bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Statistical Significance of Cluster Membership with Applications to High-Throughput Genomic Data

Neo Christopher Chung[1,2,*]

[1]Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

[2]NIH BD2K Center of Excellence in Biomedical Computing and Department of Physiology, David Geffen School of Medicine at UCLA

[*]Correspondence to nchchung@gmail.com

February 23, 2018

Contents

Abstract **2**

Introduction **3**

Methods and Algorithms **4**

Results **7**

Simulation Studies 7

Genomic Applications 9

Discussion **11**

References **14**

Supplementary Materials **17**

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Clustering is routinely applied to microarray, RNA-seq, and other genomic data, to help ascertain biological processes, disease subtypes, and cell identities. Recently, single cell RNA-seq (scRNA-seq) is used to automatically generate a large amount of gene expression profiles of unlabeled single cells. With no prior knowledge, the clustering algorithms are used to classify unlabeled single cells, that ultimately determine cell identities. However, how can we evaluate if the cluster memberships – the cell identities – are correctly assigned? To this end, we introduce the jackstraw methods for unsupervised classifications that rigorously test the assignments of genomic features into their clusters. By learning uncertainty in clustering the noisy data, the proposed jackstraw methods can identify statistically significant genomic features that truly make up the corresponding clusters. We investigated the proposed methods on scRNA-seq data from a mixture of Jurkat and 293T cell lines, where individual cell identities are unknown. The jackstraw methods evaluate cluster membership assignments of 3381 unlabeled single cells such that the majority of multiplets are identified in an unsupervised manner. We propose posterior inclusion probabilities (PIPs) for cluster membership to help select and visualize the reliable features in reduced dimensions. Additionally, we consider clustering 5981 yeast genes under cell cycle. When clustering is used in high-dimensional genomic data analysis, the proposed jackstraw tests enable rigorous evaluation of membership assignments that readily improve feature selection and visualization.

Software: `jackstraw` package in R available at `https://github.com/ncchung/jackstraw`

Keywords: scRNA-seq, RNA-seq, microarray, clustering, unsupervised classification

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

High-throughput technologies have enabled large-scale measurements of DNA, RNA, metabolites, and others. Recent technological and experimental advancements, such as single cell RNA-seq (Shalek et al., 2013; Wills et al., 2013) and mass-spectrometry (Rubakhin et al., 2011; Budnik et al., 2017), have resulted in increasing challenges and opportunities for using unlabeled data and unsupervised learning. For example, a single cell RNA-seq (scRNA-seq) technology enables gene expression measurements of thousands of blood cells in order to elucidate molecular subtypes. Unsupervised assignments of unlabeled single cells to K clusters according to their gene expression profiles provide cluster-based cell identities. Despite diverse clustering techniques available, it has not been possible to re-use data-driven clusters and test their membership in downstream statistical analyses without incurring artificially inflated significance. We have developed a novel and general method to statistically test the assignment of a data feature to a particular cluster, aiding in feature selection, dimension reduction, and visualization.

Clustering has been one of the most popular analysis methods for high-dimensional genomic data. In the absence of external and accurate labels, clustering can identify and approximate co-regulated subsets of genomic variables (e.g., genes, loci) or subtypes of related observations (e.g., patients, single cells). For example, a conventional microarray study measures gene expression of samples from either control or disease groups. Since the molecular functions of genes might not be known, the unsupervised classification of genomic variables can help identify co-varying subsets that form molecular processes (Alon et al., 1999; Golub et al., 1999; Sørlie et al., 2001). These clusters and membership assignments of genes have been extensively used in the visualization of systematic patterns and outliers (Spellman et al., 1998; Eisen et al., 1998). Recently, there have been many studies where mRNA abundances from thousands of single cells are measured en masse using scRNA-seq (Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017). Then, gene expression profiles of unlabeled single cells are clustered to obtain cell identities. These cell identities, which may be related to subtypes, lineage, or other molecular factors, are often used in downstream differential expression and other analyses. Note that a data feature refers to either a variable or an observation, since clustering can be applied on either dimension.

After automatically assigning observed features to K clusters that are summarized by K centers, we are interested in testing the membership assignments of individual features. This will improve the data-driven cell identities in scRNA-seq experiments, as well as the clustering of genomic variables to help elucidate molecular processes. To this end, we have developed an innovative data resampling and testing scheme for unsupervised classification that rigorously evaluates whether observed features are truly members of corresponding clusters. By estimating p-values and posterior inclusion probabilities (PIPs), the proposed methods can identify and visualize features that have been accurately and reliably assigned to the clusters. This bridges direct estimation of latent variables from large-scale data and fundamental hypothesis framework, which readily provides p-values, false discovery rates, and posterior probabilities crucial for data exploration and inference.

By utilizing a newly developed resampling technique called the jackstraw (Chung and Storey, 2015), the proposed methods learn overfitting inherent in using cluster centers that are estimated from the data. In other words, the proposed methods enable the accurate statistical testing of cluster membership while taking into account uncertainty in the clustering algorithms. Simulation studies demonstrate accurate and favorable operating characteristics. The joint behavior of p-values are scrutinized by conducting 100 independent simulations that satisfy the joint null criterion (Leek and Storey, 2011). Two applications are presented using two different dimensions (genomic variables and samples) available for unsupervised classification. Yeast cell cycle microarray data (Spellman et al., 1998) are used to cluster 5981 genes into $K = 6$ clusters, whose statistically significant members are identified. We also consider the scRNA-seq data from a mixture of two different cell lines (Zheng et al., 2017). By applying the proposed methods on unlabeled single cells, we show improved classification and visualization of cell identities. These proposed methods are implemented in a R package called *jackstraw* (`https://github.com/ncchung/jackstraw`), that will be available on the Comprehensive R Archive Network (CRAN).

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methods and Algorithms

The observed data $\mathbf{Y}_{(m,n)}$ contains m rows and n columns. Because either a set of variables (e.g., genes) or a set of observations (e.g., cells) may be clustered, we refer to m rows as m observed features for simplicity[1]. Then, $m_1, ..., m_K$ features are assigned into corresponding $1, ..., K$ clusters, where $\sum_{k=1}^{K}(m_k) = m$. The center $\mathbf{c}_k$ for $k = 1, ...K$ summarizes that k^{th} cluster. For example, in K-means clustering, the nearest means are used to assign observed features to the clusters. If $\mathbf{y}_i$ is assigned to k^{th} cluster with $\mathbf{c}_k$, its membership indicator $\beta_{i,k}$ is 1. By definition, the subset of features $\mathbf{y}_i$ with $\beta_{i,k} = 1$ make up $\mathbf{c}_k$.

Cluster centers and membership assignments may be viewed as approximating latent variables $\mathbf{L}$ and membership indicators $\mathbf{B}$ (i.e., dichotomous coefficients). Latent variables $\mathbf{l}_k$ for $k = 1, ..., K$ may assume a wide range of patterns including continuous or categorical structures (Linda M. Collins, 2010; Bartholomew et al., 2011). Clustering algorithms simultaneously identify the data features that contribute to the estimates of $\mathbf{L}_k$:

$$\mathbf{Y}_{(m,n)} = \mathbf{B}_{(m,K)}\mathbf{L}_{(K,n)} + \mathbf{E}_{(m,n)}$$

If a particular i^{th} feature is truly associated with a kth latent variable, its coefficient $b_{i,k}$ is 1. Otherwise, 0. Feature-specific noise $\mathbf{e}_i$ is defined as identically and independently distributed. Row-wise means are handled by centering the data, whereas row-wise variances are preserved by our proposed resampling scheme.

There have been important developments in clustering that consider mixture or latent variable models that improve our understanding and interpretation of data (Yeung et al., 2001; McLachlan and Peel, 2004; Fraley and Raftery, 2007). However, even model-based clustering approaches or regularization do not provide cluster centers and membership assignments that can be used again against the observed features, resulting in so-called "double dipping." Our proposed approach learns and incorporates inevitable uncertainty in assigning features to clusters, that are directly derived from the same set of features. This mirrors the jackstraw test when latent variables are estimated using principal component analysis (PCA) (Chung and Storey, 2015).

Jackstraw Data and Strategy

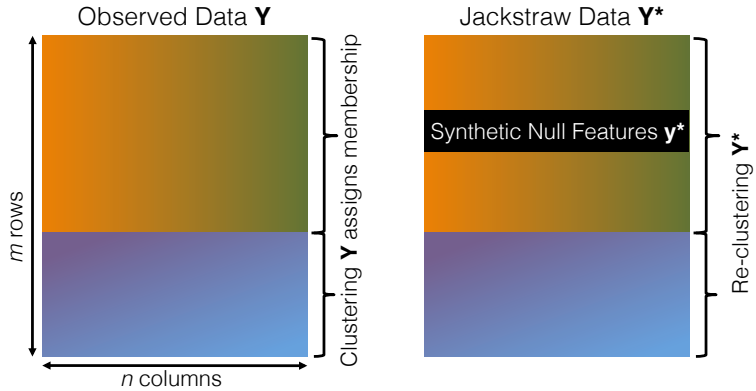

Figure 1: Illustration of the jackstraw data $\mathbf{Y}^*$ consisted of s synthetic null features ($\mathbf{y}^*$) and $m - s$ observed features. By re-clustering $\mathbf{Y}^*$, $\mathbf{y}^*$ that have been independently resampled are assigned into clusters. The jackstraw methods leverage this information to learn overfitting characteristics in clustering and to construct an empirical distribution of null statistics.

[1]This convention is also followed in the software package where the rows of input data are clustered and tested.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We apply the jackstraw strategy to clustering unlabeled features of observed data $\mathbf{Y}$. Generally, we would like to create a relatively small number s ($\ll m$ or n) of synthetic null features without disturbing the overall patterns of systematic variation. The jackstraw data $\mathbf{Y}^*$ refers to this revised data, where $m - s$ observed features are intact and s synthetic null features have been resampled with replacement (Figure 1). Applying the clustering algorithm on $\mathbf{Y}^*$ produces cluster centers $\mathbf{c}_k^*$ that are almost identical to the original cluster centers $\mathbf{c}_k$ (for $k = 1, ..., K$).

Because of the nature of clustering algorithms, all features in $\mathbf{Y}^*$, including s synthetic nulls, will be assigned to one of K clusters. When a synthetic null feature $\mathbf{y}_i^*$ is assigned to k^{th} cluster, an association statistics between $\mathbf{y}_i^*$ and $\mathbf{c}_k^*$ is under the null model that assumes independence since $\mathbf{y}_i^*$ is i.i.d. by definition. Yet, because $\mathbf{y}_i^*$ does indeed contribute to $\mathbf{c}_k^*$, we effectively learn the overfitting characteristics of the clustering algorithms. Over a large number of iterations $b = 1, ..., B$, we can form the empirical distribution of null statistics as in *Algorithm 1*.

Algorithm 1: Jackstraw Strategy for Unsupervised Classification

1. Apply the clustering algorithm to $\mathbf{Y}$, to obtain $\mathbf{C}$ and $\boldsymbol{\beta}$
2. Compute the observed statistics, relating $\mathbf{Y}$ and $\mathbf{C}$
3. Create $\mathbf{Y}^*$ with a small number of synthetic null features $\mathbf{y}^*$
4. Apply the clustering algorithm to $\mathbf{Y}^*$, to obtain $\mathbf{C}^*$ and $\boldsymbol{\beta}^*$
5. Compute the null statistics, relating $\mathbf{y}^*$ and $\mathbf{C}^*$
6. Repeat the above three steps to form an empirical distribution of null statistics

Feature-level evaluation of cluster membership requires a pre-defined number of clusters K. There is a vast amount of literature on the choice of K, which is beyond the scope of this study. In practice, a data analyst must explore the observed data, often utilizing prior knowledge, visualization, and heuristics. Methods have been proposed in the last five decades in this area of research including cluster stability or reliability statistics (Akaike, 1974; Schwarz et al., 1978; Bock, 1985; Fraley and Raftery, 1998; Pelleg et al., 2000; Tibshirani et al., 2001; Hamerly and Elkan, 2004; Liu et al., 2008; Huang et al., 2015). We recognize that data normalization, cluster stability, and other pre-classification steps are essential to sensible unsupervised learning. Through re-analysis of microarray and scRAN-seq data, we showcase the jackstraw tests in a context of broader unsupervised learning pipelines.

There are idiosyncratic outcomes of clustering that require our attention. Some clustering algorithms may generate an empty cluster or a singleton (a cluster with one feature). An empty cluster can be ignored in our methods as it does not contain any observed feature as a member. We consider the only feature of a singleton as its true member. It is possible that synthetic null features are rarely clustered into a certain cluster, such that there is a limited amount of empirical null statistics for that cluster. This likely occurs when that cluster is substantially smaller than others or has very distinct centers such that its members are tightly (and accurately) grouped in n dimensions. An increase of B would alleviate this, in tandem with examining the overall p-value distribution.

Jackstraw Tests for K-means Clustering

We now present a detailed algorithm using K-means clustering (MacQueen et al., 1967; Hartigan and Wong, 1979; Lloyd, 1982). K-means clustering is one of the most popular and well-studied algorithms that has been applied to a wide range of genomic studies. In this *Algorithm 2*, we use F-statistics where the full models include appropriate cluster centers. The use of F-statistics allows us to flexibly specify the full and null models, which may incorporate other covariates in more complex settings.

The choices of s and B controls the speed of computation, while the total number of null statistics ($s \times B$) determines the overall p-value resolution. For B iterations we need to cluster the jackstraw data B times, and for each iteration $b = 1, ..., B$, we can obtain s null statistics. Assuming $s \times B$ is hold constant,

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Algorithm 2: Jackstraw Test for Membership Assignments in K-means Clustering

1. Apply K-means clustering to the observed data $\mathbf{Y}$, resulting in cluster centers $\mathbf{c}_k$ for $k = 1, ..., K$ and membership assignments $b_{i,K}$ for $i = 1, ..., m$ and $K = 1, ..., k$
2. Compute the observed statistics $F_1, ..., F_m$, where the full models include corresponding cluster centers $\mathbf{c}_k$
3. Create s synthetic nulls by resampling a small proportion of features $s \ll m$ with replacement, resulting in a jackstraw data $\mathbf{Y}^*$, with $m - s$ observed features and s synthetic features
4. Apply the clustering algorithm to the jackstraw data $\mathbf{Y}^*$, resulting in cluster centers $\mathbf{c}_k^*$ and membership assignments $b_{i,K}^*$
5. Compute the null statistics $F_1^*, ..., F_s^*$, where the full models include corresponding cluster centers $\mathbf{c}_k^*$
6. Repeat the above three steps $b = 1, ..., B$ times to obtain a total $s * B$ of null statistics
7. Compute the p-values by empirically ranking the observed statistics among the null statistics, stratified by cluster assignments

a smaller s provides more accurate p-values, while increasing computational burdens. Therefore, we want to ensure the original clusters are preserved as much as possible, permitting the computational power. As we increase the number of synthetic null features s in $\mathbf{Y}^*$, the overall systematic variation captured by K cluster centers may be substantially disrupted (seen as an increasing proportion of $\mathbf{y}^*$ in Figure 1). While we recommend $s < .1 \times m$ for genomic data, although the number of clusters (K) and the proportion of features assigned to them ($m_1, \ldots, m_k$) must be considered. A higher value of K for a given m would need a smaller s, so that the clusters with limited members are represented in the jackstraw data.

The overwhelming disruption would further inflate null F-statistics, since a larger number of synthetic null features would make up $\mathbf{c}_k^*$. In extreme scenarios where all features have been resampled, the new cluster centers are completely dominated by independent synthetic null features. This operating characteristic allows us to guard against artificially inflated significance and to guide the input parameters for the proposed algorithm. In practice, we input $\mathbf{C}$ as the initial centers for K clusters when clustering the jackstraw data for efficient convergence. Furthermore, when a computational cost is a concern, one may correlate $\mathbf{C}$ and $\mathbf{C}^*$ to ensure comparability. We demonstrate this operating characteristics using the yeast gene expression under cell cycle, by applying the jackstraw tests with a wide range of s.

In contrast, the conventional resampling methods can be applied to the cluster centers, resulting in a "naive" significance test. After all m features are resampled with replacement, their F-statistics with respect to $\mathbf{c}_k$ are used to form an empirical distribution of null statistics. Observed F-statistics are compared to this empirical distribution to obtain naive p-values. This circular analysis inflates statistical significance, since the observed features are used twice to compute the cluster centers and to again test against the cluster centers. Essentially, this represents how the bootstrap or the permutation approaches would be applied to cluster membership assignments. We apply the conventional methods in simulation studies to demonstrate how the jackstraw approach overcomes this type of circular analysis.

Posterior Inclusion Probabilities

After the membership assignments for k^{th} cluster are tested using the jackstraw, we investigated how to harness their m_k p-values (or, the distribution of null statistics) to filter, de-noise, and visualize the clusters. When considering high-dimensional features typical in large-scale genomic studies, it is advantageous to consider a family of multiple hypotheses simultaneously (Efron, 2012). Particularly, from m_k jackstraw p-values, we propose to calculate posterior probabilities that features are included in a given cluster. A

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

discussion of posterior inclusion probabilities (PIPs) that are used for shrinkage and improvement of latent variable estimates is available in *Chapter 3* of (Chung, 2014).

Consider that the m_k jackstraw p-values $\mathbf{p}_k = p_{1,k}, \ldots, p_{m_k,k}$ are obtained for m_k features that have been assigned to k^{th} cluster. We are interested in estimating a posterior probability that $b_{i,k} \neq 0$, since non-zero coefficients imply their bona fide inclusion in the cluster:

$$\rho_{ik} = \Pr(b_{ik} \neq 0 | \mathbf{p}_{m_k}) = 1 - \Pr(b_{ik} = 0 | \mathbf{p}_{m_k}).$$

PIP can be readily obtained by estimating $\Pr(b_{ik} = 0 | \mathbf{p}_{m_k})$ through an empirical Bayes approach (Efron et al., 2001; Efron, 2007). In multiple hypothesis testing, $\Pr(b_{ik} = 0 | \mathbf{p}_{m_k})$ is called a local false discovery rate (FDR). There also exist related Bayesian methods that could be explored for specific applications and prior knowledge (Barbieri and Berger, 2004; Scott and Berger, 2005; Ghosh et al., 2006). These results in m PIPs for K families of multiple hypothesis tests corresponding to K clusters, that can be used for:

1. Retaining a subset of features $\mathbf{y}_i$ with $\boldsymbol{\rho}_i > \alpha_\rho$, where α_ρ is a user-defined threshold,
2. Visualizing features in reduced dimensions (e.g., PCA, t-SNE) where transparency $\sim \boldsymbol{\rho}_i$,
3. Improving the cluster centers by weighting the corresponding features with $\boldsymbol{\rho}_i$.

Local FDRs and PIPs from K families of multiple hypothesis tests can be flexibly combined for downstream analyses, as to aid feature selection and dimension reduction. When applying the proposed methods on microarray and scRNA-seq data, we incorporate PIPs to hard-threshold and soft-threshold the observed features. Furthermore, this approach may improve a wide range of clustering, by providing probabilistic measures and/or translating into fuzzy clustering algorithms.

Results

Unsupervised classification allows us to non-parametrically cluster large-scale data in absence of accurate external labels for data features. Because one may cluster genomic variables (such as genes) or observations (such as single cell samples), we refers to either set of data vectors to be clustered as features. Given the set of features are assigned into K clusters, we are interested in evaluating whether individual features have been correctly assigned. Based on the jackstraw approach (Figure 1), the proposed methods test their cluster membership assignments. To demonstrate its operating characteristics, we conducted simulation studies, which enabled a critical assessment using the underlying truth (*Oracle Groups*). We then applied the proposed methods on a microarray study of *Saccharomyces cerevisiae* that examines the cell cycle and another scRNA-seq data from a mixture of Jurkat and 293T cell lines whose cell identities are of interest.

Simulation Studies

In the simulation studies, we follow the latent variable model described in *Methods and Algorithms*. Latent variables $\mathbf{L}$ are drawn from the Normal($\mu = 0$, $\sigma^2 = 1$) distribution. Relationships between $\mathbf{l}_k$ and features are given by dichotomous coefficients $\mathbf{B}$ where $b_{i,k}$ indicates whether $\mathbf{y}_i$ is a member of $\mathbf{l}_k$ for $k = 1, ..., K$ and $i = 1, ..., m$. The noise $\mathbf{B}$ is drawn i.i.d. from Normal(0, σ_b^2), where its variance governs the noise level. A total of $m = 1000$ features (rows) are simulated over $n = 100$ dimensions (columns). Forming *Oracle Group A*, 500 rows are true members of the signal cluster arisen from $\mathbf{l}_1$ with $b_{i,1} = 1$ for $i = 1, ..., 500$. Other 500 rows are purely noise, in *Oracle Group B*, which can be viewed as being centered around the n-dimensional origin. Therefore, a true proportion of null features is $\pi_0 = .50$.

We simulated three scenarios using $\sigma_b^2 = 5, 10, 15$ as an increasing noise level brings these two groups closer and makes the clustering task more difficult. PCA was applied on the dataset realized from each configuration to visualize the top 2 PCs (Figure S1). Being blind to *Oracle Groups*, the K-means clustering and the jackstraw tests were applied. Theoretically, the null p-values from the features that are not related to the latent variables (corresponding to *Oracle Group B*) should form the Uniform(0,1) distribution, which can be evaluated by the Kolmogorov-Smirnov (KS) test. We repeated a given simulation configuration 100

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

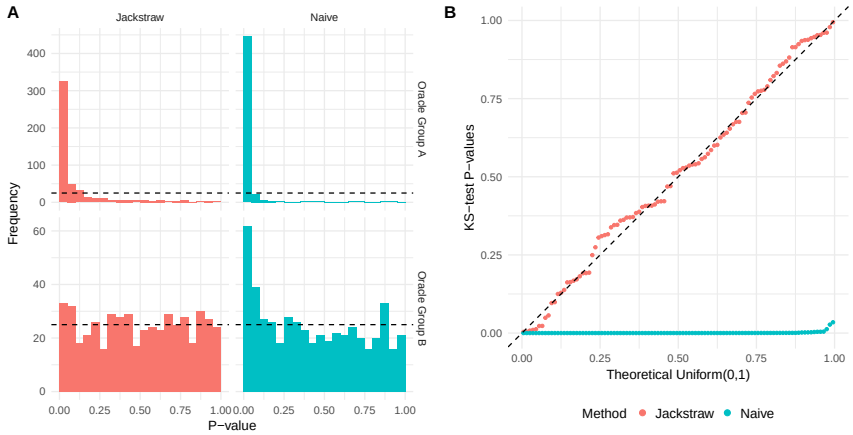

Figure 2: Evaluation of the jackstraw tests for clustering using the main simulation study with $\sigma = 10$. *Oracle Group A* contains 500 features that are derived from a latent variable $\mathbf{l} \overset{i.i.d}{\sim}$ Normal(0,1), whereas 500 features in *Oracle Group B* are noise. *(a)* Histograms of p-values stratified by methods. The jackstraw tests ($s = 100$, $B = 5000$) and the naive tests (at the same resolution) were applied without using any prior information. *(b)* This simulation study was repeated 100 times, where null p-values corresponding to *Oracle Group B* were evaluated against Uniform(0,1) distribution using Kolmogorov-Smirnov (KS) tests. QQ plot of KS p-values from the jackstraw and naive methods are shown, where valid p-values follow a diagonal line. The double KS test p-values for the proposed jackstraw and naive methods are $= 0.79$ and $< 2.2 \times 10^{-16}$, respectively.

times independently and investigated how 100 KS test p-values from 100 independent simulations meet the joint null criterion (Leek and Storey, 2011).

We describe one simulation from the main scenario involving a moderate amount of noise $\sigma_b^2 = 10$. While 1000 features were split equally between *Cluster 1* and *2*, 30 and 470 null features were members of *Cluster 1* and *2*, respectively. Because *Cluster 1* contained 470 features related to the latent variable $\mathbf{l}_1$, its center and $\mathbf{l}_1$ were highly correlated with a Pearson correlation of 0.99. The jackstraw test was then applied on the simulated data with $s = 100$ synthetic null features over $B = 5000$ iterations, while being blind to simulation parameters. Figure 2(a) shows histograms of p-values stratified by *Oracle Groups* as parametrized by dichotomous coefficients in $\mathbf{B}$. In *Oracle Group B*, the jackstraw p-values corresponding to 500 null features are uniformly distributed between 0 and 1. In contrast, the naive significance tests are highly anti-conservative, pushing towards 0. In *Oracle Group A*, the jackstraw p-values are greater than the naive p-values because the jackstraw approach learns the overfitting characteristics and fixes an anti-conservative bias (Figure 2(a)). Utilizing all m p-values, the proportion of null features are estimated to be $\widehat{\pi_0} = 0.55$ for the jackstraw and $\widehat{\pi_0} = 0.29$ for the naive methods.

We repeated this configuration to ensure accuracy and robustness across 100 independent simulations. In each simulation, we examined the joint behavior of 500 null p-values from *Oracle Group B* using a double-sided KS test. When the joint behavior of those KS test p-values follows the i.i.d. Uniform(0,1) distribution (where the double KS test p-value $> \alpha_{jnc}$), the subsequent multiple hypothesis testing procedures, including false discovery rates, hold true (Leek and Storey, 2011). In other words, meeting the stringent standard of the joint null criterion demonstrates that the proposed methods overcome "double-dipping" inherent in utilizing cluster centers and membership assignments and that the p-values are jointly and marginally accurate (Leek and Storey, 2011). A set of 100 KS test p-values, estimated from both the jackstraw and naive methods, are visualized against the Uniform(0,1) distribution (Figure 2(b)). The jackstraw tests satisfy the joint null

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

criterion, where 100 KS test p-values are uniformly distributed (double KS test p-value = 0.79). In contrast, the naive methods are strongly anti-conservative, where 100 KS test p-values are strongly skewed towards 0 (double KS test p-values $< 2.2 \times 10^{-16}$).

Results from two other simulation configurations, that are also independently repeated 100 times, are shown in Figure S2 and Figure S3. Simulated data with a relatively small noise $\sigma_b^2 = 5$ can be almost perfectly clustered. Nonetheless, the naive methods exhibit substantial overfitting where the double KS test p-value is $< 2.2 \times 10^{-16}$. The double KS test p-value for the jackstraw tests in this configuration is 0.81 (Figure S2). On the other hand, a greater noise with $\sigma_b^2 = 15$ represents a situation whose members of different clusters are substantially overlapping (Figure S1). The jackstraw tests indeed satisfy the joint null criterion with the double KS test p-value of 0.67 (Figure S3). Additional simulation studies further confirm that unlike the naive methods that overfit and produce an anti-conservative bias (downward deviations from the diagonal line), the proposed methods take account for uncertainty in clustering and result in valid p-values that enable rigorous error control.

Genomic Applications

Microarray Data from Yeast Cell Cycle Experiments

Cell cycle in *Saccharomyces cerevisiae* and other organisms are traditionally known to progress through discrete stages, such as M, G1, S, and so on (Alberts, 2017). With the advent of the microarray, gene expression levels from synchronized *S. cerevisiae* samples had been measured and analyzed in order to identify comprehensive sets of genes under cell cycle (Spellman et al., 1998; Cho et al., 1998; Tu et al., 2005; Rowicka et al., 2007). However, in these conventional studies, experimentally verified genes under cell cycle are used to identify related genes that follow similar patterns. In contrast, we re-analyzed the expression data of 5981 genes from (Spellman et al., 1998) in an unsupervised manner.

Genome-wide mRNA levels of elutriation-synchronized yeast cells were measured at 30 min intervals for 390 min (approximately 1 cell cycle) (Spellman et al., 1998). We processed and normalized this gene expression data according to (Alter et al., 2000; Chung and Storey, 2015). The number of clusters $K = 6$ was determined by the prior knowledge that there exist 6 stages of cell cycle (Alberts, 2017). While there are on-going debates on how to characterize and categorize cell cycle progression, $K = 6$ seems to be a reasonable choice. After having gotten $K = 6$ clusters from applying K-means clustering, we conducted the proposed jackstraw tests with $s = 300$ and $B = 10000$ to identify canonical genes within those clusters.

Histograms of p-values are shown in Figure 3(a), where the proportions of null features π_0 for 6 clusters are estimated to be .143, .149, .116, .178, .170, .175, .087, respectively. Note that the numeric values identifying those clusters are arbitrary without a meaningful order, but consistent within this manuscript. From a set of p-values, we calculated posterior probabilities that those genes are truly members of their assigned clusters. For example, among 709 genes that are originally assigned to the cluster 4, 45.1% (320) have posterior inclusion probabilities (PIPs) > 0.9. Repeating this analysis for all 6 clusters, a total of 3826 genes are found to be significant at the same PIP threshold (Figure 3(c)). In other words, these are the canonical genes that drive the clusters of cell cycle.

Using this gene expression data, we investigated the operating characteristics of synthetic null features $\mathbf{y}^*$ in the jackstraw test for clustering. Particularly, we applied the jackstraw tests with an increasing number of synthetic null features $s = 10, 100, 250, 500, 1000, 2500, 5000$. To obtain the same number of total null statistics $s \times B = 5 \times 10^6$, the number of iterations were set to $B = 5 \times 10^5, 5 \times 10^4, 2 \times 10^4, 1 \times 10^4, 5 \times 10^3, 2 \times 10^3, 1 \times 10^3$ correspondingly. We found that an increase in s results in slightly more conservative behavior (Figure S4). This trade-off between computational speed and accuracy of p-values is beneficial in practice since even limited computation with a big s would not incur artificially significant results.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

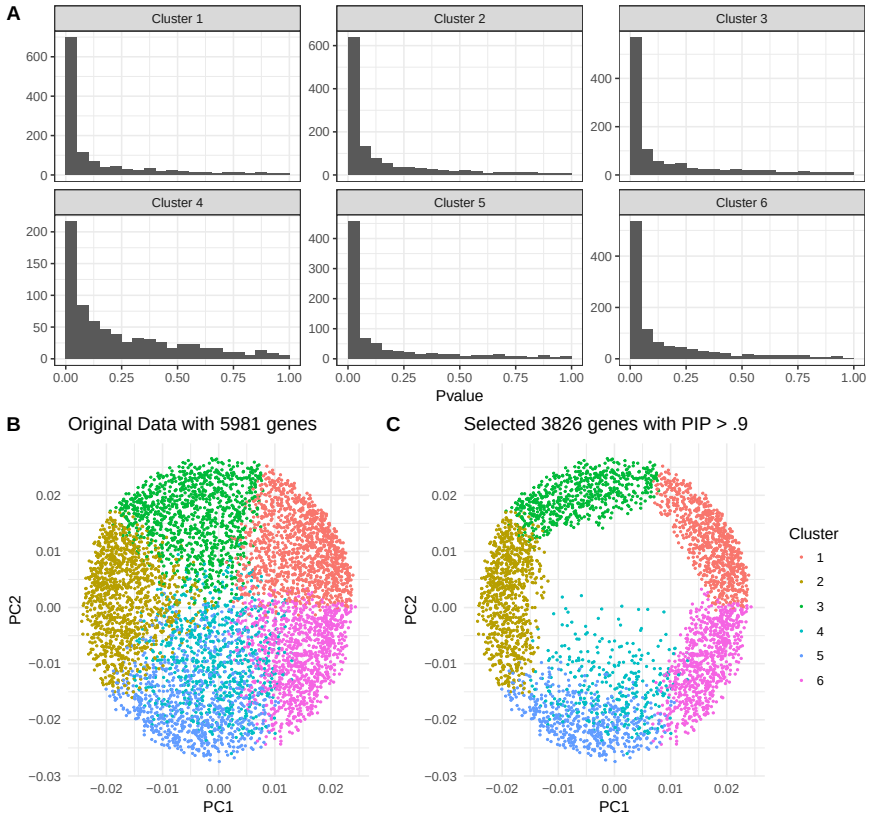

Figure 3: The jackstraw clustering analysis of microarray data of yeast cell cycle experiments. K-means clustering is applied on 5981 genes from (Spellman et al., 1998) with $K = 6$. *(a)* Histograms of p-values from the yeast cell cycle gene expression profiles. The jackstraw tests for these 6 clusters are conducted with $s = 300$ and $B = 10000$. *(b)* The top 2 PC projection using the original data with 5981 genes. *(b)* The top 2 PC projection using 3826 genes with PIP $> .9$ from the proposed methods.

Single Cell RNA-Seq Data from Jurkat and 293T Cells

Whereas conventional microarray and RNA-seq experiments obtain "bulk" gene expression from a sample that contains multiple cells, scRNA-seq enable more precise and accurate quantification from single cell samples. Recent studies using high-throughput and efficient scRNA-seq often measure gene expression from unlabeled single cells, in order to elucidate detailed molecular landscapes and identify cell identities (e.g., blood sub-types, sub-classifications of a disease) (Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017). Commonly, the cell identities are determined by applying the clustering algorithms to their gene expression profiles.

We analyzed the scRNA-seq data from (Zheng et al., 2017) that used a mixture of Jurkat and 293T cells (50:50). Note that while the mixture proportion is known, the identities of individual cells that have been sequenced are unknown. Because Jurkat (male and expressing *CD3D*) and 293T (female and expressing *XIST*) cell lines are highly distinct, we observed intelligible two groups separated along the 1st PC from their gene expression profiles (Zheng et al., 2017). However, massively parallel scRNA-seq regrettably generates multiplets (doublets, triplets, etc). The rate of multiplets increases linearly with the recovered cell number,

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

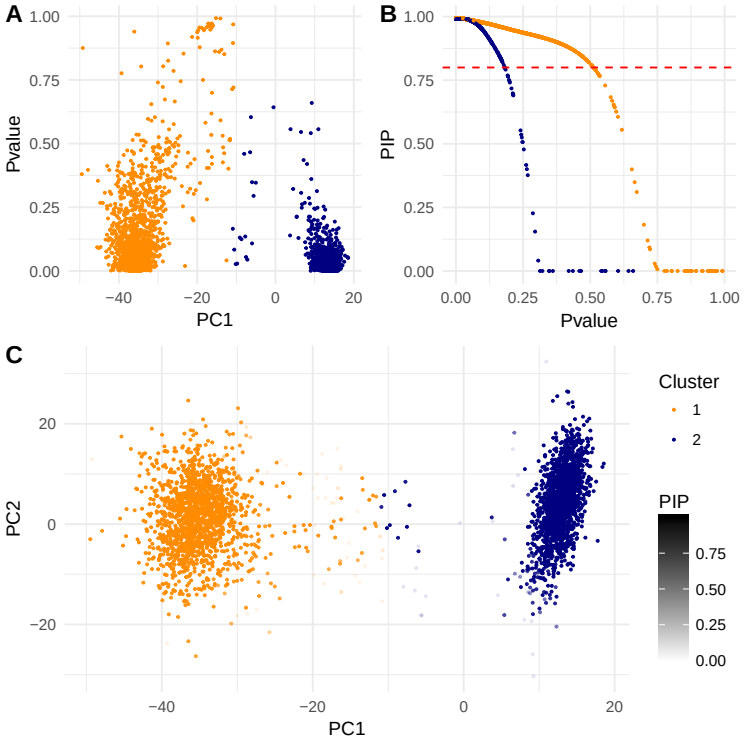

Figure 4: The jackstraw clustering analysis of scRNA-seq data of Jurkat and 293T cells (Zheng et al., 2017). *(a)* Clustering membership p-values are plotted against the 1st PC, which largely separates two cell lines. *(b)* Posterior inclusion probabilities (PIPs) are computed from p-values. At a PIP threshold of 0.80, 3.3% of 3381 single cells would be discharged from corresponding clusters. *(c)* PIPs are visualized as alpha levels on the scatterplot of the top 2 PCs. Essentially, this is identical to Figure 2(e) in (Zheng et al., 2017), except transparencies are set to PIPs from the proposed methods. Note that when PIP=0, as appeared in Figure 4(b), the data point is completely transparent.

and through single nucleotide variant (SNV) detection, they inferred a 3.1% multiplet rate for this mixture experiment (Zheng et al., 2017). For $\sim$ 10000 single cells, (Zheng et al., 2017) reports $>$ 8% multiplet rates. The ambiguous identities of singe cells would become increasingly challenging as scRNA-seq becomes more affordable and widespread.

Following the original analysis pipeline, we applied the K-means clustering on the top 10 PCs based on unique molecular identifier (UMI) counts. The jackstraw tests for those $K = 2$ clusters were conducted with $s = 100$ and $B = 10000$. We found that the jackstraw p-values capture deviation away from two centers, along the 1st PC axis (Figure 4(a)). Using the q-value methodology (Storey and Tibshirani, 2003), the proportion of null features (that are not members of the clusters) is estimated to be $\widehat{\pi_0} = 0.05$. Then, we computed the proposed PIPs from p-values (Figure 4(b)). At PIP < 0.80 (equivalent to 20% local FDRs), 3.3% of 3381 single cells would be removed from corresponding clusters, effectively and automatically removing the majority of suspected multiplets. Instead of hard-thresholding the single cell samples at an arbitrary threshold, we can also visualize posterior probabilities as levels of transparency in a conventional PCA projection, where the top 2 PCs are plotted (Figure 4(c)). Please note that because dimension reduction

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

does not fully capture local and global structures in the original high dimensions, distances in reduced dimensions (PCA, t-SNE, and alike) should be considered with caution.

Discussion

The explosion of biological data has increased the importance of unsupervised learning. Without the external and accurate labels for the observed data, unsupervised learning aims to estimate latent structure, reduce dimensions, and classify data features. In particular, clustering of high-dimensional genomic data has led to better understanding of and informative hypotheses for biological functions (Spellman et al., 1998; Jang et al., 2017), molecular subtypes (Sørlie et al., 2001; Wang and Gu, 2016), and cell identities (Wagner et al., 2016). However, data-dependent classification cannot be used in downstream analyses without incurring spurious statistical significance. Our proposed methods solve this challenge by learning the uncertainty inherent in deriving clusters from the data and conducting a statistical test using the jackstraw strategy.

There exists a wide range of clustering algorithms to automatically assign the m observed features into K clusters. The proposed methods test whether the observed features are correctly assigned to the corresponding clusters. Our key ingredient is to generate and re-cluster the jackstraw data, which include a very small number s of synthetic null features. Because of $s \ll m$, the majority of observed features are intact, resulting in cluster centers that are almost identical to the original cluster centers. Subsequently, eventual assignments of s synthetic null features into K clusters are used to derive the empirical null distribution. We have demonstrated favorable operating characteristics using simulated and real genomic data. The proposed PIP methods open new possibilities for selecting canonical cluster members, shrinking cluster centers and improving cluster algorithms. Furthermore, the proposed methods may adaptively guide the choice of stable clusters.

Our proposed strategy enables rigorous application of unsupervised learning, such that the estimated latent structure can be re-used in downstream analyses. The jackstraw test for PCA and related methods (Chung and Storey, 2015) have been used in many specialized areas of genomic studies (Macosko et al., 2015; Satija et al., 2015; Chung et al., 2017; Farré et al., 2015; Jang et al., 2017; Zheng et al., 2017). Complementing this successful approach, we have developed the jackstraw test for clustering. It may be useful to integrate both variants of the jackstraw tests, from selecting highly informative genes to deriving cell identities. Differential expression analyses based on cluster-based cell identities may become more robust by incorporating the jackstraw tests. Because the proposed methods are not limited to genomics, we anticipate its adaptation in other fields of data-intensive science.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Data Access

The proposed methods are implemented in an open source R package, `https://github.com/ncchung/jackstraw`.

Acknowledgments

This research was supported in part by the Polish National Science Centre (NCN) grants 2016/23/D/ST6/03613.

Disclosure Declaration

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

Akaike, H., 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6):716–723.

Alberts, B., 2017. *Molecular biology of the cell*. Garland science.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, **96**(12):6745–6750.

Alter, O., Brown, P. O., and Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(18):10101–10106.

Barbieri, M. and Berger, J., 2004. Optimal predictive model selection. *Annals of Statistics*, **32**:870–897.

Bartholomew, D. J., Knott, M., and Moustaki, I., 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley Series in Probability and Statistics.

Bock, H.-H., 1985. On some significance tests in cluster analysis. *Journal of classification*, **2**(1):77–108.

Budnik, B., Levy, E., and Slavov, N., 2017. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *bioRxiv*, **bioRxiv**.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., *et al.*, 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, **2**(1):65–73.

Chung, N. C., 2014. *Statistical Inference of Variables Driving Systematic Variation in High-Dimensional Biological Data*. PhD thesis, Princeton University.

Chung, N. C. and Storey, J. D., 2015. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, **31**(4):545–554.

Chung, N. C., Szyda, J., Frąszczak, M., Fries, H. R., SandøLund, M., Guldbrandtsen, B., Boichard, D., Stothard, P., Veerkamp, R., Goddard, M., *et al.*, 2017. Population structure analysis of bull genomes of european and western ancestry. *Scientific Reports*, **7**:40688.

Efron, B., 2007. Size, power and false discovery rates. *The Annals of Statistics*, **35**(4):1351–1377.

Efron, B., 2012. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V., 2001. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**(456):1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(25):14863–14868.

Farré, P., Jones, M. J., Meaney, M. J., Emberly, E., Turecki, G., and Kobor, M. S., 2015. Concordant and discordant DNA methylation signatures of aging in human blood and brain. *Epigenetics & Chromatin*, **8**(1).

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fraley, C. and Raftery, A. E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. *Computer Journal*, **41**(8):578–588.

Fraley, C. and Raftery, A. E., 2007. Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software*, **18**(6):1–13.

Ghosh, D., Chen, C., and Raghunathan, T., 2006. The false discovery rate: a variable selection perspective. *Journal of Statistical Planning and Inference*, **136**:2668 Ð 2684.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.*, 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**(5439):531–537.

Hamerly, G. and Elkan, C., 2004. Learning the k in k-means. In *Advances in neural information processing systems*, pages 281–288.

Hartigan, J. A. and Wong, M. A., 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1):100–108.

Huang, H., Liu, Y., Yuan, M., and Marron, J., 2015. Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, **24**(4):975–993.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.*, 2014. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, **343**(6172):776–779.

Jang, S., Choubey, S., Furchtgott, L., Zou, L.-N., Doyle, A., Menon, V., Loew, E. B., Krostag, A.-R., Martinez, R. A., Madisen, L., *et al.*, 2017. Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states. *eLife*, **6**.

Leek, J. T. and Storey, J. D., 2011. The joint null criterion for multiple hypothesis tests. *Statistical Applications in Genetics and Molecular Biology*, **10**(1):Article 28.

Linda M. Collins, S. T. L., 2010. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. John Wiley & Sons.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J., 2008. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, **103**(483):1281–1293.

Lloyd, S., 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, **28**(2):129–137.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al.*, 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**(5):1202–1214.

MacQueen, J. et al., 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1 of *14*, pages 281–297. Oakland, CA, USA.

McLachlan, G. and Peel, D., 2004. *Finite mixture models*. John Wiley & Sons.

Pelleg, D., Moore, A. W., et al., 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, pages 727–734.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Rowicka, M., Kudlicki, A., Tu, B. P., and Otwinowski, Z., 2007. High-resolution timing of cell cycle-regulated gene expression. *Proc Natl Acad Sci U S A*, **104**(43):16892–16897.

Rubakhin, S. S., Romanova, E. V., Nemes, P., and Sweedler, J. V., 2011. Profiling metabolites and peptides in single cells. *Nature methods*, **8**(4s):S20–S29.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, **33**(5):495–502.

Schwarz, G. et al., 1978. Estimating the dimension of a model. *The annals of statistics*, **6**(2):461–464.

Scott, J. and Berger, J., 2005. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**:2144–2162.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.*, 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**(7453):236–240.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., *et al.*, 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**(19):10869–10874.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**(12):3273–3297.

Storey, J. D. and Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16):9440–9445.

Tibshirani, R., Walther, G., and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2):411–423.

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L., 2005. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**(5751):1152–1158.

Wagner, A., Regev, A., and Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, **34**(11):1145–1160.

Wang, D. and Gu, J., 2016. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, **4**(1):58–67.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C., 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, **31**(8):748–752.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L., 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**(10):977–987.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al.*, 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**:14049.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Materials

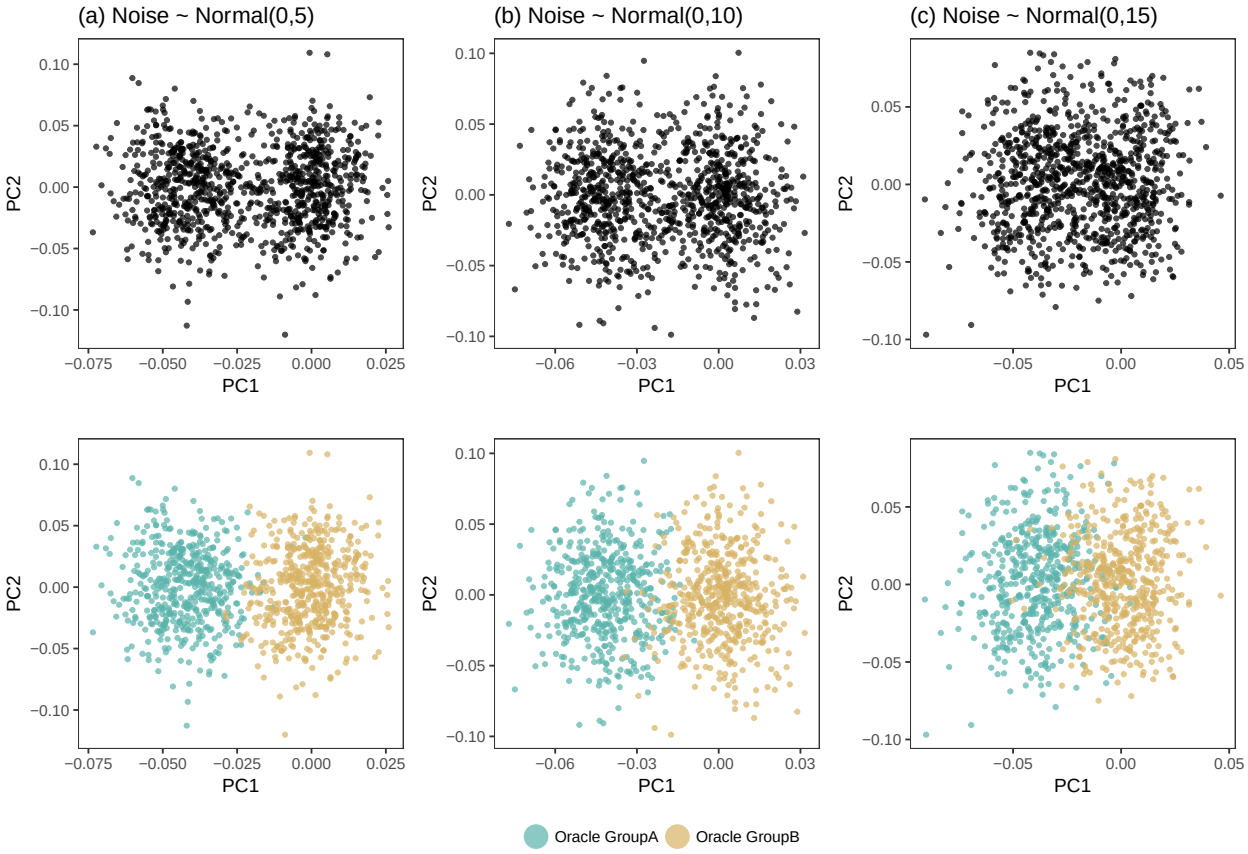

Figure S1: Scatterplots of the top 2 principal components (PCs) from the simulated data. *Oracle Groups* are shown in colors. An increasing level of noise, $\sigma^2 = 5, 10, 15$ brings data features from two different underlying structures closer together.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

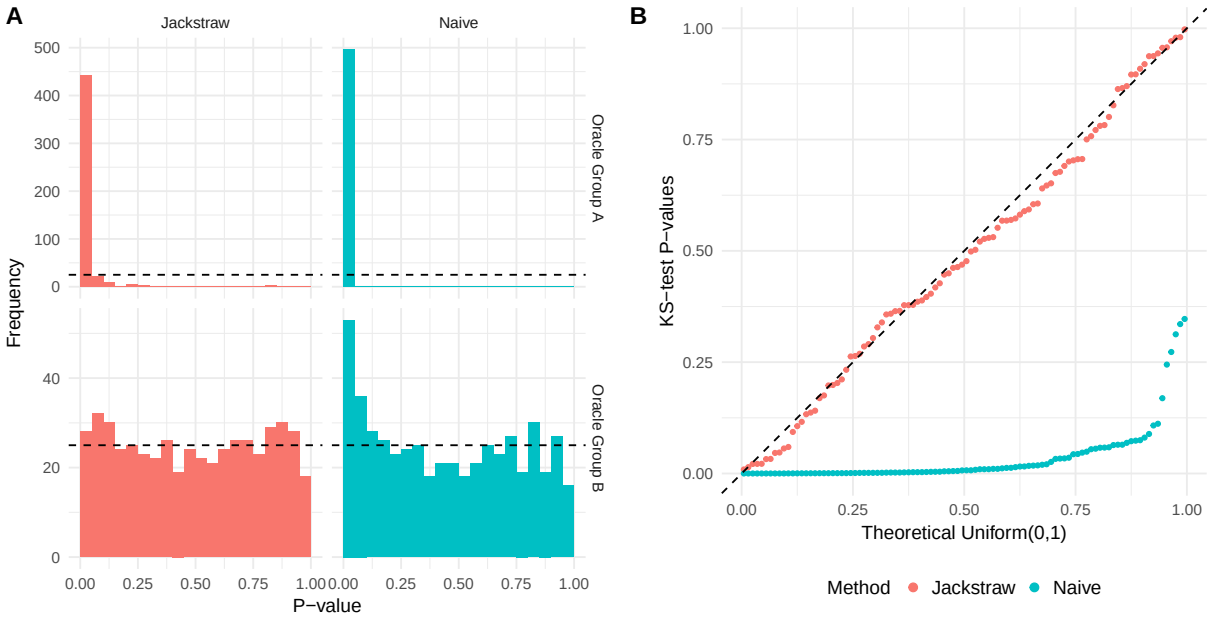

Figure S2: Simulation studies using $\sigma^2 = 5$. The jackstraw tests ($s = 100$ and $B = 5000$) or the naive tests are applied without using any information from simulation. *(a)* P-values are shown stratified by *Oracle Groups*, where the naive tests result in an anti-conservative bias. The uniformity of null p-values corresponding to *Oracle Group B* is examined by KS tests, which are independently repeated 100 times. *(b)* The total of 100 independent simulation studies are conducted, and 100 KS-test p-values are plotted against the Uniform(0,1) distribution. The proposed jackstraw tests meet the joint null criterion with a double KS test p-value of 0.81, whereas the naive tests are highly anti-conservative with a double KS test p-value of $< 2.2 \times 10^{-16}$.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

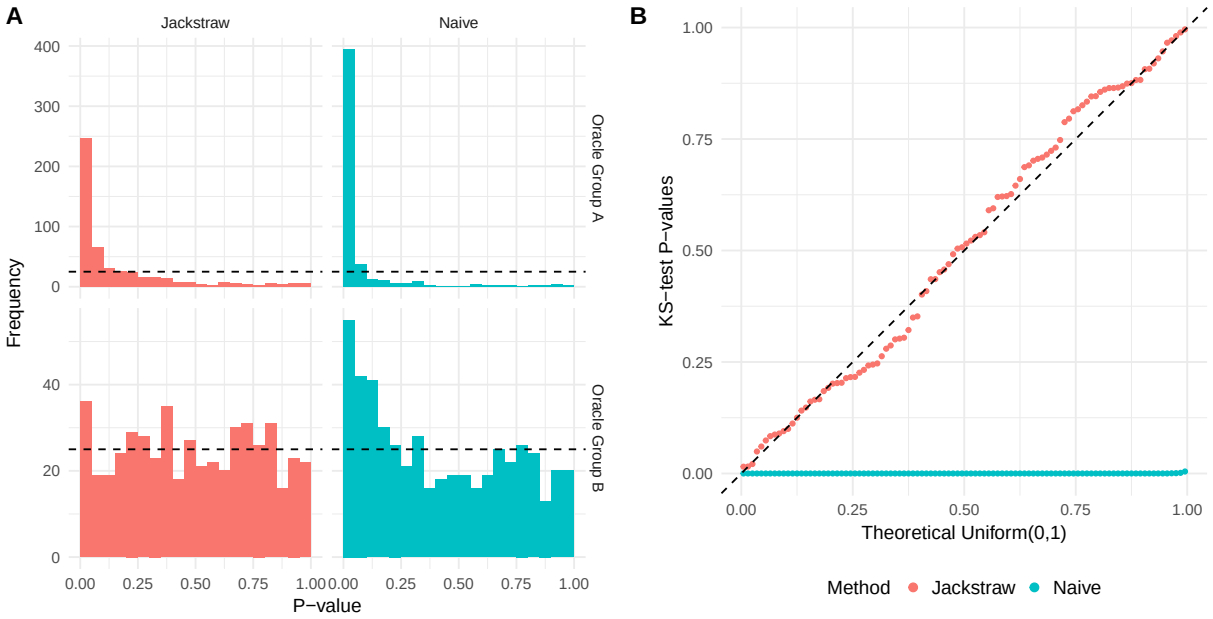

Figure S3: Simulation studies using $\sigma^2 = 15$. The jackstraw tests ($s = 100$ and $B = 5000$) or the naive tests are applied without using any information from simulation. *(a)* P-values are shown stratified by *Oracle Groups*, where the naive tests result in an anti-conservative bias. The uniformity of null p-values corresponding to *Oracle Group B* is examined by KS tests, which are independently repeated 100 times. *(b)* The total of 100 independent simulation studies are conducted, and 100 KS-test p-values are plotted against the Uniform(0,1) distribution. The proposed jackstraw tests meet the joint null criterion with a double KS test p-value of 0.67, whereas the naive tests are highly anti-conservative with a double KS test p-value of $< 2.2 \times 10^{-16}$.

bioRxiv preprint doi: https://doi.org/10.1101/248633; this version posted February 23, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

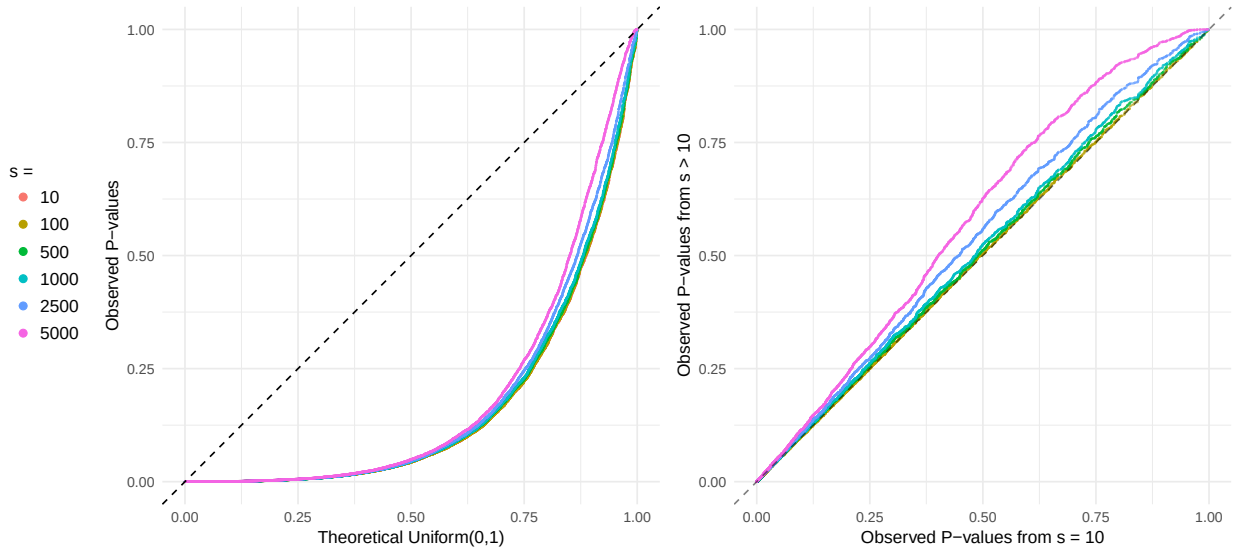

Figure S4: Comparison of the jackstraw p-values for an increasing choice of s. For the cell cycle gene expression data, the jackstraw tests are carried out with the number of synthetic null features s varying from 10 to 5000 while keeping the same number of total null statistics $s \times B = 5 \times 10^6$. On the left panel, all different sets of p-values are compared against Uniform(0,1), where as on the right panel, p-values from the jackstraw tests with $s = 10$ are used as a base line. Increasing s results in an increase in conservative behavior.