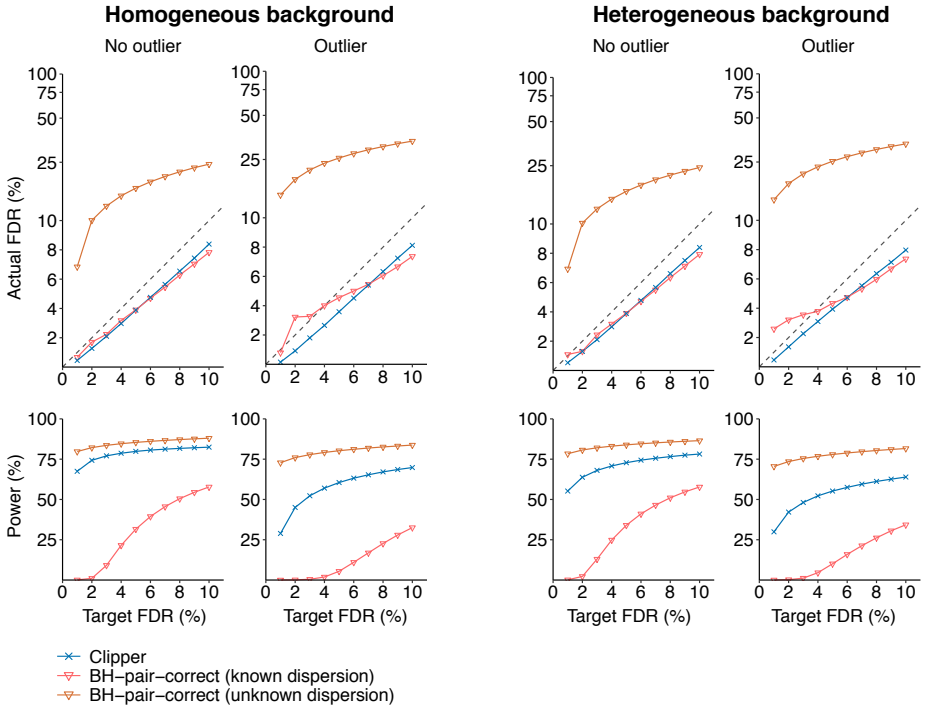


Figure S9: In 3vs3 differential analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of Clipper, BH-pair-correct (known dispersion), and BH-pair-correct (unknown dispersion) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. BH-pair-correct (unknown dispersion) cannot control the FDR in all settings. In contrast, Clipper is consistently the most powerful for homogeneous and heterogeneous background.

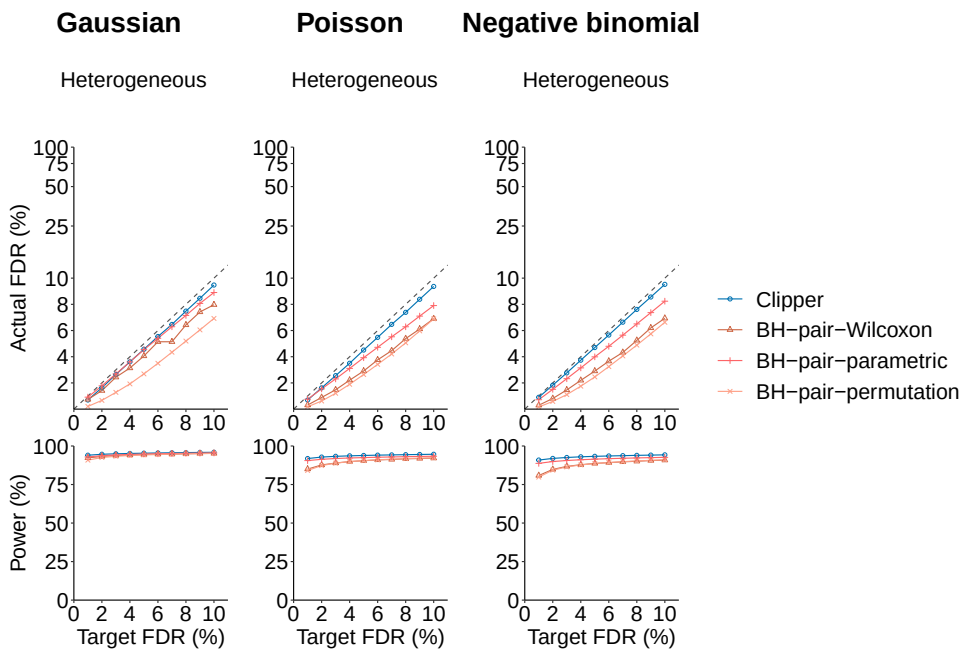


Figure S10: In 10vs10 enrichment analysis without outliers, comparison of Clipper and three generic FDR control methods (BH-pair-Wilcoxon, BH-pair-parametric, and BH-pair-permutation) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from the Gaussian distribution, the Poisson distribution, or the negative binomial distribution under heterogeneous background scenarios. Clipper achieves the highest power for all distributions.

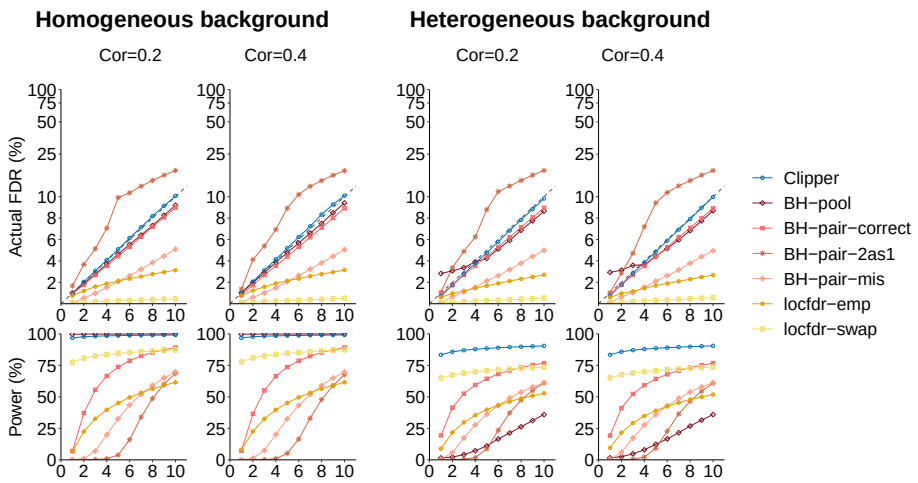


Figure S11: In 3vs3 enrichment analysis with correlated features, comparison of Clipper and six other generic FDR control methods (BH-pooled, BH-pair-correct, BH-pair-2as1, BH-pair-mis, locfdr-emp, and locfdr-swap) in terms of their FDR control and power in 3vs3 enrichment analysis. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from a multivariate Gaussian distribution with a correlation 0.2 (columns 1 and 3) or 0.4 (columns 2 and 4) between features. Among the methods that control the FDR, Clipper is the second most powerful for homogeneous background and the most powerful for heterogeneous background.

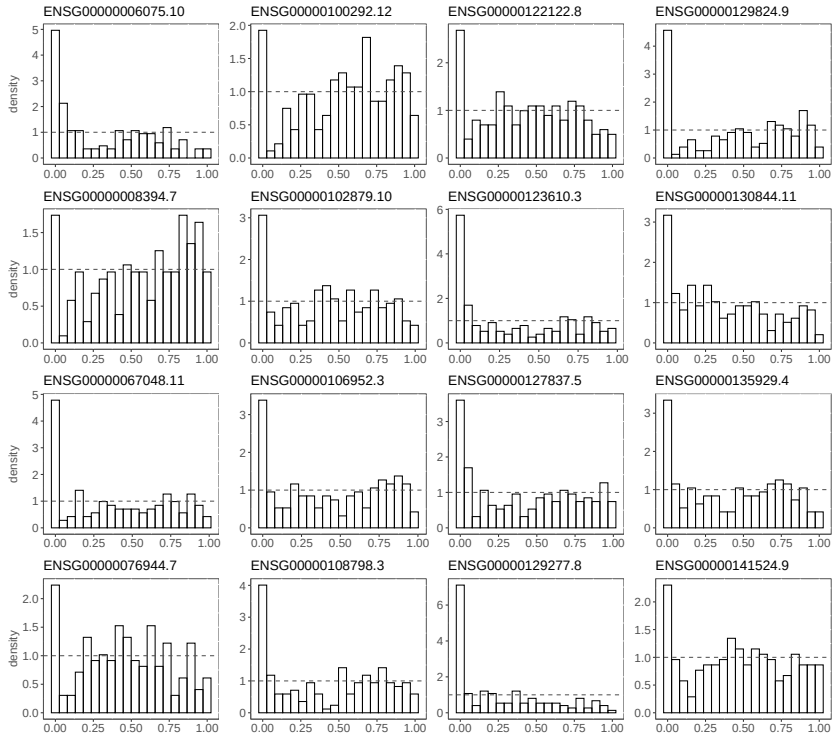


Figure S12: The p-value distributions of 16 non-DEGs that are most frequently identified by DESeq2 at $q = 5\%$ from 200 synthetic datasets. The p-values of these 16 genes tend to be overly small, and their distributions are non-uniform with a mode close to 0.

a

GO terms enriched in Clipper−specific DEGs in Clipper vs. DESeq2 comparison

GO term (ID)	qvalue (Clipper)
neutrophil activation (GO:0042119)	3.104557e−10
granulocyte activation (GO:0036230)	3.104557e−10
neutrophil degranulation (GO:0043312)	8.587750e−10
neutrophil activation involved in immune response (GO:0002283)	8.591455e−10
neutrophil mediated immunity (GO:0002446)	3.104557e−10

b

GO terms enriched in Clipper−specific DEGs in Clipper vs. edgeR comparison

GO term (ID)	qvalue (Clipper)
neutrophil degranulation (GO:0043312)	8.587750e−10
neutrophil activation involved in immune response (GO:0002283)	8.591455e−10
neutrophil activation (GO:0042119)	3.104557e−10
neutrophil mediated immunity (GO:0002446)	3.104557e−10
granulocyte activation (GO:0036230)	3.104557e−10
cellular response to chemical stress (GO:0062197)	2.157116e−03
response to oxidative stress (GO:0006979)	3.141033e−03
cellular response to oxidative stress (GO:0034599)	2.902893e−03

Figure S13: Enrichment q-values of GO terms that are found enriched in the DEGs that are uniquely identified by Clipper in pairwise comparison of **(a)** Clipper vs. edgeR and **(b)** Clipper vs. DESeq2. These GO terms are all related to immune response and thus biologically meaningful.

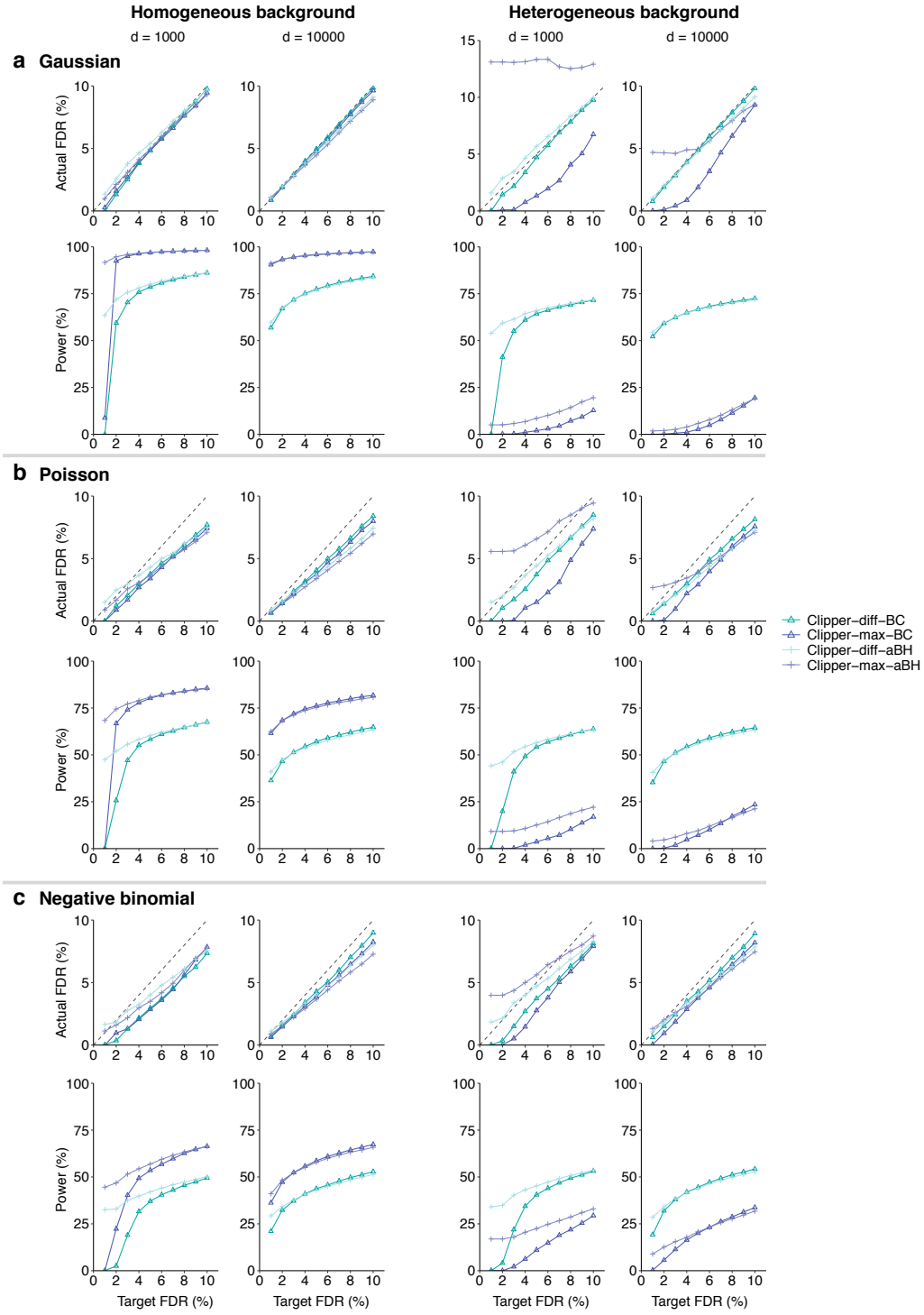


Figure S14: In 1vs1 enrichment analysis, comparison of four Clipper variant algorithms (Clipper-diff-BC, Clipper-diff-aBH, Clipper-max-BC, and Clipper-max-aBH) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 1000$ or $10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Clipper-diff-BC is chosen as the default implementation under this scenario.

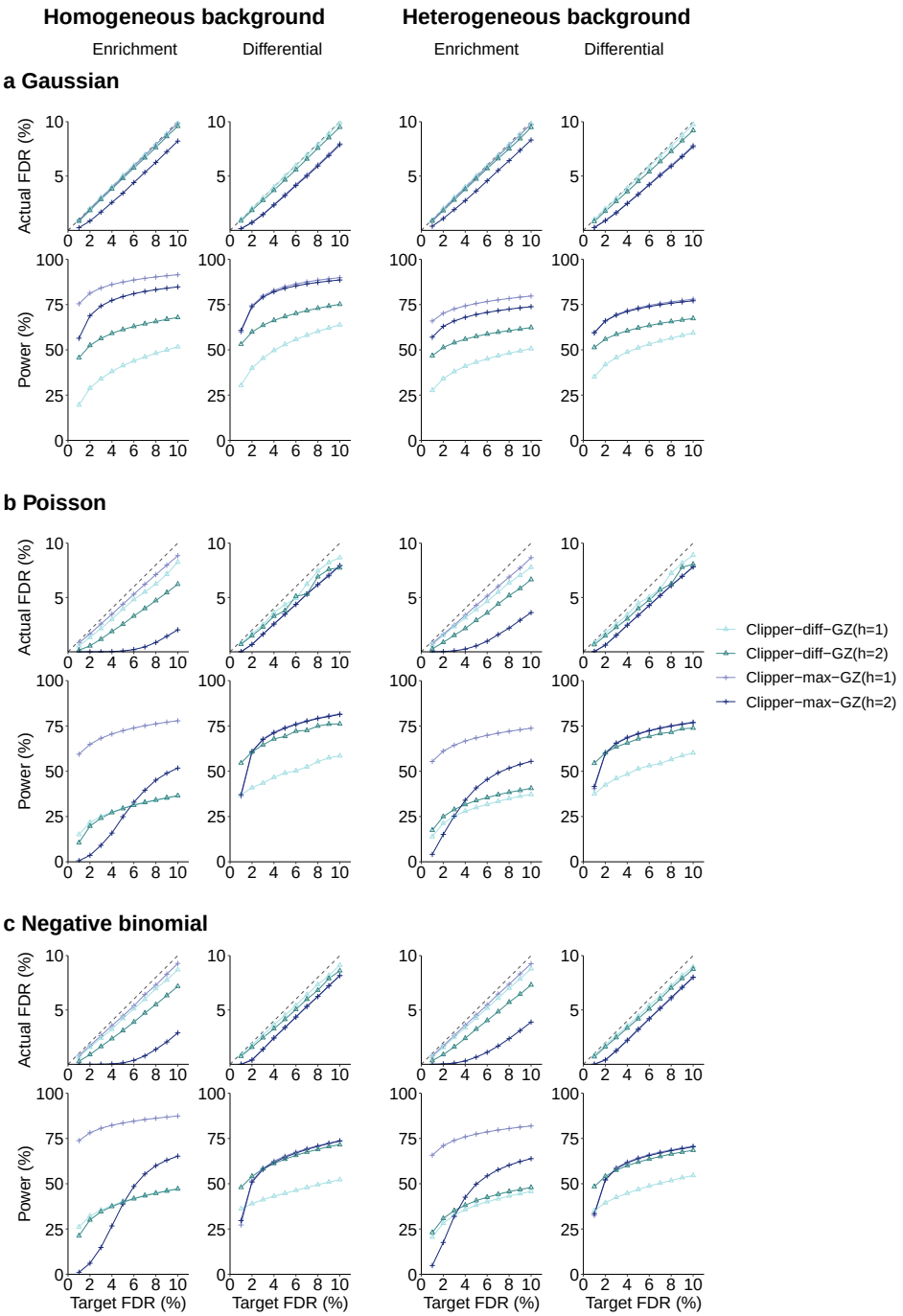


Figure S15: In 2vs1 enrichment analysis (columns 1 and 3) and differential analysis (columns 2 and 4), comparison of four Clipper variant algorithms (Clipper-diff-GZ(h=1), Clipper-diff-GZ(h=2), Clipper-max-GZ(h=1), and Clipper-max-GZ(h=2)) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Clipper-max-GZ(h=1) is chosen as the default implementation under this scenario.

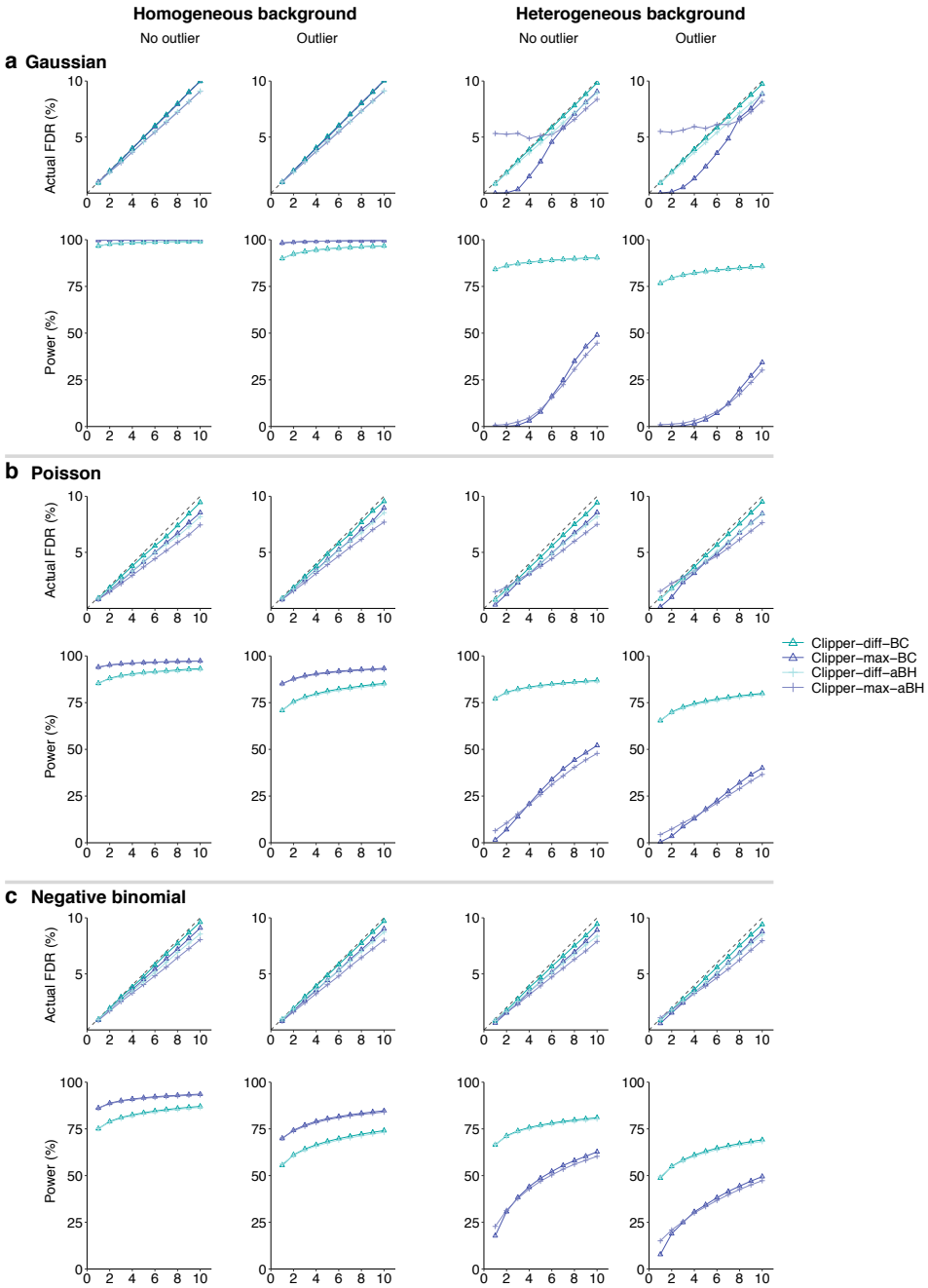


Figure S16: In 3vs3 enrichment analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of four Clipper variant algorithms (Clipper-diff-BC, Clipper-diff-aBH, Clipper-max-BC, and Clipper-max-aBH) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Clipper-diff-BC is chosen as the default implementation under this scenario.

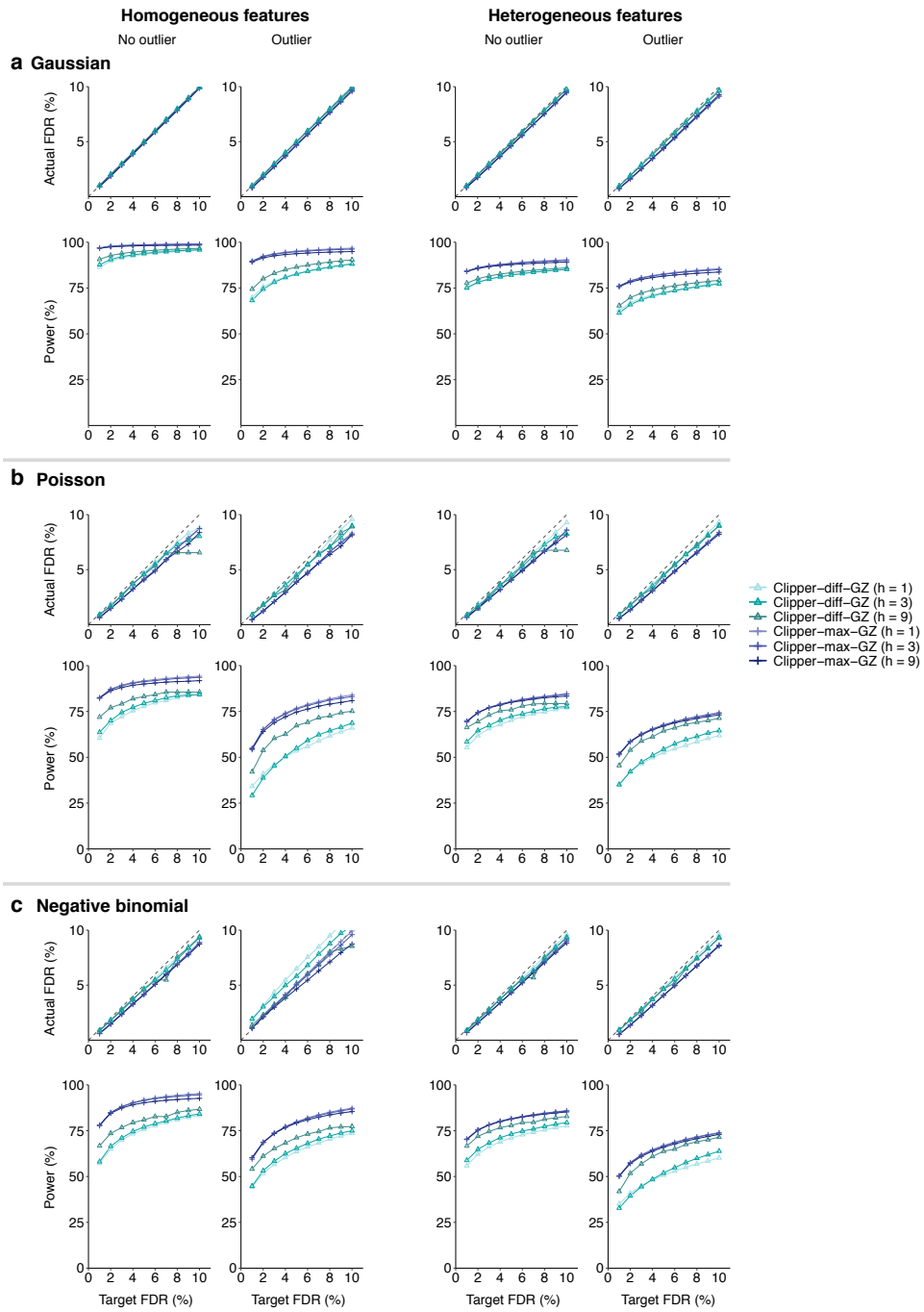


Figure S17: In 3vs3 differential analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of six Clipper variant algorithms (Clipper-diff-GZ(h=1), Clipper-diff-GZ(h=3), Clipper-diff-GZ(h=9), Clipper-max-GZ(h=1), Clipper-max-GZ(h=3), and Clipper-max-GZ(h=9)) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Clipper-max-GZ(h=1) is chosen as the default implementation under this scenario.

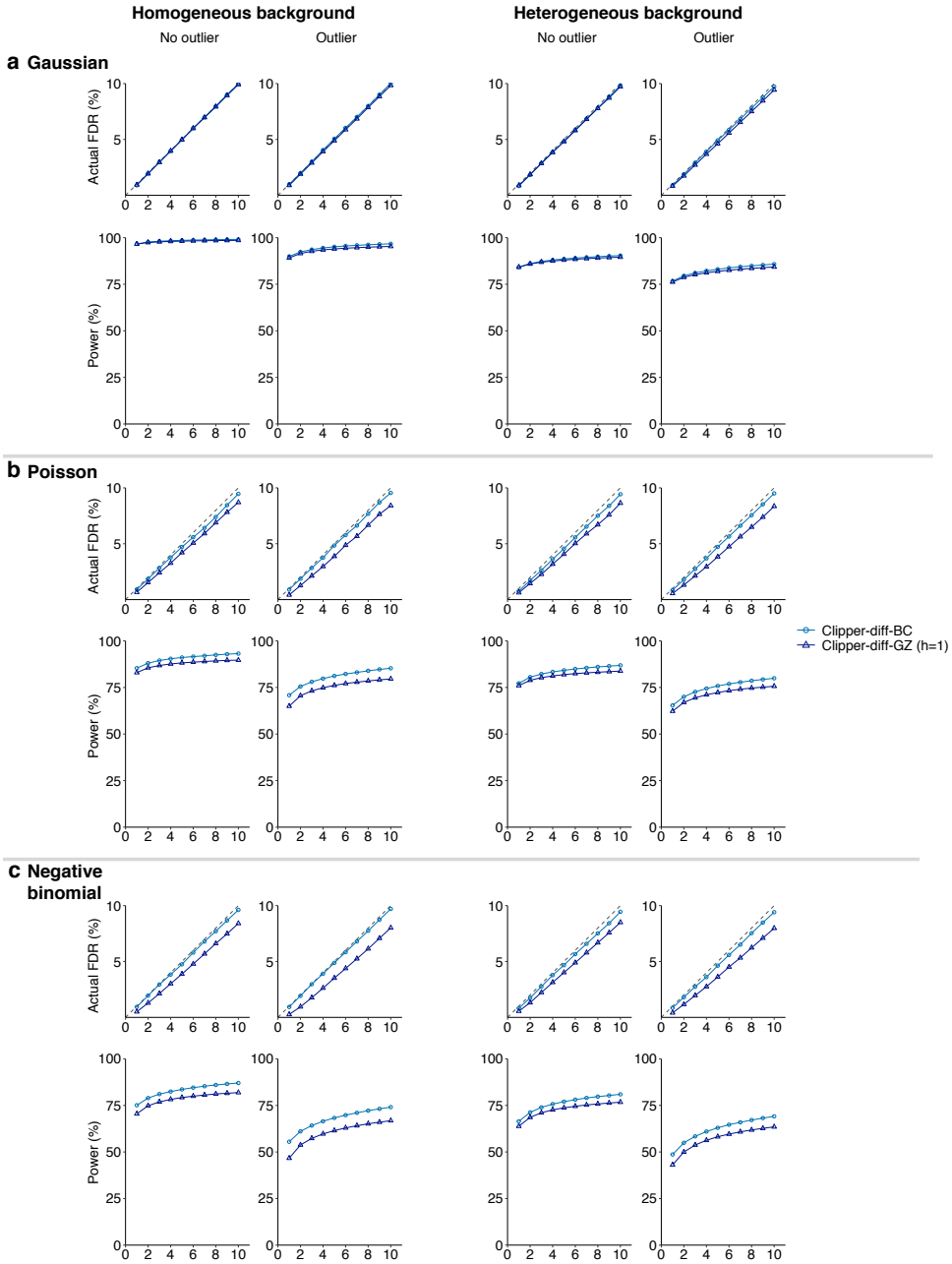


Figure S18: In 3vs3 enrichment analysis without (columns 1 and 3) or with outliers (columns 2 and 4), comparison of two Clipper variant algorithms (Clipper-diff-BC, Clipper-max-GZ(h=1)) in terms of their FDR control and power. At target FDR thresholds $q \in \{1\%, 2\%, \cdots, 10\%\}$, each method's actual FDRs and power are evaluated on 200 simulated datasets with $d = 10{,}000$ features generated from (**a**) the Gaussian distribution, (**b**) the Poisson distribution, or (**c**) the negative binomial distribution under homogeneous (two left columns) and heterogeneous (two right columns) background scenarios. Clipper-diff-BC is chosen as the default implementation under this scenario.

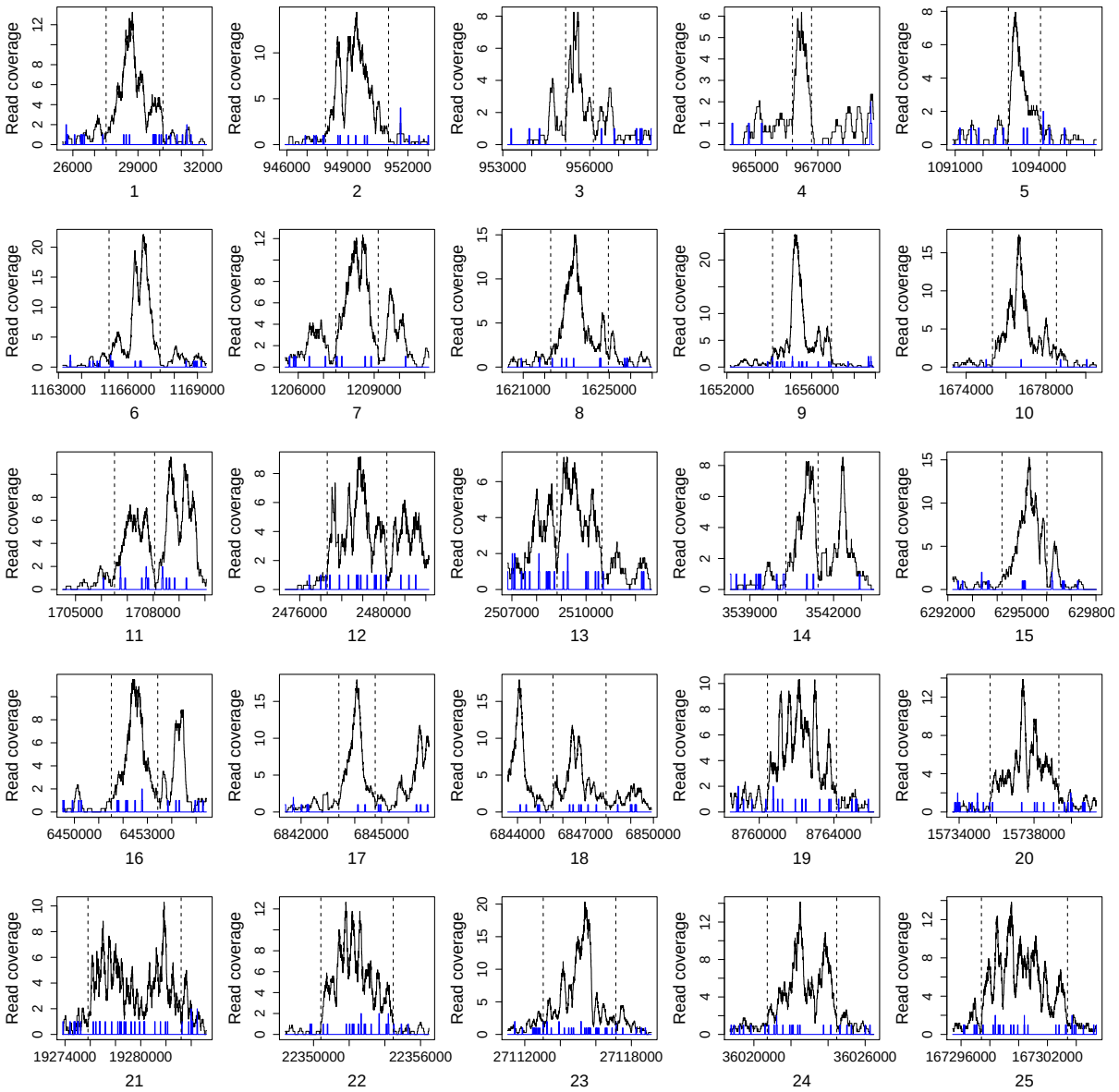


Figure S19: 25 "true peaks" from H3K4me3 ChIP-seq data of cell line GM12878. Black and blue curves indicate the read coverages in the experimental and control samples, respectively. Vertical dashed lines indicate the peak boundaries.

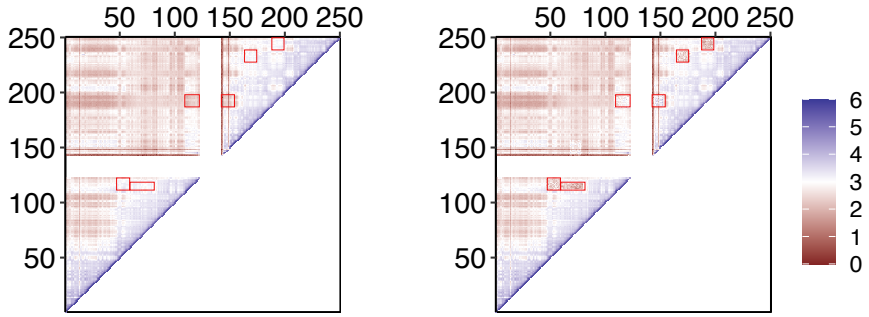


Figure S20: $\log_{10}$-transformed mean Hi-C interaction matrices (μ_X and μ_Y in Section S6.4) under the two conditions. DIR regions are highlighted in red squares.

References

[1] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[2] John D Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[3] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[4] Bradley Efron. “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis”. In: *Journal of the American Statistical Association* 99.465 (2004), pp. 96–104.

[5] Bradley Efron et al. “Empirical Bayes analysis of a microarray experiment”. In: *Journal of the American statistical association* 96.456 (2001), pp. 1151–1160.

[6] Abhishek K Sarkar and Matthew Stephens. “Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis”. In: *BioRxiv* (2020).

[7] Ery Arias-Castro and Shiyun Chen. “Distribution-free multiple testing”. In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1983–2001.

[8] Rina Foygel Barber, Emmanuel J Candès, et al. “A knockoff filter for high-dimensional selective inference”. In: *The Annals of Statistics* 47.5 (2019), pp. 2504–2537.

[9] Jaime Roquero Gimenez and James Zou. “Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization”. In: *arXiv preprint arXiv:1810.11378* (2018).

[10] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), pp. 1–9.

[11] Sven Heinz et al. “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities”. In: *Molecular cell* 38.4 (2010), pp. 576–589.

[12] Marina Spivak et al. “Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets”. In: *Journal of proteome research* 8.7 (2009), pp. 3737–3745.

[13] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[14] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), p. 550.

[15] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. “multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments”. In: *Bioinformatics* 35.17 (2019), pp. 2916–2923.

[16] John C Stansfield et al. “HiCcompare: an R-package for joint normalization and comparison of HI-C datasets”. In: *BMC bioinformatics* 19.1 (2018), p. 279.

[17] Aaron TL Lun and Gordon K Smyth. “diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data”. In: *BMC bioinformatics* 16.1 (2015), pp. 1–11.

[18] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. “FIND: difFerential chromatin INteractions Detection using a spatial Poisson process”. In: *Genome research* 28.3 (2018), pp. 412–422.

[19] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[20] Claire R Williams et al. “Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq”. In: *BMC bioinformatics* 18.1 (2017), p. 38.

[21] Mark D Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. In: *Genome biology* 11.3 (2010), pp. 1–9.