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], Ning Wang[2], 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 features such as 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), which is primarily controlled 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. Clipper is a general statistical framework for FDR control without relying on p-values or specific data distributions. Clipper outperforms existing methods for a broad range of applications in high-throughput data analysis.

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., 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. Technically, under the frequentist statistical paradigm, the FDR is defined as the expected proportion of uninteresting features among the discoveries. In parallel, under the Bayesian statistical 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 FDR control 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] and/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 the difficulty of valid p-value calculation 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 a theoretical guarantee for FDR control. (Although later works have studied some theoretical properties of SAM, they are not about the exact control of the FDR [33, 34].) 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 using 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 Methods and Additional File 1: Section S2). 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 Methods and Additional File 1: Section S2).

To verify Clipper's performance, we designed comprehensive simulation studies to benchmark Clipper against existing generic FDR control methods (Additional File 1: 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 bulk and single-cell RNA-seq data, and DIR identification from Hi-C data. Notably, our benchmarking results for peptide identification are based on our in-house data, 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 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. By simulation setting, we refer to a combination of an experimental design, a distributional family, and a background scenario. The details of simulation settings are described in Additional File 1: 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 cal-

culation approaches, respectively. See Methods for detail.

The comparison results are in Fig. 2 and Additional File 1: Figs. S1–S14. A good FDR control method should have its actual FDR no larger than the target FDR threshold and achieve high power. The results show that Clipper controls the FDR and is overall more powerful than the 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 the 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 Additional File 1: 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 both 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 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 Additional File 1: Figs. S1–S8). These results confirm that 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 Methods)—fail to control the FDR in the existence of outliers (Fig. 2e and Additional File 1: Figs. S3 and S7) or for the negative binomial distribution with unknown dispersion (Fig. 2f and Additional File 1: 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 Additional File 1: Figs. S4 and S8).

Fourth, the two local-fdr-based methods—locfdr-emp and locfdr-swap—achieve the FDR control for all designs and analyses; however, they are less powerful than Clipper in most cases (Additional File 1: 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 (Additional File 1: Fig. S10).

Moreover, the above five phenomena are consistently observed across the three distributions (Gaussian, Poission, and negative binomial) that we have examined, further confirming the robustness of Clipper.

In addition, for the 3vs3 enrichment analysis, we also varied the proportion of interesting features as 10%, 20%, and 40%. The comparison results in Additional File 1: Fig. S3 (columns 1 and 3 for 10%) and Additional File 1: Fig. S11 (for 20% and 40%) show that the performance of Clipper is robust to the proportion of interesting features.

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 the other generic FDR control methods. The comparison results in Additional File 1: Fig. S12 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 bulk or single-cell 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 an indispensable analysis is to identify genomic regions with enriched sequence reads in ChIP-seq data. These regions are likely 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 HOMER [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 (with 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 (Additional File 1: Section S5.1), we used ENCODE ChIP-seq data of cell line GM12878 [44] and ChiPulate [45], a ChIP-seq data simulator, to generate semi-synthetic data with spiked-in peaks (Additional File 1: 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 (Additional File 1: 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 [46]. Besides, MACS2 uses one-sample tests to compute p-values when two-sample tests should be performed. As a result, the p-values of MACS2 and HOMER are questionable, so using their p-values for FDR control has no guaranteed 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 em-

phasize 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.

As a side note, it is known that uninteresting regions tend to have larger read counts in the control sample than in the experimental (ChIP) sample; however, this phenomenon does not violate Clipper's theoretical assumption for FDR control (Additional File 1: Section S7.1).

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 [47]. 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, 48, 49]. 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 target or decoy PSM. Discoveries are the target PSMs whose q-values are no greater than the target FDR threshold q.

We generated the first comprehensive benchmark dataset from an archaea species *Pyrococcus furiosus*, and we used it to examine the FDR control and power of a popular database search algorithm SEQUEST [3] (Additional File 1: Section S5.2). Using this benchmark dataset (Additional File 1: Section S6.2), we demonstrate that, as an add-on, Clipper improves the power of SEQUEST. Specifically, Clipper treats mass spectra as features. For each mass spectrum, Clipper considers the experimental (or background) measurement as the $-\log_{10}$-transformed SEQUEST q-value of the target (or decoy) PSM that includes the mass spectrum. Then Clipper decides which mass spectra and their corresponding target PSMs are discoveries (Additional File 1: 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 SEQUEST 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 SEQUEST and SEQUEST+Clipper both control the FDR, SEQUEST+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 bulk 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: edgeR [4] and DESeq2 [5] (Additional File 1: Section S5.3). Specifically, when we implemented Clipper, we first performed the trimmed mean of M values (TMM) normalization [50] to correct for batch effects; then we treated genes as features and their normalized expression levels as measurements under two conditions (Additional File 1: Section S7.3). We also implemented two versions of DESeq2 and edgeR: with or without IHW, a popular procedure for boosting the power of p-value-based FDR control methods by incorporating feature covariates [27]. In our implementation of the two versions of DESeq2 and edgeR, we used their standard pipelines, including normalization, model fitting, and gene filtering (edgeR only). To verify the FDR control, we generated four realistic semi-synthetic datasets from two real RNA-seq datasets—one from classical and non-classical human monocytes [51] and the other from yeasts with or without *snf2* knockout [52]—using simulation strategies 1 and 2 (Additional File 1: Section S6.3).

In detail, in simulation strategy 1, we used bulk RNA-seq samples from two conditions to compute a fold change for every gene between the two conditions; then we defined true DEGs as the genes whose fold changes exceeded a threshold; next, we randomly drew three RNA-seq samples and treated them as replicates from each condition ($m = n = 3$ as in Methods); using those subsampled replicates of two conditions, we preserved the true DEGs' read counts and permuted the read counts of the true non-DEGs, i.e., the genes other than true DEGs, between conditions. In summary, simulation strategy 1 guarantees that the measurements of true non-DEGs are i.i.d., an assumption that Clipper relies on for theoretical FDR control.

In simulation strategy 2, which we borrowed from a benchmark study [53], we first randomly selected at most 30% genes as true DEGs; next, we randomly drew six RNA-seq samples from one condition (classical human monocytes and yeasts without knockout) and split the samples into two "synthetic conditions," each with three replicates ($m = n = 3$ as in Methods); then for each true DEG, we multiplied its read counts under one of the two synthetic conditions (randomly and independently picked for each gene) by a randomly generated fold change (see Additional File 1: Section S6.3); finally, for the true non-DEGs, we preserved their read counts in the six samples. In summary, simulation strategy 2 preserves batch effects, if existent in real data, for the true non-DEGs (the majority of genes). As a result, the semi-synthetic data generated under strategy 2 may violate the Clipper assumption for theoretical FDR control and thus can help evaluate the robustness of Clipper on real data.

The four semi-synthetic datasets have ground truths (true DEGs and non-DEGs) to evaluate each DEG identification method's FDR and power for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Our results in Fig. 4a and Additional File 1: Figs. S15a–S17a show that Clipper consistently controls the FDR and achieves high power on all four semi-synthetic datasets. In contrast, DESeq2 and edgeR cannot consistently control the FDR except for the yeast semi-synthetic dataset generated under simulation strategy 2. Given the fact that DESeq2 and edgeR do not consistently perform well on the three other semi-synthetic datasets, we hypothesize that their parametric distributional assumptions, if violated on real data, hinder valid FDR control, in line with our motivation for developing Clipper. By examining whether true non-DEGs' p-values calculated by DESeq2 or edgeR follow the theoretical Uniform$[0, 1]$ distribution, we find that the answer is no for many non-DEGs, as indicated by the small p-values (one per non-DEG) of uniformity tests (Additional File 1: Fig. S18); this issue is more serious for DESeq2, consistent with the worse FDR control of DESeq2 (Fig. 4a and Additional File 1: Figs. S15a–S17a). Furthermore, we observe that adding IHW to edgeR and DESeq2 has negligible effects on the four

semi-synthetic datasets.

To further explain why DESeq2 fails to control the FDR, we examined the p-value distributions of 16 non-DEGs that were most frequently identified (from the 100 semi-synthetic datasets generated from the human monocyte dataset using simulation strategy 1) by DESeq2 at the target FDR threshold $q = 0.05$. Our results in Additional File 1: Fig. S19 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 as discoveries by DESeq2.

In addition, we compared the DEG ranking by Clipper, edgeR, and DESeq2 in two ways. First, for true DEGs, we compared their ranking by each method with their true ranking based on true expression fold changes (from large to small, as in semi-synthetic data generation in Additional File 1: Section S6.3). Specifically, we ranked true DEGs using Clipper's contrast scores (from large to small), edgeR's p-values (from small to large), or DESeq2's p-values (from small to large). Our results in Fig. 4b and Additional File 1: Figs. S15b–S17b show that Clipper's contrast scores exhibit the most consistent ranking with the ranking based on true fold changes. Second, to compare the power of Clipper, edgeR, and DESeq2 based on their DEG rankings instead of nominal p-values, we calculated their power under the actual FDRs, which only depend on gene rankings (for the definition of actual FDR, see Additional File 1: Section S6.3). Fig. 4a and Additional File 1: Figs. S15a–S17a show that, when Clipper, edgeR, and DESeq2 have the same actual FDR, Clipper consistently outperforms edgeR and DESeq2 in terms of power, i.e., Clipper has the largest number of true DEGs in its top ranked genes.

We also compared the reproducibility of Clipper, edgeR, and DESeq2 in the presence of sampling randomness. Specifically, we used two semi-synthetic datasets (generated independently from the same procedure in Additional File 1: Section S6.3) as technical replicates and computed Clipper's contrast scores, edgeR's p-values, and DESeq's p-values on each dataset. For each method, we evaluated its reproducibility between the two semi-synthetic datasets by computing three criteria—the irreproducibility discovery rate (IDR) [54], the Pearson correlation, and the Spearman correlation—using its contrast scores or $-\log_{10}$-transformed p-values. Fig. 4c and Additional File 1: Figs. S15c–S17c show that Clipper's contrast scores have higher reproducibility by all three criteria compared to edgeR's and DESeq2's p-values.

Finally, we compared Clipper with DESeq2 and edgeR on the real RNA-seq data of classical and non-classical human monocytes [51]. 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. 5a 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 Clipper finds biologically meaningful DEGs missed by DESeq2 or edgeR, 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` [55]. Fig. 5b ("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. 5c 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 Clipper's identified DEGs than in edgeR and DESeq2's identified DEGs. 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. 5b and Additional File 1: Fig. S20 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. In addition, we examined the DEGs that were identified by Clipper only but missed by both edgeR and DESeq2. Fig. 5d and Additional File 2

show that these genes include multiple key immune-related genes, including *CD36*, *DUSP2*, and *TNFAIP3*. We further performed pathway analysis on these genes and the DEGs that were identified by DEseq2 only but missed by both edgeR and Clipper, using the R package `limma` [10]. Additional File 1: Fig. S21a shows that the DEGs that were only identified by Clipper have significant enrichment for immune-related pathways including phagosome, a key function of monocytes and macrophages. On the contrary, Additional File 1: Fig. S21b shows that fewer immune-related pathways are enriched in DEGs that were only identified by DESeq2. 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.

DEG identification from single-cell RNA-seq data (differential analysis II)

Single-cell RNA sequencing (scRNA-seq) technologies have revolutionized biomedical sciences by enabling genome-wide profiling of gene expression levels at an unprecedented single-cell resolution. DEG analysis is widely applied to scRNA-seq data for discovering genes whose expression levels change between two conditions or between two cell types. Compared with bulk RNA-seq data, scRNA-seq data have many more "replicates" (i.e., cells, whose number is often in hundreds) under each condition or within each cell type.

We compared Clipper (Additional File 1: Section S7.4) with edgeR [4], MAST [56], Monocle3 [57], the two-sample t test, and the Wilcoxon rank-sum test (Additional File 1: Section S5.4), five methods that are either popular or reported to have comparatively top performance from a previous benchmark study [58]. To verify the FDR control, we used scDesign2, a flexible probabilistic simulator, to generate scRNA-seq count data with known true DEGs [59]. scDesign2 offers three key advantages that enable the generation of realistic semi-synthetic scRNA-seq count data: (1) it captures distinct marginal distributions of different genes; (2) it preserves gene-gene correlations; (3) it adapts to various scRNA-seq protocols. Using scDesign2, we generated two semi-synthetic scRNA-seq datasets from two real scRNA-seq datasets of peripheral blood mononuclear cells (PBMCs) [60]: one using 10x Genomics [61] and the other using Drop-seq [62]. Each semi-synthetic dataset contains two cell types, CD4+ T cells and cytotoxic T cells, which we treated as two conditions (Additional File 1: Section S6.4). Having true DEGs known, the semi-synthetic datasets allow us to evaluate the FDRs and power of Clipper and the other five methods for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3c and Additional File 1: Fig. S22 show that on both 10x Genomics and Drop-seq semi-synthetic datasets, Clipper consistently controls the FDR and remains the most powerful (except for $q = 1\%$ and 2%) among all the methods that achieve FDR control. These results demonstrate Clipper's robust performance in scRNA-seq DEG analysis.

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

Hi-C experiments are widely used to investigate spatial organizations of chromosomes and to 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 contact 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] (Additional File 1: Section S5.5). Specifically, we applied Clipper to DIR identification by treating pairs of genomic regions as features and their contact intensities as measurements. To verify

the FDR control of Clipper (Additional File 1: Section S7.5), diffHic, FIND, and multiHiCcompare, we generated realistic semi-synthetic data from real interaction matrices of ENCODE cell line GM12878 [44] with true spiked-in DIRs to evaluate the FDR and power (Additional File 1: Section S6.5). 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 $q = 1\%$ and 2%, 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 provides a powerful add-on to existing bioinformatics tools to improve the reliability of FDR control and thus the reproducibility of scientific discoveries.

Clipper's FDR control procedures (BC and GZ procedures in Methods) are motivated by the Barber-Candès (BC)'s knockoff paper [35] and the Gimenez-Zou (GZ)'s multiple knockoff paper [63], but we do not need to construct knockoffs in enrichment analysis when two conditions have the same number of replicates; the reason is that the replicates under the background condition serve as natural negative controls. For differential analysis and enrichment analysis with unequal numbers of replicates, in order to guarantee the theoretical assumptions for FDR control, Clipper uses permutations instead of the complicated knockoff construction because Clipper only examines features marginally and does not concern about features' joint distribution.

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, which are widely used to validate newly developed computational frameworks [64]. 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 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 vari-

ances between two conditions. Regarding the contrast score, Clipper currently makes careful choices between two contrast scores, minus and maximum, based on the number of replicates and the analysis task (enrichment or differential).

Notably, Clipper achieves FDR control and high power using those two simple contrast scores, which are calculated for individual features without borrowing information from other features. However, Clipper does leverage the power of multiple testing by setting a contrast score threshold based on all features' contrast scores. This is a likely reason why Clipper achieves good power even with simple contrast scores. An advantage of Clipper is that it allows other definitions of contrast scores, such as the two-sample t statistic that considers within-condition variances. Empirical evidence (Additional File 1: Figs. S13 and S14) shows that the Clipper variant using the two-sample t statistic is underpowered by the default Clipper, which uses the minus summary statistic (difference of two conditions' sample means) as the contrast score in the 3vs3 enrichment analysis or as the degree of interestingness in the 3vs3 differential analysis (see Methods). Here is our current interpretation of this seemingly counter-intuitive result.

- First, both the minus statistic and the t statistic satisfy Clipper's theoretical conditions (Lemmas 1 and 3 in Additional File 1: Section S2), which guarantee the FDR control by the BC and GZ procedures; this is confirmed in Additional File 1: Figs. S13 and S14. Hence, from the FDR control perspective, Clipper does not require the adjustment for within-condition variances by using a t statistic.

- Second, Clipper is different from the two-sample t test or the regression-based t test, where the t statistic was purposely derived as a pivotal statistic so that its null distribution (the t distribution) does not depend on unknown parameters. Since Clipper does not require a null distribution for each feature, the advantage of the t statistic being pivotal no longer matters.

- Third, the minus statistic only requires estimates of two conditions' mean parameters, while the t statistic additionally requires estimates of the two conditions' variances. Hence, when the sample sizes (i.e., the numbers of replicates) are small, the two more parameters that need estimation in the t statistic might contribute to the observed power loss of the Clipper t statistic variant. Indeed, the power difference between the two statistics diminishes as the sample sizes increase from 3vs3 in Additional File 1: Figs. S13–S14 to 10vs10 in Additional File 1: Fig. S10 (where we compared the default Clipper with BH-pair-parametric, which is based on the two-sample t test and is highly similar to the Clipper t statistic variant).

- Fourth, we observe empirically that a contrast score would have better power if its distribution (based on its values of all features) has a larger range and a heavier right tail (in the positive domain). Compared to the minus statistic, the t statistic has a smaller range and a lighter right tail due to its adjustment for features' within-condition variances (Additional File 1: Fig. S23). This observation is consistent with the power difference of the two statistics.

Beyond our current interpretation, however, we admit that future studies are needed to explore alternative contrast scores and their power with respect to data characteristics and analysis tasks. Furthermore, we may generalize Clipper to be robust against sample batch effects by constructing the contrast score as a regression-based test statistic that has batch effects removed.

Our current version of Clipper allows the identification of interesting features between two conditions. However, there is a growing need to generalize our framework to identify features across more than two conditions. For example, temporal analysis of scRNA-seq data aims to identify genes whose expression levels change along cell pseudotime [31]. To tailor Clipper for such analysis, we could define a new

contrast score that differentiates the genes with stationary expression (uninteresting features) from the other genes with varying expression (interesting features). Further studies are needed to explore the possibility of extending Clipper to the regression framework so that Clipper can accommodate data of multiple conditions or even continuous conditions, as well as adjusting for confounding covariates.

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 SEQUEST 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 [58, 65]. 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. [66], 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 minus and maximum contrast scores used in Clipper. While we leave the optimization of contrast scores to future work, we observed that the minus contrast score outpowers the two-sample t statistic in our analysis (Additional File 1: Figs. S13 and S14), and we hypothesize 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 small-sample-size high-throughout biological data.

Conclusion

In high-throughput biological data analysis, which aims to identify interesting features by comparing two conditions, existing bioinformatics tools 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 bulk or single-cell RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data. Our results demonstrate Clipper's flexibility and reliability for FDR control, as well as its broad applications in high-throughput data analysis.

Methods

Clipper: 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 is the case with 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 $\{\mu_{Yj}\}_{j=1}^d$,

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

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

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

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 $t(\cdot, \cdot) : \mathbb{R}^m_{\geq 0} \times \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ to extract data information regarding whether a feature is interesting or not:

$$t^{\mathrm{minus}}(\boldsymbol{x}, \boldsymbol{y}) := \bar{x} - \bar{y}\,; \tag{3}$$

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

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.

Notably, other summary statistics can also be used to construct contrast scores. For example, an alternative summary statistic is the t statistic from the two-sample t test:

$$t^{\mathrm{t}}(\boldsymbol{x}, \boldsymbol{y}) := \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}}}\,. \tag{5}$$

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 $m + n > 2$).

Clipper: 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:

$$C_j := t^{\mathrm{minus}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{minus contrast score}\,, \tag{6}$$

or

$$C_j := t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{maximum contrast score}\,. \tag{7}$$

Fig. 6a 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*

$$T^{\mathrm{BC}} := \min\left\{t \in \mathcal{C} : \frac{\mathrm{card}(\{j : C_j \leq -t\}) + 1}{\mathrm{card}(\{j : C_j \geq t\}) \vee 1} \leq q\right\} \tag{8}$$

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

Clipper: 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 $\left\{(\boldsymbol{X}_j^\sigma, \boldsymbol{Y}_j^\sigma)\right\}_{j=1}^d$, where

$$\begin{aligned} \boldsymbol{X}_j^\sigma &:= \left(W_{j\sigma(1)}, \ldots, W_{j\sigma(m)}\right)^\top, \\ \boldsymbol{Y}_j^\sigma &:= \left(W_{j\sigma(m+1)}, \ldots, W_{j\sigma(m+n)}\right)^\top. \end{aligned} \tag{9}$$

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, $\ell = 0, 1, \ldots, h$. Then Clipper computes $T_j^{\sigma_\ell} := t^{\text{minus}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ 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^{minus} instead of $t^{\max}$ because empirical evidence shows that t^{minus} 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:

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

or

$$C_j := \begin{cases} \left|T_j^{(0)}\right| & \text{if } T_j^{(0)} = T_j^{\sigma_0} > T_j^{(1)} \\ 0 & \text{if } T_j^{(0)} = T_j^{(1)} \\ -\left|T_j^{(0)}\right| & \text{otherwise} \end{cases} \qquad \textbf{maximum contrast score}. \tag{11}$$

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 $T_j^{\sigma_1}, \ldots, T_j^{\sigma_h}$ (calculated 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. 6b (right) shows a cartoon illustration of contrast scores when $m = 2$ and $n = 1$. Motivated by Gimenez and Zou [63], 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 [63]) *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:*

$$T^{\text{GZ}} := \min\left\{t \in \mathcal{C} : \frac{\frac{1}{h} + \frac{1}{h}\text{card}\left(\{j : C_j \leq -t\}\right)}{\text{card}\left(\{j : C_j \geq t\}\right) \vee 1} \leq q\right\} \tag{12}$$

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

Clipper: 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 $h_{\text{total}} := \binom{m+n}{m}$ (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 $h_{\max} := h_{\text{total}} - 1$ 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 $T_j^{\sigma_\ell} := \left|t^{\text{minus}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})\right|$, where $\boldsymbol{X}_j^{\sigma_\ell}$ and $\boldsymbol{Y}_j^{\sigma_\ell}$ are the permuted data defined in (9), and it defines C_j as the contrast score of feature j, $j = 1, \ldots, d$, in the same ways as in (10) or (11). Fig. 6b (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 [63] to set a cutoff on contrast scores to control the FDR under the target level $q \in (0, 1)$.

Clipper: discussion on contrast scores

Granted, when we use permutations to construct contrast scores in the GZ procedure, we can convert contrast scores into permutation-based p-values (see Additional File 1: Section 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 [67], 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 (Additional File 1: Section S1.1.2).

For both differential and enrichment analyses, the two contrast scores (minus and maximum) can both control the FDR. Based on the power comparison results in Additional File 1: Section S3 and Additional File 1: Figs. S24–S28, Clipper has the following default choice of contrast score: for the enrichment analysis with $m = n$, Clipper uses the BC procedure with the minus contrast score; for the enrichment analysis with $m \neq n$ or the differential analysis, Clipper uses the GZ procedure with the maximum contrast score.

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 (Additional File 1: Definition S1) or Storey's qvalue procedure (Additional File 1: Definition S2) to make discoveries (Additional File 1: 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 [68]. 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 based on the feature’s observed test statistic under the experimental condition 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 (Additional File 1: 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). Additional File 1: 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, 69].

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.

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) [71]: used for qvalue-pool, qvalue-pair-correct, qvalue-pair-mis, and qvalue-pair-2as1.
- **locfdr** R package (v 1.1-8 with default arguments) [72]: 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.
- **ChIPulate** software package [45]: available at https://github.com/vishakad/chipulate.
- **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.
- **SEQUEST** 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.
- **limma** R package (v 3.44.3 with default arguments) [10]: available at https://www.bioconductor.org/packages/release/bioc/html/limma.html.
- **MAST** R package (v 1.14.0 with default arguments) [56]: available at https://www.bioconductor.org/packages/release/bioc/html/MAST.html.
- **monocle3** R package (v 0.2.3.0 with default arguments) [57]: available at https://github.com/cole-trapnell-lab/monocle3.
- **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/.

Availability of Data and Materials

- The **Clipper** R package is available at https://github.com/JSB-UCLA/Clipper/ [73].
- The code and processed data for reproducing the figures are available at https://zenodo.org/record/5115468 (DOI number 10.5281/zenodo.5115468) [74].
- A video introduction of Clipper is available at https://youtu.be/-GXyHiJMpLo.
- 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
  wgEncodeBroadHistoneGm12878H3k4me3StdAlnRep1.bam
  and the two control samples
  wgEncodeBroadHistoneGm12878ControlStdAlnRep1.bam
  wgEncodeBroadHistoneGm12878ControlStdAlnRep2.bam.
  The processed dataset is available at https://zenodo.org/record/5115468.
- The Pfu mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [70] partner repository with the dataset identifier PXD028558 (https://www.ebi.ac.uk/pride/archive/projects/PXD028558).
  The processed MS benchmark dataset is available as
  data/pfu/Archaea Sequest DECOYS 100_FDR.xlsx for the decoy PSMs and
  data/pfu/Archaea Sequest Targets 100_FDR.xlsx for the target PSMs
  at https://zenodo.org/record/5202768.
- The human monocyte RNA-seq dataset (SRA accession number SRP082682) 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 non-classical 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/5115468.
- The Hi-C dataset from the cell line GM12878 (GEO accession number GSE63525) 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/5115468.
- The PBMC single-cell RNA-seq dataset (GEO accession number GSE132044) is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132044. The processed dataset is available at https://zenodo.org/record/5115468.

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 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, 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).

Author Contribution

J.J.L., X.G., and Y.E.C. developed the methodology. X.G. performed the simulation analysis and the DEG analysis on semi-synthetic single-cell RNA-seq data. X.G. did the peak calling analysis, advised by W.L.. Y.E.C. did the DEG analysis on semi-synthetic bulk RNA-seq data, the DIR analysis, and the peptide identification analysis. The real data used in peptide identification analysis was generated by M.M., K.W., and A.M., supported by L.D.W.. X.G. and Y.E.C. performed DEG analysis on real bulk RNA-seq data, the results of which were interpreted by X.G., D.S., and N.W.. X.G. and Y.E.C. developed the R package. Y.E.C. wrote the mathematical proof. X.G., Y.E.C., and J.J.L. wrote the manuscript. The authors read and approved the final manuscript.

Ethics Approval

Not applicable.

Additional Files

Additional file 1

Supplementary materials. It includes a review of generic FDR control methods (Section S1 [75]), the detailed Clipper methodology (Section S2), the comparison of Clipper variant algorithms (Section S3), data generation and detailed implementation of the paired approach (a p-value calculation approach) in simulation studies (Section S4), bioinformatic methods with FDR control functionality (Section S5 [76, 77]), benchmark data generation in omics data applications (Section S6 [78–82]), implementation of Clipper in omics data applications (Section S7), proofs (Section S8), and Figures S1–S29.

Additional file 2

Supplementary table. Biological functions of the DEGs (between classical and non-classical human monocytes) that were identified by Clipper only but missed by both edgeR and DESeq2 from the human monocyte RNA-seq dataset (SRA accession number SRP082682).

References

[1] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.

[2] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[3] Michael P Washburn, Dirk Wolters, and John R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology*, 19(3):242–247, 2001.

[4] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[5] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[6] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[7] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[8] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):1–14, 2010.

[9] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[10] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[11] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multihiccompare: joint normalization and comparative analysis of complex hi-c experiments. *Bioinformatics*, 35(17):2916–2923, 2019.

[12] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. Find: differential chromatin interactions detection using a spatial poisson process. *Genome research*, 28(3):412–422, 2018.

[13] Aaron TL Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):1–11, 2015.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1):289–300, 1995.

[15] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[16] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[17] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[18] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[19] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[20] Bing Yang, Yan-Jie Wu, Ming Zhu, Sheng-Bo Fan, Jinzhong Lin, Kun Zhang, Shuang Li, Hao Chi, Yu-Xin Li, Hai-Feng Chen, et al. Identification of cross-linked peptides from complex samples. *Nature methods*, 9(9):904–906, 2012.

[21] James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, 5(4):e1000352, 2009.

[22] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[23] Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in bioinformatics*, 20(1):210–221, 2019.

[24] Ye Zheng and Sündüz Keleş. Freehi-c simulates high-fidelity hi-c data for benchmarking and data augmentation. *Nature Methods*, 17(1):37–40, 2020.

[25] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[26] Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[27] Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7): 577–580, 2016.

[28] Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *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. *PeerJ*, 6:e6035, 2018.

[30] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond p values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566, 2019.

[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. *bioRxiv*, 2020.

[32] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 (9):5116–5121, 2001.

[33] Jesse Hemerik and Jelle J Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155, 2018.

[34] Jesse Hemerik, Aldo Solari, and Jelle J Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649, 2019.

[35] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[36] Ery Arias-Castro and Shiyun Chen. Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001, 2017.

[37] Yoav Benjamini. Selective inference: The silent killer of replicability. *Issue 2.4*, 2(4), 2020.

[38] Kristen Emery, Syamand Hasam, William Stafford Noble, and Uri Keich. Multiple competition-based fdr control and its application to peptide detection. In *International Conference on Research in Computational Molecular Biology*, pages 54–71. Springer, 2020.

[39] Abhishek K Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single cell rna-seq analysis. *BioRxiv*, 2020.

[40] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.

[41] Pamela J Mitchell and Robert Tjian. Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, 245(4916):371–378, 1989.

[42] Mark Ptashne and Alexander Gann. Transcriptional activation by recruitment. *Nature*, 386(6625): 569–577, 1997.

[43] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS Comput Biol*, 9(11):e1003326, 2013.

[44] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[45] Vishaka Datta, Sridhar Hannenhalli, and Rahul Siddharthan. Chipulate: A comprehensive chip-seq simulation pipeline. *PLoS computational biology*, 15(3):e1006921, 2019.

[46] Aaron Diaz, Kiyoub Park, Daniel A Lim, and Jun S Song. Normalization, bias correction, and peak calling for chip-seq. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

[47] Boris Bogdanow, Henrik Zauber, and Matthias Selbach. Systematic errors in peptide and protein identification and quantification by modified peptides. *Molecular & Cellular Proteomics*, 15(8): 2791–2801, 2016.

[48] David N Perkins, Darryl JC Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, 20(18):3551–3567, 1999.

[49] Marshall Bern, Yong J Kil, and Christopher Becker. Byonic: advanced peptide and protein identification software. *Current protocols in bioinformatics*, 40(1):13–20, 2012.

[50] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

[51] Claire R Williams, Alyssa Baccarella, Jay Z Parrish, and Charles C Kim. Empirical assessment of analysis workflows for differential expression analysis of human samples using rna-seq. *BMC bioinformatics*, 18(1):38, 2017.

[52] Marek Gierliński, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon Simpson, Tom Owen-Hughes, et al. Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 31 (22):3625–3630, 2015.

[53] Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21, 2019.

[54] Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, 5(3):1752–1779, 2011.

[55] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5): 284–287, 2012.

[56] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

[57] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mrna quantification and differential analysis with census. *Nature methods*, 14(3):309–315, 2017.

[58] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[59] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: an interpretable simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *bioRxiv*, 2020.

[60] Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology*, 38 (6):737–746, 2020.

[61] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

[62] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[63] Jaime Roquero Gimenez and James Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *arXiv preprint arXiv:1810.11378*, 2018.

[64] Ning Wang, Diane Lefaudeux, Anup Mazumder, Jingyi Jessica Li, and Alexander Hoffmann. Identifying the combinatorial control of signal-dependent transcription factors. *PLOS Computational Biology*, 17(6):e1009095, 2021.

[65] Jonathan Thorsen, Asker Brejnrod, Martin Mortensen, Morten A Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rrna gene amplicon data analysis methods used in microbiome studies. *Microbiome*, 4(1):62, 2016.

[66] Kun He, Mengjie Li, Yan Fu, Fuzhou Gong, and Xiaoming Sun. Null-free false discovery rate control using decoy permutations for multiple testing. *arXiv preprint arXiv:1804.08222*, 2018.

[67] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2): 171–196, 1999.

[68] Alexey I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, 73(11):2092–2123, 2010.

[69] Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome biology*, 15(12):554, 2014.

[70] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, *et al.*, "The pride database and related tools and resources in 2019: improving support for quantification data," *Nucleic acids research*, vol. 47, no. D1, pp. D442–D450, 2019.

[71] John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2020. URL http://github.com/jdstorey/qvalue. R package version 2.20.0.

[72] Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[73] Xinzhou Ge, Yiling Chen, and Jingyi Jessica Li. Clipper: p-value-free FDR control on high-throughput data from two conditions. URL https://github.com/JSB-UCLA/Clipper. R package version 0.1.0.

[74] Xinzhou Ge, Yiling Chen, and Jingyi Jessica Li. Clipper: p-value-free FDR control on high-throughput data from two conditions, Jul 2021. URL https://zenodo.org/record/5115468.

[75] Rina Foygel Barber, Emmanuel J Candès, et al. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

[76] Marina Spivak, Jason Weston, Léon Bottou, Lukas Käll, and William Stafford Noble. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *Journal of proteome research*, 8(7):3737–3745, 2009.

[77] John C Stansfield, Kellen G Cresswell, Vladimir I Vladimirov, and Mikhail G Dozmorov. Hiccompare: an r-package for joint normalization and comparison of hi-c datasets. *BMC bioinformatics*, 19(1):279, 2018.

[78] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.

[79] Dattatreya Mellacheruvu, Zachary Wright, Amber L Couzens, Jean-Philippe Lambert, Nicole A St-Denis, Tuo Li, Yana V Miteva, Simon Hauri, Mihaela E Sardiu, Teck Yew Low, et al. The crapome: a contaminant repository for affinity purification–mass spectrometry data. *Nature methods*, 10(8): 730–736, 2013.

[80] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26 (12):1367–1372, 2008.

[81] Anton A Goloborodko, Lev I Levitsky, Mark V Ivanov, and Mikhail V Gorshkov. Pyteomics—a python framework for exploratory data analysis and rapid software prototyping in proteomics. *Journal of The American Society for Mass Spectrometry*, 24(2):301–304, 2013.

[82] Lev I Levitsky, Joshua A Klein, Mark V Ivanov, and Mikhail V Gorshkov. Pyteomics 4.0: five years of development of a python proteomics framework. *Journal of proteome research*, 18(2):709–714, 2018.

Figure legends

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 with 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 Additional File 1: Eq. (S14)) 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 semi-synthetic ChIP-seq data; **(b)** peptide identification on real proteomics data; **(c)** DEG analysis on semi-synthetic 10x Genomics scRNA-seq data; **(d)** DIR

analysis on semi-synthetic Hi-C data. In all four panels, the target FDR threshold q ranges from 1% to 10%. In the "Actual FDR vs. Target FDR" plot of each panel, points above the dashed diagonal line indicate failed FDR control; when this happens, the power of the corresponding methods is not shown, including MACS2 and HOMER in (a), edgeR 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.

Figure 4. Comparison of Clipper and two popular DEG identification methods—edgeR and DESeq2—in DEG analysis on semi-synthetic bulk RNA-seq data (generated from human monocyte real data using simulation strategy 2 in Additional File 1: Section S6.3). (a) FDR control, power given the same target FDR, and power given the same actual FDR. **(b)** Ranking consistency of the true DEGs among the top 100 DEGs identified by each method. The consistency is defined between the genes' ranking based on edgeR/DESeq2's p-values or Clipper's contrast scores and their ranking based on true expression fold changes. **(c)** Reproducibility between two semi-synthetic datasets as technical replicates. Three reproducibility criteria are used: the IDR, the Pearson correlation, and the Spearman correlation. Each criterion is calculated for edgeR/DESeq2's p-values or Clipper's contrast scores on the two semi-synthetic datasets. Among the three methods, only Clipper controls the FDR, and Clipper achieves the highest power, the best gene ranking consistency, and the best reproducibility.

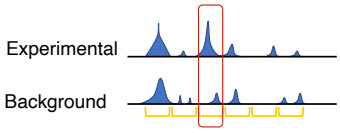
Figure 5. Application of Clipper, DESeq2, and edgeR to identifying DEGs from the classical and non-classical human monocyte dataset. (a) A Venn diagram showing the overlaps of the identified DEGs (at the FDR threshold $q = 5\%$) by the three DE methods. **(b)** Numbers of GO terms enriched (with enrichment 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. **(d)** A scatterplot of the claimed FDR of Clipper against that of edgeR for all the DEGs identified by Clipper, edgeR or DESeq2. The 46 DEGs identified by Clipper only at 5% FDR are highlighted in red.

Figure 6. 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 minus 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 minus contrast score.

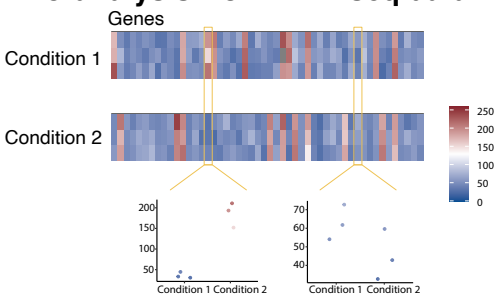
Figures

a High-throughput omics data analyses

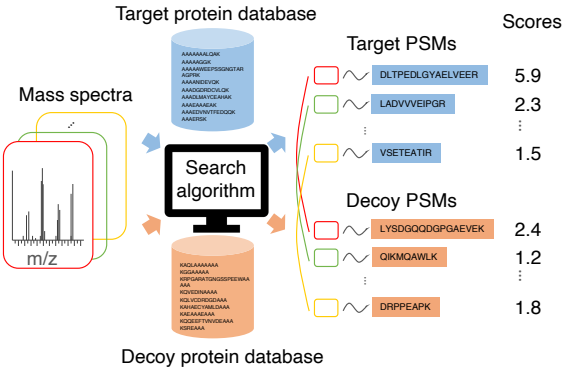
Peak calling from ChIP-seq data



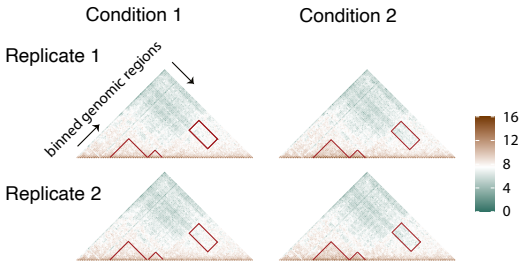
DEG analysis from RNA-seq data



Peptide identification from MS data



DIR analysis from Hi-C data



b Generic FDR control methods

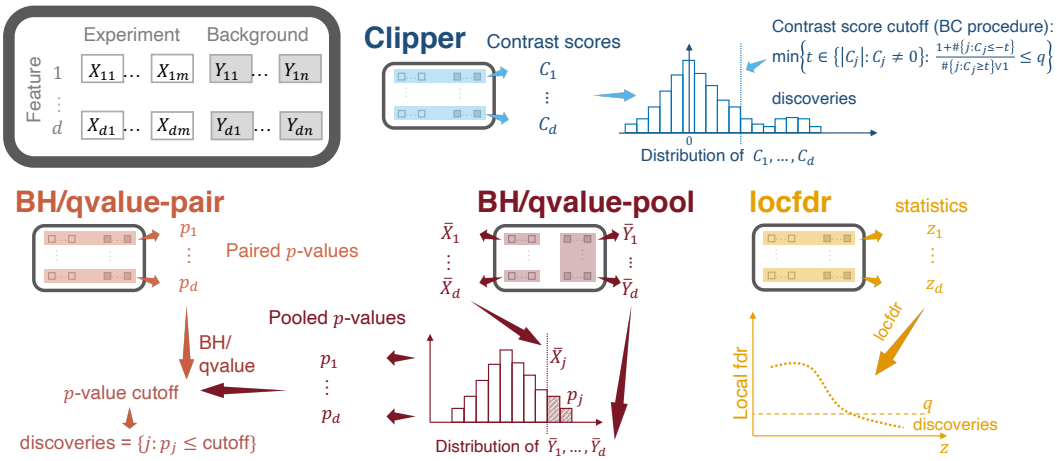


Figure 1

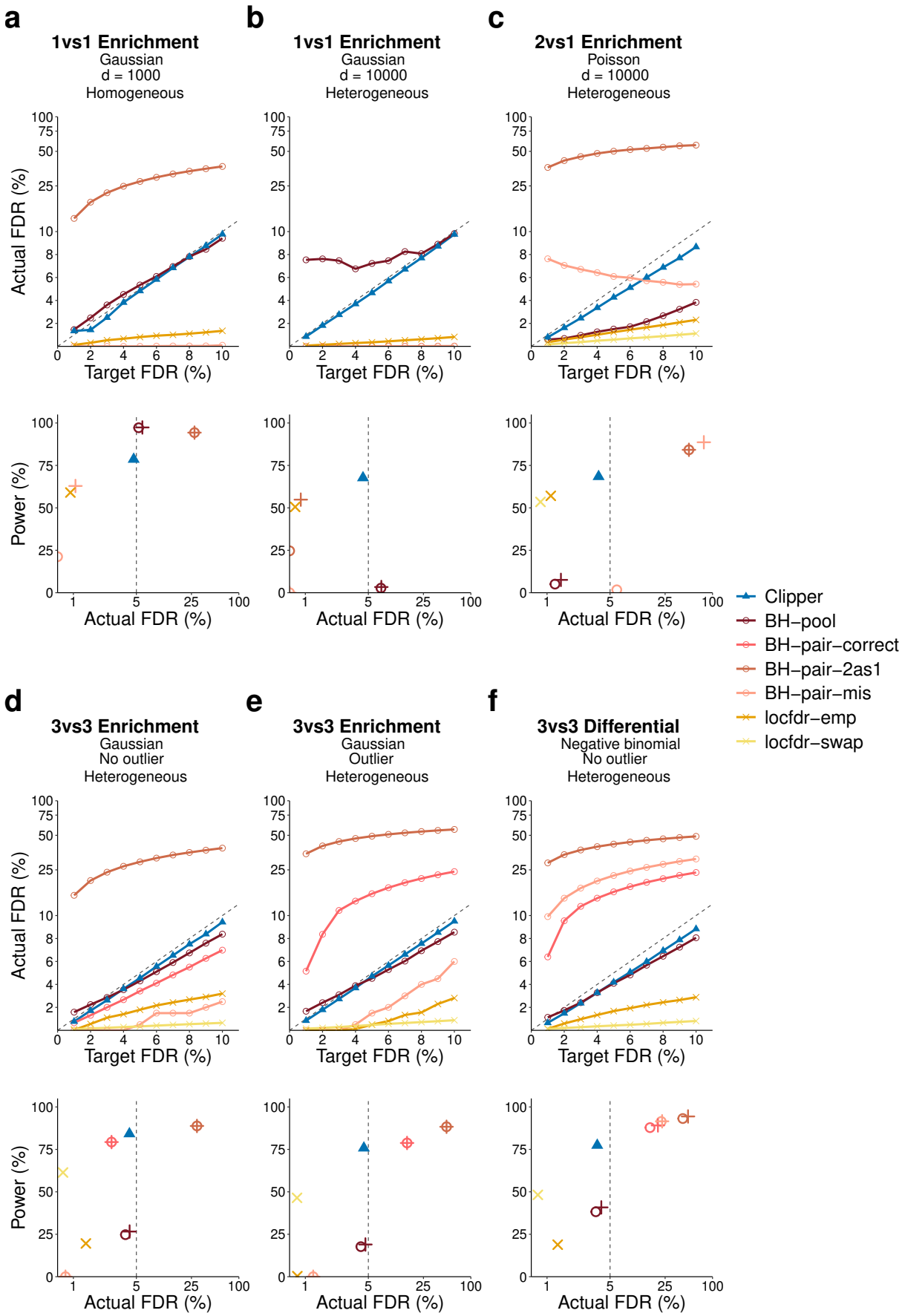

a
1vs1 Enrichment
Gaussian
d = 1000
Homogeneous
b
1vs1 Enrichment
Gaussian
d = 10000
Heterogeneous
c
2vs1 Enrichment
Poisson
d = 10000
Heterogeneous
d
3vs3 Enrichment
Gaussian
No outlier
Heterogeneous
e
3vs3 Enrichment
Gaussian
Outlier
Heterogeneous
f
3vs3 Differential
Negative binomial
No outlier
Heterogeneous
Actual FDR (%)
Target FDR (%)
Power (%)
Actual FDR (%)
Clipper
BH−pool
BH−pair−correct
BH−pair−2as1
BH−pair−mis
locfdr−emp
locfdr−swap

Figure 2

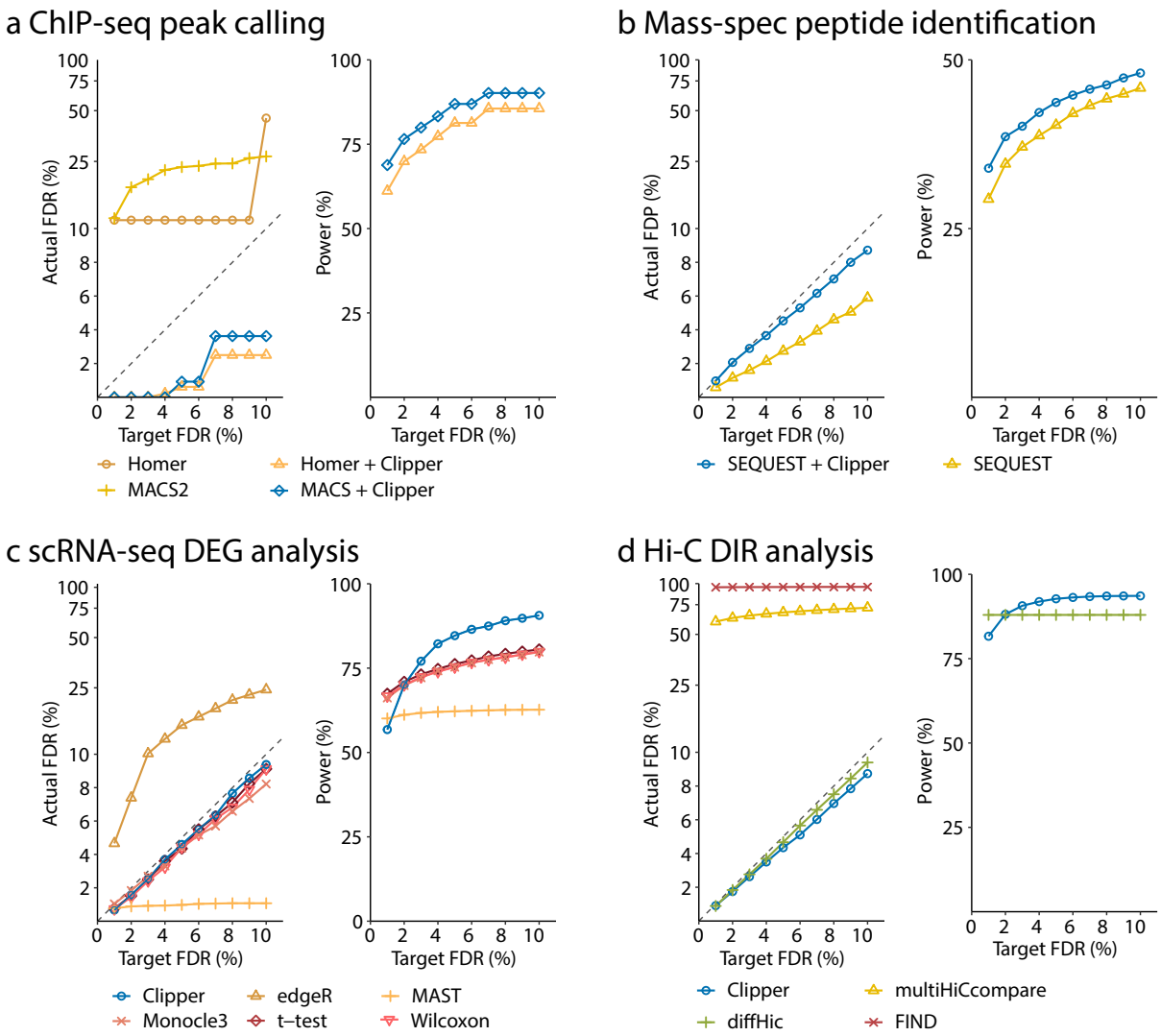

a ChIP-seq peak calling
Actual FDR (%)
Power (%)
Target FDR (%)
Homer
MACS2
Homer + Clipper
MACS + Clipper
b Mass-spec peptide identification
Actual FDP (%)
Power (%)
Target FDR (%)
SEQUEST + Clipper
SEQUEST
c scRNA-seq DEG analysis
Actual FDR (%)
Power (%)
Target FDR (%)
Clipper
Monocle3
edgeR
t–test
MAST
Wilcoxon
d Hi-C DIR analysis
Actual FDR (%)
Power (%)
Target FDR (%)
Clipper
diffHic
multiHiCcompare
FIND

Figure 3

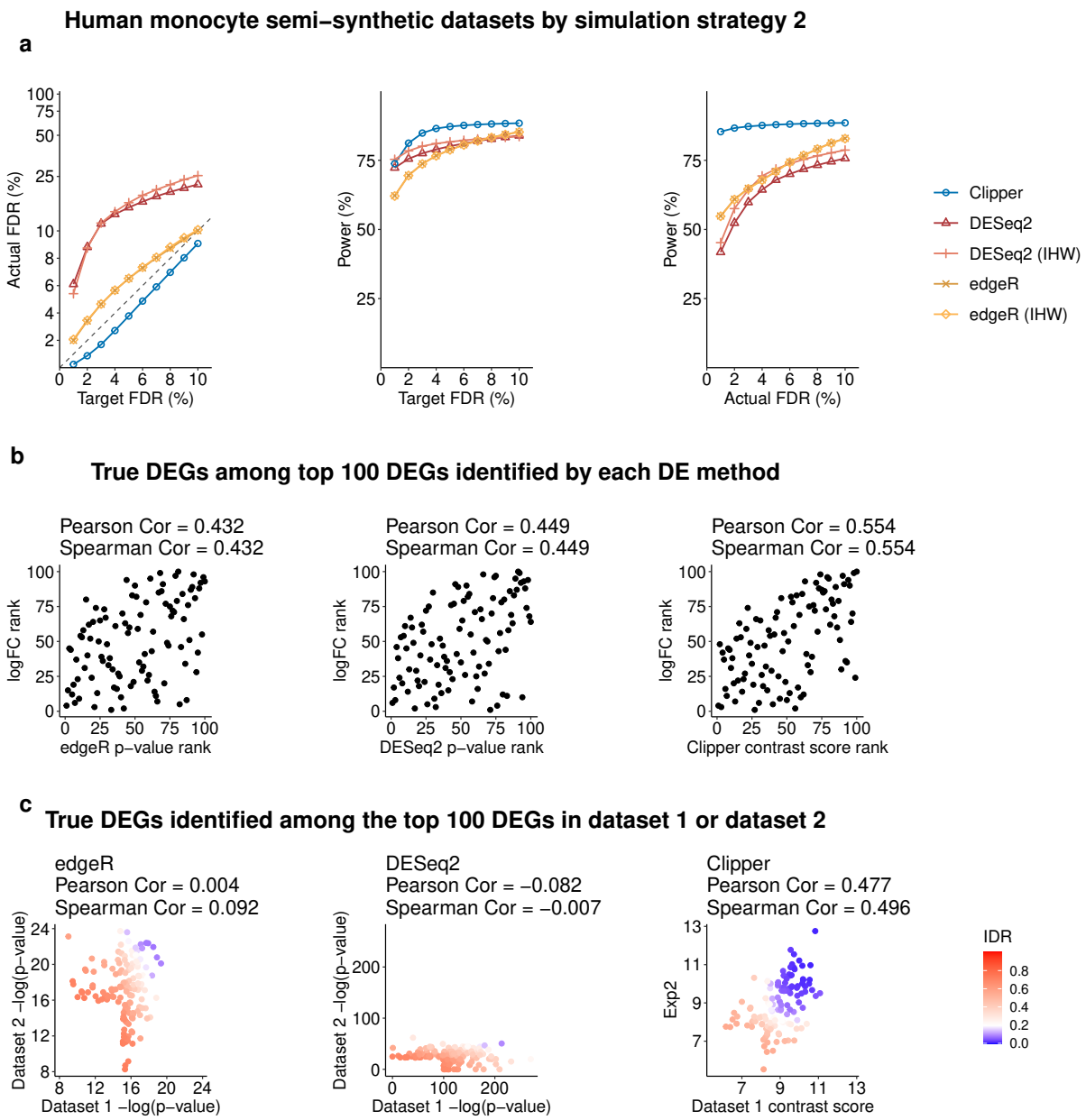
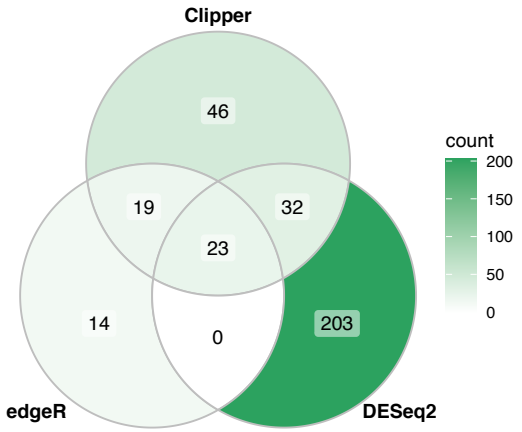
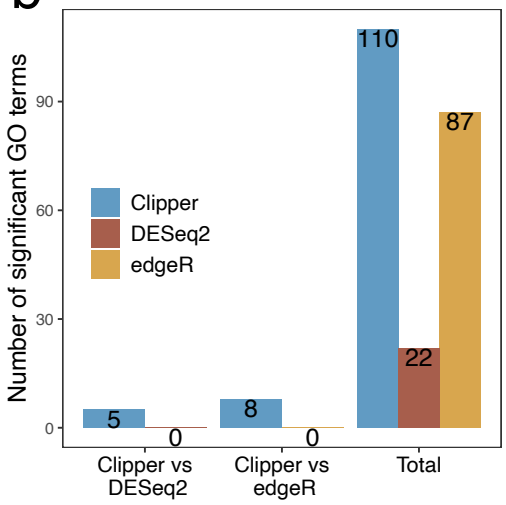

Human monocyte semi−synthetic datasets by simulation strategy 2
a
Actual FDR (%)
Target FDR (%)
Power (%)
Target FDR (%)
Power (%)
Actual FDR (%)
Clipper
DESeq2
DESeq2 (IHW)
edgeR
edgeR (IHW)
b
True DEGs among top 100 DEGs identified by each DE method
Pearson Cor = 0.432
Spearman Cor = 0.432
logFC rank
edgeR p−value rank
Pearson Cor = 0.449
Spearman Cor = 0.449
logFC rank
DESeq2 p−value rank
Pearson Cor = 0.554
Spearman Cor = 0.554
logFC rank
Clipper contrast score rank
c
True DEGs identified among the top 100 DEGs in dataset 1 or dataset 2
edgeR
Pearson Cor = 0.004
Spearman Cor = 0.092
Dataset 2 −log(p−value)
Dataset 1 −log(p−value)
DESeq2
Pearson Cor = −0.082
Spearman Cor = −0.007
Dataset 2 −log(p−value)
Dataset 1 −log(p−value)
Clipper
Pearson Cor = 0.477
Spearman Cor = 0.496
Exp2
Dataset 1 contrast score
IDR

Figure 4

a



b



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

d

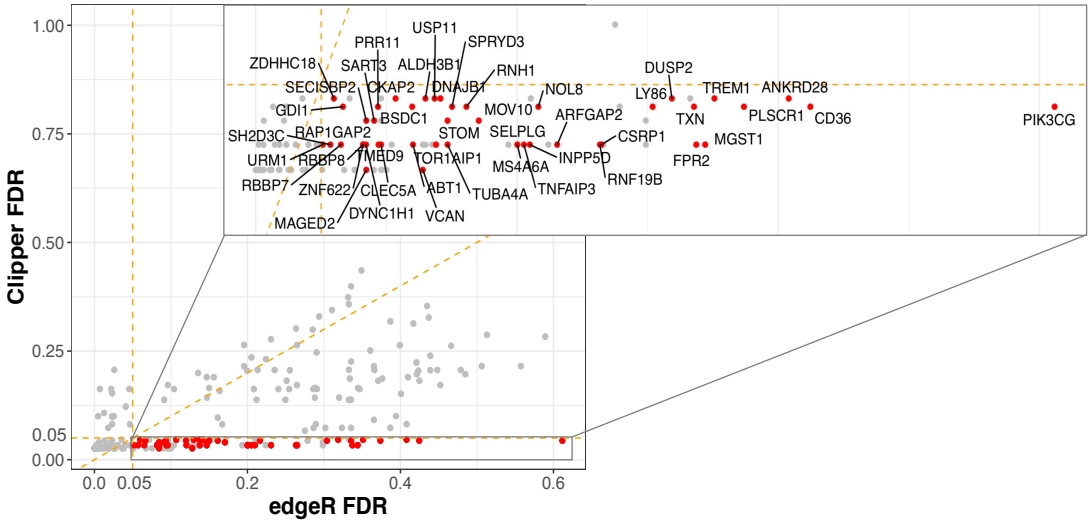


Figure 5

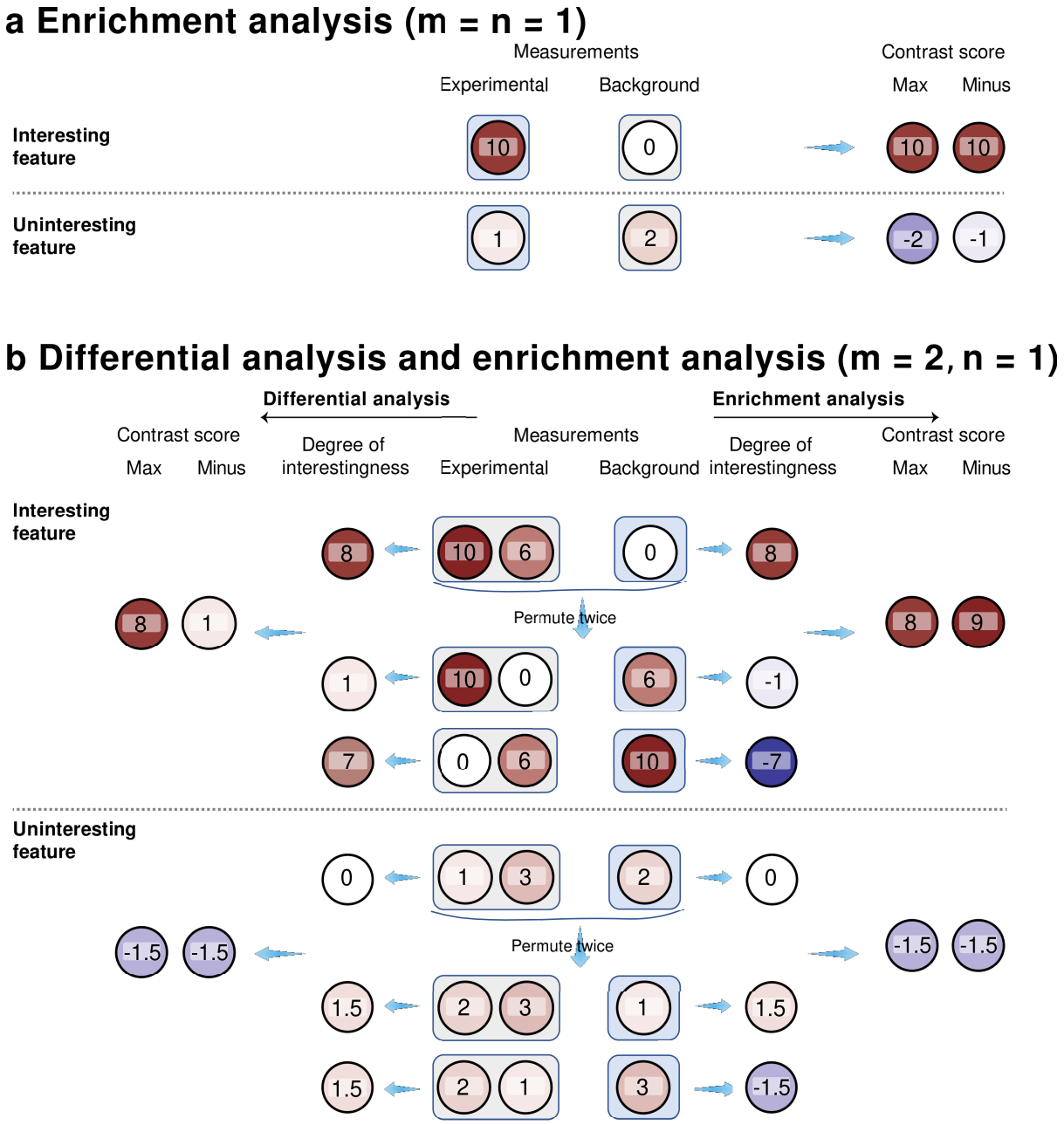

a Enrichment analysis (m = n = 1)
Measurements
Contrast score
Experimental
Background
Max
Minus
Interesting feature
10
0
10
10
Uninteresting feature
1
2
-2
-1
b Differential analysis and enrichment analysis (m = 2, n = 1)
Differential analysis
Enrichment analysis
Contrast score
Max
Minus
Degree of interestingness
Measurements
Experimental
Background
Degree of interestingness
Contrast score
Max
Minus
Interesting feature
8
10
6
0
8
8
1
Permute twice
8
9
1
10
0
6
-1
7
0
6
10
-7
Uninteresting feature
0
1
3
2
0
-1.5
-1.5
Permute twice
-1.5
-1.5
1.5
2
3
1
1.5
1.5
2
1
3
-1.5

Figure 6