Heterogeneity-Preserving Discriminative Feature Selection for Subtype Discovery

Abdur Rahman M. A. Basher[1,2*], Caleb Hallinan[1*], and Kwonmoo Lee[1,2‡]

[1] Vascular Biology Program, Boston Children's Hospital, Boston, MA 02115, USA.
[2] Department of Surgery, Harvard Medical School, Boston, MA 02115, USA.
*Equal Contributors
‡To whom correspondence should be addressed

Abstract

Subtype discovery is crucial to disease diagnosis and targeted therapy as each cell or patient can display a wide range of responses to specific treatments. It is also vital to investigate the heterogeneity, or diverse variety, of disease states to better understand pathological processes. While there are many new opportunities for subtype discovery due to the explosion of single-cell data (single-cell RNA-seq, proteomic, and imaging datasets), selecting features for disease subtyping from high-dimensional datasets is challenging. Feature selection is the process of reducing the number of features for downstream computational analyses. Most feature selection algorithms are focused on identifying features to aid in the classification of known disease phenotypes. However, this effectively eliminates heterogeneous features and collapses the disease feature space, hindering valuable subtyping. Our work aimed to identify feature sets that preserve heterogeneity while maintaining the discrimination of known disease states. We initially applied a data-driven approach to determine the statistical characteristics of features essential in preserving heterogeneity by combining feature clustering and deep metric learning. This analysis revealed that features with a significant difference in interquartile range (IQR) between classes could contain critical subtype information. Utilizing this knowledge, we developed a statistical method, PHet (Preserving Heterogeneity), that performs recurrent sub-sampling differential analysis of IQR between classes to identify a minimal set of heterogeneity-preserving features while maximizing the quality of subtype clustering. Using public datasets of microarray and single-cell RNA-seq, we demonstrated that PHet effectively identifies disease subtypes and outperforms the previous outlier-based methods. In summary, our study provides a novel feature selection method for disease subtyping, which will enable not only personalized medicine but also a detailed understanding of the underlying heterogeneous mechanisms of diseases.

1 Introduction

Profiling subpopulations, or subtype discovery, of patient samples through analyzing molecular expression profiles (e.g., genes, transcriptomics, proteomics) is a fundamental research problem with opportunities in the field of cancer study [1]. A specific challenge in identifying subtypes of patients is to analyze the differences in the molecular patterns of patients under two experimental conditions, such as comparing cancer patients with healthy individuals. In cancer studies, it is well known that chromosomal translocations may result in the fusion of a gene's regulatory region with another gene that has oncogenic potential. Subsequently, this may lead to the over-expression of the oncogene and the disruption of normal cellular functions. This translocation is one of the many existing sources of biological heterogeneity of cancer cells [2]. However, not all cancer patients have the same chromosomal translocations, and some may have more extreme levels of oncogene expression than others, leading to the need for different treatment plans. Therefore, it is crucial to identify and classify cancer patients into subtypes that reflect their molecular and clinical characteristics, which can help to guide personalized and effective therapies [3, 4, 5].

Traditional differential expression methods, such as DESeq2 [6], edgeR [7], and LIMMA [8], assume the homogeneous nature of samples within conditions and were not designed to address genes expressed in a subset of cancer samples. Therefore, Tomlins and colleagues [9] developed a method called "cancer outlier profile analysis" (COPA) to detect outliers in cancer patients based on the median absolute deviation of a gene expression profile given the two experimental conditions. Since then, an active and growing body of works has expanded COPA to introduce more robust, statistical, and machine-learning approaches for identifying homogeneous patient subgroups in both experimental conditions [10, 11, 12, 13, 14, 15]. These approaches, including COPA, estimates statistics corresponding to aberrantly expressed genes between the two conditions and then ranks them. However, in the process of statistics computation, they often do not rely on differentially expressed genes, resulting in an admixture of samples of two conditions. Moreover, applications of these approaches are not focused on more complex problems, such as heterogeneous population reconstructions from

cells and discovering the changes in cell states among experimental conditions (e.g., two-time points in a cell developmental process and biological systems before and after treatment).

In parallel, there exists a line of research works, sharing similar goals, that attempt to solve the heterogeneity recovery problem under different experimental conditions using expression data, also known as omics data, in the context of differentially abundant cell subpopulation detection [16, 17, 18, 19, 20, 21]. These approaches, however, are primarily designed to detect cell composition changes among conditions and do not provide a mechanism to screen for highly expressed genes within each condition or differentially expressed genes among conditions. With high-throughput high-dimensional molecular data, sparseness and variability among gene expressions are still unresolved problems. It is imperative to analyze for genes that are useful for subpopulation or subtype detection, and these methods are not equipped to handle this problem. Importantly, a subtype detection algorithm, therefore, should possess the property to identify genes that have distinct patterns of expression or variability across different subtypes of a biological system. Broadly speaking, the genes, or features, are divided into three categories: 1)- Differentially Expressed (DE) features that are discriminative, 2)- Highly Variable (HV) or heterogeneous features that indicate the existence of outliers in a condition, and 3)- a composition of both or Hetero-Discriminative (HD).

To identify the key features that can distinguish different subtypes of diseases, we developed a computational framework that uses an iterative subsampling strategy to extract relevant features from omics data. Subsampling is a simple and effective way to detect outliers by analyzing a small subset of samples while capturing the three types of informative features: heterogeneous, discriminative, and hetero-discriminative. The main challenge with this approach is that subsampling may introduce redundant and unnecessary features for subtype detection. It is, therefore, important to select a minimum set of features that can help reconstruing subpopulations and detect outliers. We account against redundant feature collection by scoring each feature based on absolute differences in the interquartile range (IQR) [22] and the Fishers method [23] between samples of two conditions. The use of IQR has important implications in preserving heterogeneous features while the Fishers method can easily define discriminative features. Moreover, our method, which we call PHet (Preserving Heterogeneity), uses the Kolomorogov-Simirnov test between the two biological conditions to reduce scores for discriminative feature selection. By doing so, PHet attempts to minimize the number of informative features collected in order to maximize the quality of subtype detection. This min-max optimization is rarely addressed in the literature, and PHet is the first method to solve this problem in the context of subtype detection.

In the present study, we demonstrate the power of PHet by conducting a comprehensive benchmarking on 28 simulated, microarray, and single-cell transcriptomics datasets. We compared PHet with 24 existing computational methods using various performance metrics to ensure a fair and comprehensive assessment. PHet consistently outperformed the other baselines in identifying outliers and subtypes under binary experimental conditions. PHet also displayed superior scalability and robustness to high-dimensional data compared to alternative methods.

2 Results

2.1 Characterizing features in omics data extrapolated by deep-metric learning

Biotechnology has made remarkable strides in the last ten years, allowing researchers to collect and analyze various types of omics data from diverse sources, such as genomics (DNA sequences), transcriptomics (RNA expression), proteomics (protein abundance and interactions), metabolomics (metabolite levels and pathways), and more. This data can help better uncover the underlying mechanisms of diseases, identify subgroups of patients, and reveal the heterogeneity of cells, which can ultimately inform the development of personalized and targeted treatments. However, omics data also posess significant challenges for analysis, such as high dimensionality, noise, redundancy, and complexity. Many existing methods for analyzing omics data are not suitable or optimal for dealing with these challenges and may fail to capture the relevant structure or patterns in the data.

One common approach to address the high dimensionality of omics data is to perform dimensionality reduction, which projects the data onto a lower-dimensional space that preserves some aspects of the original data, such as variance or distance. Dimensionality reduction can facilitate data visualization and exploration, as well as downstream analysis tasks such as clustering or classification. However, dimensionality reduction also has some drawbacks, such as information loss and reduced interpretability. Another approach to deal with high-dimensional omics data is to perform feature selection, which selects a subset of features that are relevant for subtype discovery. Feature selection can reduce noise and redundancy in the data, improve computational efficiency, and enhance interpretability. Yet certain challenges in feature selection exist like choosing an appropriate criterion or metric for selecting features and balancing between different types of features that may have different roles or importance in the data. For example, when analyzing omics data for subtype detection, two types of features that may be relevant are highly variable (HV) features and differentially expressed (DE)

features. HV features are features that have high variability across samples, which may reflect biological diversity or heterogeneity in the data. DE features are features that have significant differences in expression or abundance between different experimental conditions or groups, such as disease status or treatment response. These features may reflect biological processes or pathways that are associated with the conditions or groups. However, HV features may not necessarily be DE features, and vice versa. Therefore, selecting only one type of feature may miss important information or signals in the data. We argue that a balanced selection of both HV and DE features should be considered in a supervised setting for subtype detection.

This idea requires an in-depth analysis to understand the characteristics of these heterogeneity-preserving discriminatory features. Artificial intelligence (AI), such as deep learning, can provide valuable insights into features from complex and diverse data sets. Moreover, these models are powerful in reducing dimensions that can preserve relationships among features to some extent. One such model, called deep-metric learning (DML) with triplet loss (Fig. 1(a)), can be applied to complex omics single cell transcriptomics data to obtain reduced dimensions of cells. However, instead of reducing the dimensions of cells, we take the opposite route and treat cells as variables. This is an intuitive way to perform feature selection as we shall notice. After applying DML, we reduce the dimensions of each feature for each experimental condition while simultaneously clustering features to learn a meaningful distance between clusters of feature embeddings. The triplet loss function is utilized to maximize the distance of clusters of features that are different while relocating similar clusters together. Therefore, if two features have a similar representation over cells in two different conditions they are pulled closer together while two features having different distributions over cells will be pushed further away. In doing this, we can obtain feature scores using the Euclidean distance between feature embeddings of the same feature but of different conditions, where features with higher scores are considered to be informative. We find that treating features in this way enables us to interpret and characterize features differently compared to previous methods. To give more context to this discussion, we used the Patel data [26], which is a single-cell RNA-seq dataset with 5 subtypes of primary glioblastomas, to compute the Euclidean distance between feature embeddings of different conditions to score features based on their informativeness. We illustrated this by comparing feature embeddings from DML for two experimental conditions in two dimensions (Figs 1(b)-(d)). We observed that features with high scores are far from the center and have large mean and IQR differences (Fig. 1(b)-(d)). We further classify these features into three types: discriminatory features (Fig. 1(f)), heterogeneous features (Fig. 1(g)), and Hetero-Discriminative (HD) features that combine both properties (Fig. 1(h)).

Since HD features comprise both HV and DE properties, it is important and necessary to capture these features. To demonstrate the advantages of HD features, we compared subtyping results using all, DE (using LIMMA [27]), HV (using dispersion-based selection [28]), and HD features on the Patel dataset [26] (Figs 1(i)-(l)). Results show that HD features have the smallest number of features but show the best subtyping outcomes among these feature classes. While differential expression analysis toolkits, such as LIMMA [27], can identify DE and HD features, they also select redundant or uninformative features. Dispersion-based analysis can locate HV features[28], but they can't locate discriminate properties of data in supervised settings. Since there exist no tools equipped to select HD features and based on the results using these novel features, we developed PHet, which is designed to extract the optimal combination of differentially expressed, HD, and to a lesser extent HV features to perform effective subtyping from omics data in supervised settings.

2.2 Overview of PHet

PHet is a method for finding informative features for the purpose of subtype discovery from omics expression data. Initially, the data is annotated with two conditions: control and case (e.g., responders and non-responders to drug treatment) and is preprocessed by removing low-quality samples and features and normalizing the expression values to have zero mean and unit variance. This processed data is then fed to the PHet pipeline which is composed of six major steps (Fig. 2 and Section 3). (i) Iterative subsampling to calculate p-values, using t-test or z-test, and absolute interquartile range differences (ΔIQR) for each feature between two subsets of control and case samples. The size of both subsets is fixed in an advance and is determined by the closest integer to the square root of $\min(n, m)$, where n and m correspond to the number of samples in control and case groups, respectively. The p-values measure the statistical significance of the difference in expression levels between the two groups, while the ΔIQR values indicate the differences in variability of the expression between each group. To ensure capturing sufficient diversity of features among samples that help identify subtypes, the subsampling procedure is repeated for a predefined number of iterations.

Immediately, the Fisher's combined probability test (ii) is then applied to summarize the collected p-values. The results from this test serve as prior information to calculating features statistics and ranking. (iii) A weighted features profile is constructed using the nonparametric two-sample Kolmogorov–Smirnov (KS) test between control and case samples to map each feature to its profile (Section 3). The KS test is used to identify discriminative features that exhibit maximum difference from long-run distribution (cumulative distribution functions) between control and case samples. At the same time, this test has the potential to eliminate redundant DE features by minimizing their scores. This is achieved by binning the p-values from the KS test

into a predefined number of intervals, where each interval is associated with a weight denoting a feature profile. By default, the number of bins is set to four. (iv) Feature statistics are estimated using a combination of the ΔIQR values, feature profiles, and Fisher's combined probability scores. During this step, the three types of features (DE, HV, and HD) will be identified and given high weights for the purpose of subtype detection while reducing scores for features that are less significant to discovering subpopulations. (v) Feature statistics are fitted using the gamma distribution, and features exceeding a user threshold (α) are trimmed. By default, α is set to 0.01. Finally, appropriate dimensionality reduction and clustering methods (vi) are performed on the reduced omics data, having only those selected features from step (v), to reconstruct data heterogeneity.

2.3 PHet can effectively detect changes in simulated samples

We evaluated the performance of 25 algorithms on two sets of simulated data (ten in total). These two sets of data correspond to the 'minority' and 'mixed' model schemes, as proposed by Campos and colleagues [15] in order to capture the heterogeneity structure of samples under the supervised settings. In the 'minority' model, only a small fraction of the case samples exhibit a change in a certain feature, while in the 'mixed' model, there is a substantial intra-group variation in both the case and control samples for that feature. Each data consists of 40 samples that were split evenly between control ($n = 20$) and case ($m = 20$) groups and $p = 1100$ features, of which 93 are differentially expressed between these two groups. We constructed 5 datasets corresponding to the minority model for which proportions of outliers in the case group varied from 5% (1 in 20) to 45% (9 in 20). Outliers were generated by adjusting the expressions of 100 randomly selected, including DE, features. The same approach was followed in constructing the five datasets for the mixed model, but outliers were evenly distributed between a subset of both case and control groups. Details about these datasets are outlined in Section 3. These simulated datasets are ideal to compare the performance of the ten algorithms under balanced situations, where the number of case and control samples is exactly the same. For quantitative analysis, the F1 metric was used to compare the top 100 predicted features by each method, with the top 100 true features. With the exception of DECO, features statistics from each method were fitted using the gamma distribution, and features with probabilities at a significance level $\alpha < 0.01$ were selected to show the number of informative features for each method (Section 3). For model-specific parameters, we followed the best practices suggested by the authors of each method to ensure that our experimental process is conducted in a fair and consistent manner without introducing any bias or error due to inappropriate parameter selection. An overall best-performing algorithm should attain high F1 scores under the two schemes while also predicting a minimum set of important features indicated by the gamma distribution.

Both PHet and PHet (ΔDispersion) retained a minimum set of significant features (at $\alpha < 0.01$) under all conditions for both minority and mixed changes (Figs 3(a)-(d)). Although, in specific simulated datasets, Dispersion (per condition), ΔDispersion and ΔDispersion+ΔMean methods achieve similar performance, they have significantly low F1 scores (< 0.4 on average) compared to PHet (0.85) for both model schemes, which shows that dispersion based feature selection methods require more statistical power to detect outliers. However, PHet (ΔDispersion) has a high F1 score despite this model utilizing dispersion-based features selection. This shows that iterative subsampling, indeed, has the ability to detect feature variation across small sets of samples.A majority of DE-based methods, DIDS, and DECO seem to perform on par with PHet, however, these methods, in particular DECO, retrieve more false positive features ($\alpha < 0.01$) than PHet. This implies that these methods may introduce some noise or redundancy in the feature screening process.

One of the challenges in biological studies and microarray data is the imbalanced sample size between groups. This can lead to biased estimates of differential expression, reduced statistical power, and wrong interpretation of the data [29]. Therefore, it is important to investigate the predictive capability of a subtype detection algorithm in retrieving a set of known markers in its top-ranked predicted features when the sample size is small in either the control or the case group. Here, we used HER2 data (Section 3), consisting of HER2 non-amplified samples ('control'; $n = 178$) admixed with HER2-amplified samples ('case'; $m = 10$), as described by de Ronde and colleagues [14]. In this dataset, we selected 20 differential markers that are located on ch17q12 or ch17q21 chromosome to compare the performance of 25 algorithms. We computed the F1 score for each algorithm based on how many of the true markers were among the top predicted features by an algorithm, varying the number of top features from 1 to 500. Despite not being among the top-ranked methods, PHet demonstrated a competitive performance with a mean F1 score above 35% (Fig. 3(e)). Furthermore, PHet is designed to infer a small set of features that are relevant for subtypes detection, unlike DE-based and other outliers detection algorithms (e.g., LSOSS and DIDS). Overall, PHet showed consistently lower sensitivity in selecting a set of significant differential features and showed consistently lower sensitivity to identifying subpopulations in samples under binary conditions.

2.4 Systematic benchmarking demonstrates PHet's superior performances in reconstructing heterogeneous subpopulations of cells and subtyping of patients

To demonstrate the performance of PHet on the cell subpopulation detection task, we benchmarked PHet against six single-cell transcriptomics (scRNA-seq) datasets (Supplementary Table 3). Cells in scRNA-seq datasets were divided into binary experimental conditions: control and case. An effective algorithm should be able to identify cell subsets while also able to preserve biological integrity using a minimum number of highly variable and expressive features. Prior to executing algorithms, we visualized each scRNA-seq data with all features using UMAP and observed that cells in Baron, Darmanis, and Patel data were randomly scattered with no indication of the existence of subclusters of cells (Supplementary Figs 15-20). These datasets reflect the nature of real-world scenarios where recovering cells heterogeneity would be the prime objective. On the contrary, by using all the features, we can already identify multiple cell subpopulations in Camp, Li, and Yan scRNA-seq data. These data give us an opportunity to evaluate subtype detection methods for selecting a minimal set of interpretable and meaningful features that can manage to retain the true cell subpopulations. In both cases, PHet and its variant PHet (ΔDispersion) were evaluated against 23 algorithms for cell subpopulation detection in these data using multiple performance metrics (e.g., F1, adjusted Rand index, homogeneity, and completeness).

Compared to outlier detection methods, PHet demonstrated its superiority on all scRNA-seq datasets in terms of F1 metric with a mean score reaching above 80% (Fig. 4(a)). F1 scores were obtained by comparing the top 100 differentially expressed features using LIMMA [27] with the top 100 discriminative features predicted by each method on scRNA-seq datasets (see Section 3 for details on metrics). Moreover, PHet was able to strike a balance between minimizing the number of predicted features for reconstructing the heterogeneity of cells (having an average of fewer than 450 features) and maximizing the discriminative or DE features for all single-cell transcriptomics datasets ($<$ 450 number of predicted features on average; Fig. 4(b)). Since the differentially expressed features were obtained using LIMMA, the average F1 score for LIMMA is 1. However, by fitting feature statistics using the gamma distribution, most DE-based methods failed to match PHet's performance, implying an improper estimation of feature statistics from these methods.

With respect to the average adjusted Rand index (ARI) metric, PHet achieved $>$ 10% gain over the second-best performing algorithm PHet (ΔDispersion) (Fig. 4(c)), indicating IQR-based approach is appropriate over dispersion based feature selection for better clustering quality and, hence, effectively recovering cell types. By exploiting the ground truth cell types corresponding to the scRNA-seq datasets, UMAP visualizations (using only predicted features) for each method also ensured that PHet attained a high level of biological integrity on all transcriptomics datasets. For example, among the 14 cell types in the Baron dataset (Supplementary Table 2), PHet was able to detect 736 features contributing to at least 7 clusters of cells. Previously, Baron and colleagues [30] showed that four cell types: alpha, beta, gamma, and delta, representing the endocrine cells, have inherent hierarchical relationships among themselves because they share similar expression profiles. Of these endocrine cell types, the alpha and gamma cells were seen to be proximally located. This finding suggests that these cells may have a close functional relationship in regulating glucose metabolism. The alpha cells secrete glucagon, which raises blood sugar levels, while the gamma cells secrete a pancreatic polypeptide, which inhibits both glucagon and insulin secretion. The co-localization of these cells may allow them to coordinate their actions and maintain glucose homeostasis [31]. Ductal and endothelial cells were remotely localized based on their expression, supporting the hierarchical integrity proposed by Baron and colleagues [30]. No other algorithms were able to accurately distinguish populations of cells in this data (Supplementary Fig. 15). On the basis of performance metrics using scRNA-seq datasets (Supplementary Fig. 14), PHet consistently produced high-quality clustering results for Baron [30], Darmanis [32], and Patel [26] data (Fig. 4(g); Supplementary Figs 15, 17, and 19). T-statistic, Wilcoxon, KS, LIMMA, ΔIQR and DECO methods were found to detect the five subsets for Camp [33] data (Supplementary Fig. 16) albeit at the cost of retrieving almost all feature sets (Fig. 4(d); Fig. 4(e); Supplementary Fig. 14) whereas PHet reaches the same conclusion with less predicted features ($\sim$ 313). Similar trends can be noticed for Li [34] and Yan [35] scRNA-seq datasets (Supplementary Figs 18 and 20). Of note, PHet utilized the same configurations across all scRNA-seq datasets.

We then extended our tests and evaluated the performance of PHet on patient subclass discovery using 11 well-known microarray gene expression datasets (Supplementary Table 2). Patient samples were apriori divided into control and case groups based on their clinical diagnosis. PHet ranked top in terms of an average ARI score $>$ 60%, substantially outperforming other baselines (Fig. 4(f)) while predicting fewer DE features ($>$ 130 on average; Figs 4(d) and (e)). For example, the MLL dataset [36] (in Supplementary Table 2) consists of 72 patients categorized into acute lymphoblastic leukemia (ALL), mixed-lineage leukemia (MLL), and acute myeloid leukemia (AML). By assigning ALL patients in the control group and MLL and AML in the case group, PHet recognized three clusters of patients with a near-perfect match (ARI = 88%) about their true sample types. None of the other algorithms that we compared it with were able to identify three subsets of patients (Supplementary Fig. 12), reassuring that PHet is a robust and reliable algorithm for detecting sample heterogeneity and subtypes. Similar observations can be interpreted for GSE1987 (Supplementary Fig. 11).

PHet did underperform on the Lung data (Supplementary Fig. 10), yet achieved a competitive performance on the remaining benchmark microarray datasets and over multiple metrics (Supplementary Fig. 2; Supplementary Figs 3-9, and 13). In both sets of data, PHet achieved superior clustering quality than other baseline methods. Moreover, our approach showed consistent results on various metrics and measures (e.g., sample size, number of features, and number of subtypes), while other methods did not deliver this level of quality and stability.

2.5 PHet retains cell lineages in HBECs and MTECs data

Preserving inherent complex relationships among cells is a fundamental challenge in cell differentiation analysis. To understand how cells differentiate, it is essential to capture the interactions and dependencies among different cell types with respect to changes in feature expression. We assessed PHet's capacity to detect subpopulations constituting trajectories of cells on two published scRNA-seq data of the conducting airway epithelium, which is a key component of the respiratory system [25]: 14,163 mouse tracheal epithelial cells (MTECs) and 2,970 primary human bronchial epithelial cells (HBECs). The MTECs data consists of injured and uninjured mice collected at 1, 2, 3, and 7 days after polidocanol-induced injury. Previous studies of lineage tracing and the regeneration process of post-injury have confirmed that basal cells differentiate into a heterogeneous population that contains the epithelial stem cells and other rare cell populations (PNECs, brush cells, and pulmonary ionocytes) [25, 24, 37]. We considered basal cells, for both datasets, as the control group while the remaining cell types were assembled under the case group (Supplementary Table 4). Compared with using annotated markers [25] (Figs 5(a) and (g)), PHet, using its predicted features, identified two distinct cell clusters representing basal and secretory cells for both HBECs and MTECs data (Figs 5(b) and (h)). Clusters representing rare cell populations (including brush and PNECs, pulmonary ionocytes, and SLC16A7) were also detected on UMAP visualizations.

To determine the extent to which PHet's predicted heterogeneous and discriminative features establish lineages of populations, we performed SPRING visualizations [38]. SPRING is a graph-based method that constructs a cell similarity network based on their expression profiles, subsequently, facilitating the process to reveal the structure and dynamics of cell populations. We used the known annotated markers and PHet's predicted features to visualize HBECs and MTECs data. This allowed us to examine which approaches is better in capturing the differentiation trajectories of cell populations. Although the markers and PHet's features displayed similar cell differentiation processes, PHet exclusively showed two distinct lineages for the HBECs spanning basal-to-luminal differentiation, including rare cells (Fig. 5(c)-(d)). By analyzing the metadata of HBECs, we observed that a particular lineage corresponds to a subset of donors (Fig. 5(e)). For example, cells in the black cluster were associated with a subset of donors (donor 1) that are enriched in cytokeratin genes KRT4 and KRT13. These cells may serve as progenitors to luminal differentiation. Moreover, the relative abundance of cell populations shows that the secretory cells are overrepresented in the red cluster (donors 2 and 3) and have a higher expression of the CYP2F1 gene in comparison to cells in the black cluster (Fig. 5(f), Supplementary Fig. 21 (g)). The CYP2F1 gene encodes a cytochrome P450 enzyme that is involved in the metabolism of xenobiotics and endogenous compounds which are major roles for the secretory club cells [39]. We also identified several features with diverse functions that were differentially expressed between the clusters, including a member of a BPI fold protein (BPIFA1) and KRT4 (Supplementary Fig. 21 (g)).

UMAP visualizations of the MTECs data using markers and PHet's features displayed similar observations regarding subpopulations of cells (Figs 5(g)-(h)). Similar to HBECs analysis, we applied SPRING and clearly identified two distinct epithelial cell lineages in MTECs (Figs 5(i)-(j)). However, the SPRING plot based on PHet's features offered more transparent clustering results than the one based on markers, suggesting that PHet's features capture more information about the cell states and transitions. We then characterized cell populations in these two trajectories (Fig. 5(k)) using cell-specific gene signatures. The bottom cluster of cells was enriched with basal cells while the top cluster contained a significant proportion of cycling basal cells ($> 20\%$) (Fig. 5(l)). This suggests that these cells may have a regenerative function and respond to injury by proliferating and differentiating into other cell types (Supplementary Figs 22(g)-(j); Supplementary Fig. 24(a)). Transitional cells that uniquely express KRT4 and KRT13 genes were also found abundant in the top cluster. These cells may represent an intermediate population between tracheal basal stem cells and differentiated secretory cells [25, 24, 40]. We further examined the heterogeneity of basal cells in the murine trachea corresponding to two clusters. We used SCANPY [41], an unsupervised method, to cluster basal cells based on the features detected by PHet. This approach enabled us to characterize four clusters of basal cells with distinct gene expression profiles and biological functions (Supplementary Fig. 23). The canonical basal cell markers, including TPR63 (tumor protein P63) and the cytokeratin 5 (KRT5) were highly expressed in Basal-1 and Basal-3 clusters, while cells in Basal-2 and Basal-4 showed enrichment for SCGB3A2 and the serpin family (e.g. TSPAN1), respectively, suggesting that they represent separate basal cells transitioning to a luminal secretory phenotype (Supplementary Figs 24(b)-(c)). These results support the previous findings [25, 42], confirming that PHet can detect cell subpopulations and its predicted features can accurately infer the developmental trajectories of cell differentiation. We note that in contrast to PHet's results, markers and

dispersion-based approaches were suboptimal in detecting these four basal subsets (Supplementary Figs. 25 and 26).

3 Discussion

Subtypes detection in the ever-increasing number of omics expression datasets, involving experimental or biological conditions, is important to explore tissue heterogeneity, discover cellular differentiation trajectories, and reveal the molecular signatures of different biological states or phenotypes associated with complex diseases (e.g., cancer and diabetes). Inaccurate subtypes detection, therefore, may have detrimental impacts on clinical decision-making, developing targeted therapeutics, and treatment planning for patients. However, locating molecular signatures in high-dimensional omics data impose several challenges associated with noise, sparseness, heterogeneity, or confounding factors. Therefore, a preliminary and essential step is to diagnose features that are linked to a specific subtype. In the supervised setting, where the experimental conditions are specified, features can be compartmented into three classes: discriminative, heterogeneous, and hetero-discriminative. Discriminative features are those that have different expression levels across experimental conditions. Heterogeneous features are those that have high variability within each condition. Hetero-discriminative features are those that are both discriminative and heterogeneous. These features play a pivotal role in capturing the heterogeneity and diversity of subtypes between/within each condition.

Several approaches have been developed to extract relevant features from data, such as DIDS [14] and DECO [15]. However, these methods have substantial limitations, such as focusing on a specific type of feature, obtaining redundant features, or failing to capture subtype related features. PHet overcomes these drawbacks by using recurrent subsampling and statistical tests. PHet assigns scores to features based on their heterogeneity and discriminability across experimental conditions and filters out irrelevant features by fitting a gamma distribution to the scores. The resulting features can then be used to cluster the data into subtypes that reflect the underlying biological variation. Based on case studies, PHet, indeed, retained a minimum set of features while ensuring optimal clustering quality for subtypes detection.

We note that the current framework also works for integrating with downstream differential abundance analysis tools, such as Milo [18]. Specifically, the selected features may be applied to track cellular changes in multiple experimental conditions. Apart from binary experimental conditions, PHet, as a flexible and versatile framework, can also be extended to multiple conditions (e.g., basal, secretory, and ciliated cells), over different omics measurements (e.g., single-cell RNA-seq, proteomics, or metabolomics), and different experimental designs (e.g., time-series or multi-factorial experiments). For optimum subtypes detection using PHet, the data should be batch-corrected and prepossessed beforehand, and the framework is not designed to account for confounding factors or artifact noise that may affect the expression measurements [43, 44, 45, 46].

The development of comprehensive atlas datasets for various organisms and tissues (e.g., Human Cell Atlas [47] and the Mouse Cell Atlas [48]) has enabled the discovery and characterization of new cell types [49, 50, 51]. These studies provide novel insights into the diversity and function of cellular phenotypes across different biological contexts and conditions and investigate how cell populations change in disease conditions [52, 53]. We envision that PHet will facilitate researchers to discover unknown cancer types and reveal the cellular diversity of healthy and diseased tissues using large-scale omics expression data, in particular, at single-cell resolution.

References

[1] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for precision medicine. *N Engl J Med*, 366(6):489–491, 2012.

[2] Marc A Hogenbirk, Marinus R Heideman, Iris de Rink, Arno Velds, Ron M Kerkhoven, Lodewyk FA Wessels, and Heinz Jacobs. Defining chromosomal translocation risks in cancer. *Proceedings of the National Academy of Sciences*, 113(26):E3649–E3656, 2016.

[3] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.

[4] Yasmin A Lyons, Sherry Y Wu, Willem W Overwijk, Keith A Baggerly, and Anil K Sood. Immune cell profiling in cancer: molecular approaches to cell-specific identification. *NPJ precision oncology*, 1(1):26, 2017.

[5] Rosie Fisher, Lazos Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[6] 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):1–21, 2014.

[7] 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.

[8] 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.

[9] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *science*, 310(5748):644–648, 2005.

[10] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8, 2007.

[11] Baolin Wu. Cancer outlier differential gene expression detection. *Biostatistics*, 8(3):566–575, 2007.

[12] Heng Lian. Most: detecting cancer differential gene expression. *Biostatistics*, 9(3):411–418, 2008.

[13] Yupeng Wang and Romdhane Rekaya. Lsoss: detection of cancer outlier differential gene expression. *Biomarker insights*, 5:BMI–S5175, 2010.

[14] Jorma J de Ronde, Guillem Rigaill, Sven Rottenberg, Sjoerd Rodenhuis, and Lodewyk FA Wessels. Identifying subgroup markers in heterogeneous populations. *Nucleic acids research*, 41(21):e200–e200, 2013.

[15] Francisco J Campos-Laborie, Alberto Risueño, M Ortiz-Estévez, B Rosón-Burgo, Conrad Droste, Celia Fontanillo, Remco Loos, Jose Manuel Sanchez-Santos, MW Trotter, and Javier De Las Rivas. Deco: decompose heterogeneous population cohorts for patient stratification and discovery of sample biomarkers using omic data profiling. *Bioinformatics*, 35(19):3651–3662, 2019.

[16] Xin Gao, Deqing Hu, Madelaine Gogol, and Hua Li. Clustermap: compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics*, 35(17):3038–3045, 2019.

[17] Michael A Skinnider, Jordan W Squair, Claudia Kathe, Mark A Anderson, Matthieu Gautier, Kaya JE Matson, Marco Milano, Thomas H Hutson, Quentin Barraud, Aaron A Phillips, et al. Cell type prioritization in single-cell data. *Nature biotechnology*, 39(1):30–34, 2021.

[18] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022.

[19] Jun Zhao, Ariel Jaffe, Henry Li, Ofir Lindenbaum, Esen Sefik, Ruaidhrí Jackson, Xiuyuan Cheng, Richard A Flavell, and Yuval Kluger. Detection of differentially abundant cell subpopulations in scrna-seq data. *Proceedings of the National Academy of Sciences*, 118(22):e2100293118, 2021.

[20] Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications*, 11(1):4318, 2020.

[21] Chanwoo Kim, Hanbin Lee, Juhee Jeong, Keehoon Jung, and Buhm Han. Marcopolo: a method to discover differentially expressed genes in single-cell rna-seq data without depending on prior clustering. *Nucleic acids research*, 50(12):e71–e71, 2022.

[22] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer, 2005.

[23] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press, 1925.

[24] Daniel T Montoro, Adam L Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, et al. A revised airway epithelial hierarchy includes cftr-expressing ionocytes. *Nature*, 560(7718):319–324, 2018.

[25] Lindsey W Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M Klein, and Aron B Jaffe. A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature*, 560(7718):377–381, 2018.

[26] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[27] 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.

[28] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[29] Chunming Xu and Scott A Jackson. Machine learning and complex biological data, 2019.

[30] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[31] Pia V Röder, Bingbing Wu, Yixian Liu, and Weiping Han. Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine*, 48(3):e219–e219, 2016.

[32] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

[33] J Gray Camp, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

[34] Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708–718, 2017.

[35] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, 2013.

[36] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47, 2002.

[37] Marie Deprez, Laure-Emmanuelle Zaragosi, Marin Truchi, Christophe Becavin, Sandra Ruiz García, Marie-Jeanne Arguel, Magali Plaisant, Virginie Magnone, Kevin Lebrigand, Sophie Abelanet, et al. A single-cell atlas of the human healthy airways. *American journal of respiratory and critical care medicine*, 202(12):1636–1645, 2020.

[38] Caleb Weinreb, Samuel Wolock, and Allon M Klein. Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, 2018.

[39] Jordan D Davis and Tomasz P Wypych. Cellular and functional heterogeneity of the airway epithelium. *Mucosal immunology*, 14(5):978–990, 2021.

[40] Richard J Hewitt and Clare M Lloyd. Regulation of immune responses by the airway epithelial cell landscape. *Nature Reviews Immunology*, 21(6):347–362, 2021.

[41] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[42] Gianni Carraro, Justin Langerman, Shan Sabri, Zareeb Lorenzana, Arunima Purkayastha, Guangzhu Zhang, Bindu Konda, Cody J Aros, Ben A Calvert, Aleks Szymaniak, et al. Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. *Nature medicine*, 27(5):806–814, 2021.

[43] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

[44] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542, 2018.

[45] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

[46] 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.

[47] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

[48] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[49] Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, 17(12):1200–1206, 2020.

[50] Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B Altman. Leveraging the cell ontology to classify unseen cell types. *Nature communications*, 12(1):5556, 2021.

[51] Jan Dohmen, Artem Baranovskii, Jonathan Ronen, Bora Uyar, Vedran Franke, and Altuna Akalin. Identifying tumor cells at the single-cell level using machine learning. *Genome Biology*, 23(1):123, 2022.

[52] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

[53] Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nature methods*, 14(3):297–301, 2017.

[54] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[55] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14(1):1–13, 2014.

[56] Ilker Etikan and Kabiru Bala. Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, 5(6):00149, 2017.

[57] Erich Leo Lehmann and Howard J D'Abrera. *Nonparametrics: statistical methods based on ranks*. Holden-day, 1975.

[58] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[59] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[60] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[62] Harald Cramér. *Mathematical methods of statistics*, volume 26. Princeton university press, 1999.

[63] Somnath Datta and Glen A Satten. Rank-sum tests for clustered data. *Journal of the American Statistical Association*, 100(471):908–915, 2005.

[64] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[65] James W MacDonald and Debashis Ghosh. Copa-cancer outlier profile analysis. *Bioinformatics*, 22(23):2950–2951, 2006.

[66] 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.

[67] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[68] Daniel R Rhodes, Jianjun Yu, Kalyan Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pander, and Arul M Chinnaiyan. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, 2004.

[69] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson Jr, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001.

[70] Federico Marini, Jan Linke, and Harald Binder. ideal: an R/Bioconductor package for interactive differential expression analysis. January 2020.

[71] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[72] Sven Rottenberg, Marieke A Vollebergh, Bas de Hoon, Jorma de Ronde, Philip C Schouten, Ariena Kersbergen, Serge AL Zander, Marina Pajic, Janneke E Jaspers, Martijn Jonkers, et al. Impact of intertumoral heterogeneity on predicting chemotherapy response of brca1-deficient mammary tumorsdrug response prediction of brca1-deficient tumors. *Cancer research*, 72(9):2350–2361, 2012.

[73] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[74] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[75] Doulaye Dembélé. A flexible microarray data simulation model. *Microarrays*, 2(2):115–130, 2013.

Methods

Throughout this paper, the default vector is considered to be a column vector and is represented by a boldface lowercase letter (e.g., $\mathbf{x}$) while matrices are denoted by boldface uppercase letters (e.g., $\mathbf{X}$). If a subscript letter i is attached to a matrix (e.g., $\mathbf{X}_i$), it indicates the i-th row of $\mathbf{X}$, which is a row vector. A subscript character to a vector (e.g., $\mathbf{x}_i$) denotes an i-th cell of $\mathbf{x}$. Occasional superscript, $\mathbf{x}^{(i)}$, suggests a sample or an iteration index. With these notations in mind, let us describe the data and the PHet framework.

Data preprocessing

Omics data, denoted by $\mathbf{M} \in \mathbb{R}^{n+m,p}$, takes expression profile (e.g., gene) as the input and was categorized into *control* (n samples) and *case* (m samples) experimental groups, where both groups share the same set of features ($p \in \mathbb{N}$). Formally, let the data matrix $\mathbf{X} \subset \mathbf{M}$ be control samples of size $n \in \mathbb{N}$ where each element in $\mathbf{X}_{i,j}$ represents the expression value for sample $i \in \{1, ..., n\}$ and feature $j \in \{1, ..., p\}$ while the data matrix $\mathbf{Y} \subset \mathbf{M}$ be a set of case samples of size $m \in \mathbb{N}$ where $\mathbf{Y}_{k,j}$ represents the expression value for sample $k \in \{1, ..., m\}$ and feature $j \in \{1, ..., p\}$. We filtered out low-quality samples and features in the omics data. Specifically, features

expressed (as non-zero) in more than 1% of samples and samples expressed as non-zero in more than 1% of features were retained. All the data were log-transformed. We did not scale datasets to unit variance and zero mean, as scaling is an intrinsic property of methods. No other additional preprocessing and normalization was performed on the data. Note that in this manuscript, samples and cells were used interchangeably.

PHet framework

The PHet framework is a method for finding relevant features that distinguish the subpopulations of samples in a control-case study. PHet accomplishes this goal in a pipeline procedure that is composed of six stages (see Fig. 2): (1)- iterative sampling, (2)- fisher's combined probability test, (3)- recovering feature profiles, (4)- estimating features statistics, (5)- feature ranking, and 6)- dimensionality reduction and clustering.

Iterative subsampling

Given an annotated omics dataset ($\mathbf{M}$), PHet computes the following: i)- p-values and ii)- absolute differences in the interquartile range (ΔIQR) between control and case samples to collect feature signals. A p-value indicates the statistical significance of how likely it is that any observed difference, for a given feature, between the two sample groups is due to chance. Being a probability, the p-value can take any value within the $[0, 1]$ range. For a feature, a low p-value means that the difference is unlikely to be explained by chance, and therefore suggests that the feature is differentially expressed [54], whereas a p-value close to 1 suggests no difference between the groups other than due to chance (or statistically insignificant). Two main factors contribute to the estimation of p-values: (1) the sample size of both groups, which affects the power of the test; and (2) the spread of samples among the two groups, which reflects the variability of the data. Generally, having an uneven sample size between the two groups can introduce bias and affect the accuracy of the p-value calculation. Therefore, instead of using all samples from control and case groups, PHet randomly selects subsets of samples from each group with a fixed size. The size of both subsets is determined by the closest integer to the square root of $\min(n, m)$, which is considered to be the optimum subset size. This process, called "iterative subsampling", is repeated for a predefined number of iterations $t \in \mathbb{N}$ to obtain a distribution of p-values for each feature, which are stored in a matrix $\mathbf{P} \in \mathbb{R}^{p \times t}$ to be used for identifying differentially expressed (DE) or discriminative features. PHet employs the two-sample Student t-test (or Z-test if the subset size exceeds 30) as a test statistic to compute p-values.

However, p-values alone are not sufficient to rank the features by their biological relevance, as they do not reflect changes in expression for subtype detection. As pointed out, an omics dataset may exhibit data dispersion which impairs p-value calculation, even with the iterative subsampling process. Therefore, PHet employs another measurement, the absolute difference of interquartile range (ΔIQR) between each condition. The IQR, for a feature, is defined as the difference between the first and third quartiles of the data, i.e., $q_1 - q_3$, where q_1 and q_3 are the 25th and 75th percentiles of the data [55]. A high IQR, for a feature, indicates a large dispersion of expression values within each condition, which may suggest a heterogeneous conserved feature or a highly variable feature (HV). On the other hand, a higher ΔIQR indicates a large dispersion of expression values between samples of the two groups. It is probable that a ΔIQR value, for a feature, may be close to 0 while the IQR values for each condition may be high. This will affect the ranking of the corresponding feature. Fortunately, during the subsampling process, every sample in the omics data will have an equal (non-zero) opportunity to be included in the subset [56], enabling a selection of highly variable features using ΔIQR scores. Moreover, this process can substantially reduce the redundant and unnecessary features that can not be achieved using the IQR test. The formula for ΔIQR of a subset s is:

$$\Delta\text{IQR}^s = |\text{IQR}(\mathbf{X}^s) - \text{IQR}(\mathbf{Y}^s)| \tag{1}$$

Initially, the ΔIQR values are stored in $\mathbf{R} \in \mathbb{R}^{p \times t}$. Then, the max operation for each feature is applied to collapse values to a vector of size p, i.e., $\mathbf{R} \in \mathbb{R}^p$. Upon accomplishing this step, two outputs will be delivered: the p-values ($\mathbf{P}$) and the absolute ΔIQR values ($\mathbf{R}$). They both serve as inputs in the PHet pipeline. Note that instead of computing ΔIQR values, PHet has the functionality to collect normalized dispersion analysis or ΔDispersion in an iterative subsampling manner. See Section 3 for ΔDispersion.

Fisher's combined probability test

In the second step of the pipeline, PHet summarizes the obtained p-values from the previous step into test statistics using Fisher's method for combined probability [23]:

$$\mathbf{I} = -2 \times \text{diag}(\log \mathbf{P} \log \mathbf{P}^{\intercal}) \tag{2}$$

where diag corresponds the matrix-to-vector diagonal operation. The Fisher's scores, represented by $\mathbf{I} \in \mathbb{R}^p$, provide a measure of differential consistency of probabilities across t number of independent tests or iterations.

When p-values for a feature, across t rounds, tend to be small (or non-consistent), the corresponding Fisher score will be large, suggesting that this feature is statistically significant, and vice versa. This fact can be used to rank and eliminate non-informative features, resulting in the reduction of high dimensional omics data size to a smaller set of discriminative features. The final output of this step is Fisher's combined probability scores ($\mathbf{I}$).

Constructing weighted feature profiles

In this step, PHet groups features into corresponding profiles. At first, PHet estimates the p-values of features using a nonparametric two-sample Kolmogorov–Smirnov (KS) test between control and case samples. The KS test measures the maximum difference, for a feature, between the two cumulative distributions without considering the type of distribution of control and case samples [57], which determines if both samples belong to the same population or not. To address the uneven sample size, PHet applies a random permutation of both sample groups and employs a fixed sample size, determined by the closest integer to the square root of $\min(n, m)$. Then, PHet compartments (also known as binning) p-values into a predefined number of intervals $b \in \mathbb{N}$. Each interval or bin is associated with a weight denoting a feature profile:

$$\mathbf{w} = \{w_i : 0 \leq w_i \leq 1\}_{i \in \{1,\ldots,b\}} \text{ and } \mathbf{w}^{\intercal} 1 = 1 \tag{3}$$

By default, the number of bins is set to four. Finally, PHet obtains the weighted feature profile matrix as:

$$\mathbf{O} = \widehat{\mathbf{O}}\mathbf{w} \tag{4}$$

where $\mathbf{O} \in \mathbb{R}^p$ and $\widehat{\mathbf{O}} \in \{0, 1\}^{p \times b}$. Each row in $\widehat{\mathbf{O}}$ represents a one-hot encoding vector where all cells are set to 0 except a single 1 in a cell used to identify the feature profile. The final output of this step is the weighted feature profile matrix ($\mathbf{O}$).

Estimating feature statistics

In this step, PHet computes overall feature scores, referred to as the $\mathbf{H}$ statistics, using the outputs from previous steps: (1)- the ΔIQR scores ($\mathbf{R}$), which highlights the differences in the IQR of two samples to collect a set of heterogeneity preserving features (or Dispersion), (2)- the Fisher's scores ($\mathbf{I}$), which indicates how discriminant each feature is (DE), and (3)- the weighted feature profile matrix ($\mathbf{O}$), which links features to its profile:

$$\mathbf{H} = \mathbf{R} + (\mathbf{I} \odot \mathbf{O}) \tag{5}$$

where $\odot$ corresponds to the Hadamard (element-wise) product. As can be seen that Fisher's scores are adjusted by the weighted feature profile matrix to prevent the overestimation of discriminative features. The matrix $\mathbf{H} \in \mathbb{R}^p$ holds the final statistics, where a high feature score indicates its importance for subtype detection.

Features selection

The statistics in $\mathbf{H}$, from the previous step, can be used to rank and select discriminative (DE), heterogeneous (HV), and hetero-discriminative (HD) features. By using a histogram, one can show that scores follow a gamma distribution with shape γ, scale β, and location μ parameters, i.e., gamma(μ, γ, β). As such, we fit a gamma distribution to estimate these parameters after scaling features statistics to unit variance and zero means; then, features with probabilities exceeding a user-defined $\alpha \in [0, 1]$ (set to 0.01 by default) are removed from the analysis. The output of this step is a set of discriminative and hetero-discriminative features that are important to discriminate and detect subpopulations in samples.

Clustering and Dimensionality reduction

In this step, PHet employs the identified DE, HV, and HD features to reduce the omics data and performs clustering to locate subtypes in sample groups. For single-cell data, spectral clustering [58] is a convenient approach that groups a subset of densely connected cells, representing subpopulations, on a graph. Then, PHet compresses the reduced expression matrix into a low-dimensional representation using an appropriate dimensionality reduction algorithm to optimally preserve the structure and distribution of the data in two or three dimensions. The default dimensionality reduction algorithm in PHet is UMAP [59] which is arguably the best method to collapse high-dimensional features in omics data other than PCA [60] and t-SNE [61]. The UMAP algorithm uses the stochastic gradient descent approach to minimize the difference between distances among samples i)- in a higher-dimensional space and ii)- in their projected lower-dimensional space. PHet also provides outputs in a tabular format that can be used in SPRING application [38] to explore cell trajectories. This step is common to all subpopulation detection algorithms (Section 3).

Benchmark evaluation compared to existing tools

The performance of the PHet algorithm was evaluated in comparison with four discriminative features analysis tools and their variants: the standard Student t-test [62], Wilcoxon rank-sum test [63], Kolmogorov–Smirnov test [64], and LIMMA [27]; two heterogeneous feature identification methods and their variants: Dispersion [28] and IQR [22]; and seven outliers detection algorithms: COPA [65], OS [10], ORT [11], MOST [12], LSOSS [13], DIDS [14], and DECO [15] (Table 1). For all these methods, features were selected using either p-values or their scores with a cutoff threshold determined by the gamma distribution. Descriptions of each method are included below.

Student t-test (t-statistic). The Student t-test or t-statistic is a conventional statistical test that compares the mean difference between two groups of samples for each feature [62]. This test assumes that the data follows a normal distribution and that the variance is equal across the two groups. The t-statistic has been used to identify features that are differentially expressed between the two groups, which can provide insights into the biological mechanisms underlying the disease or the response to treatment. The two-class t-statistic for a feature j is defined by:

$$\text{statistic} = \frac{\overline{\mathbf{Y}}_{:,j} - \overline{\mathbf{X}}_{:,j}}{s_j} \tag{6}$$

where $\overline{\mathbf{Y}}_{:,j}$ and $\overline{\mathbf{X}}_{:,j}$ are the mean expression values of case and control samples, respectively. s_j is the pooled standard error estimate defined as:

$$s_j^2 = \frac{\sum_{1\leq i\leq n}(\mathbf{X}_{i,j} - \overline{\mathbf{X}}_{:,j})^2 + \sum_{1\leq i\leq m}(\mathbf{Y}_{i,j} - \overline{\mathbf{Y}}_{:,j})^2}{n+m-2} \tag{7}$$

A high statistic means that the feature j has a large difference in mean expression between the groups, while a low statistic means that the feature j has a similar expression in both groups. To select important features, we followed two approaches: 1)- a p-value cutoff based on false discovery rate (FDR) at < 0.01 significance level using the Benjamini & Hochberg method [66] and 2)- fit a gamma distribution to statistics and then trim features at < 0.01 significance level. The first approach is termed as t-statistic while the later strategy is called t-statistic+Gamma. For both approaches, we applied a two-sided test using the 'SciPy' package [67].

Wilcoxon rank-sum test. This is a nonparametric statistical approach to test the null hypothesis that samples in control and case groups are drawn from distributions with equal medians, against the alternative that they are not. Although this test compares medians, it does not use the expression values of the medians to perform the comparison. For a small sample size in the range of 10 to 15 for both groups and a one-sided test for the control group, the Wilcoxon rank-sum statistic has the following formula for the feature j:

$$\begin{aligned} \text{statistic} &= \frac{W_j - \mu}{\sigma} \\ W_j &= \sum_{i\in n} \text{rank}(\mathbf{X}_{i,j}) \\ \mu &= \frac{n(n+m+1)}{2} \\ \sigma &= \sqrt{\frac{nm(n+m+1)}{12}} \end{aligned} \tag{8}$$

The approximated p-value for the Wilcoxon rank-sum statistic for the feature j can be pulled from the standard normal distribution to decide on the null hypothesis. A similar formula can be derived for the large sample size and two-sided test [63]. Having obtained the p-value for all features, the important features can be determined based on FDR at < 0.01 significance level using the Benjamini & Hochberg method [66]. Alternatively, we can fit a gamma distribution to statistics and then select features at < 0.01 significance level. We refer to the previous approach as Wilcoxon while for the later workflow, it is called Wilcoxon+Gamma. For both approaches, we applied a two-sided test using the 'SciPy' package [67]. It is worth mentioning that the Wilcoxon rank-sum test is more resilient to outliers and does not require assumptions about data distributions; however, this test is less effective when the sample size per condition is less than 8.

Kolmogorov–Smirnov (KS) test. This is a nonparametric two-sample test that is used to determine if samples from control and case groups belong to the same population or not [64]. The KS test, for a feature j, is based on the calculation of the maximum difference between the two empirical cumulative distributions of control and case samples [57]:

$$\text{statistic} = \sup_j |F(\mathbf{X}_{:,j}) - G(\mathbf{Y}_{:,j})| \tag{9}$$

where $F(.)$ and $G(.)$ are the empirical distribution functions of control and case samples, respectively, and sup is the supremum function. To collect important features, we follow a similar approach as with the t-statistic:

Method	Strategy	Down-regulated features	Outlier weighting	Subtype discovery	Multiple groups	Reference
t-statistic	Mean difference	✓	×	×	×	[62]
t-statistic+Gamma	Mean difference and the gamma distribution	✓	×	×	×	[62] and this work
Wilcoxon	Sum of the ranks	✓	×	×	×	[63]
Wilcoxon+Gamma	Sum of the ranks and the gamma distribution	✓	×	×	×	[63] and this work
KS	Empirical cumulative distribution function	✓	×	×	×	[64]
KS+Gamma	Empirical cumulative distribution function and the gamma distribution	✓	×	×	×	[64] and this work
LIMMA	Feature-wise linear models and empirical Bayes to obtain posterior variance estimators	✓	×	×	×	[27]
LIMMA+Gamma	Feature-wise linear models, empirical Bayes to obtain posterior variance estimators, and the gamma distribution	✓	×	×	×	[27] and this work
Dispersion (composite)	Normalized dispersion of all samples	–	–	✓	–	[28]
Dispersion (per condition)	Normalized dispersion of samples of each condition	–	–	✓	✓	[28] and this work
ΔDispersion	Differences of normalized dispersion between samples of each condition	–	–	✓	✓	[28] and this work
ΔDispersion+ΔMean	Differences of normalized dispersion and mean between samples of each condition	–	–	✓	✓	[28] and this work
IQR (composite)	Z-score and IQR of all samples	–	–	✓	–	[22] and this work
IQR (per condition)	Z-score and IQR of samples of each condition	–	–	✓	✓	[22] and this work
ΔIQR	Z-score and ΔIQR	–	–	✓	✓	[22] and this work
ΔIQR+ΔMean	Z-score, ΔIQR, and ΔMean	–	–	✓	✓	[22] and this work
COPA	Percentile MAD	×	×	×	×	[9]
OS	Quantile ordered & cut-off	✓	×	×	×	[10]
ORT	Robust t-statistic	×	×	×	×	[11]
MOST	Maximum ordered	✓	×	×	×	[12]
LSOSS	Least sum of ordered subset square t-statistic	✓	×	×	×	[13]
DIDS	Maximum value from control group	✓	✓	×	×	[14]
DECO	RDA and NSCA	✓	✓	✓	✓	[15]
PHet (ΔDispersion)	Log transform, iterative subsampling, Fisher's method, and ΔDispersion	✓	✓	✓	✓	This work
PHet	Z-score, iterative subsampling, Fisher's method, and ΔIQR	✓	✓	✓	✓	This work

Table 1: **Methods used in the benchmark.** MAD: median absolute deviation of expression values; RDA: recurrent differential analysis; and NSCA: non-symmetrical correspondence analysis.

1)- a p-value cutoff based on FDR at < 0.01 significance level using the Benjamini & Hochberg method [66] and 2)- fit a gamma distribution to statistics and then trim features at < 0.01 significance level. While the first workflow is referred to as KS, the latter strategy is called KS+Gamma. We employed a two-sided test using the 'SciPy' package [67] for these two strategies.

LIMMA. LIMMA (Linear Models for Microarray Data) [27] is a package designed to perform differential expression analysis for large-scale expression studies while remaining reasonably easy to utilize for simple experiments and small sample sizes. The input to LIMMA is expression data which can be log ratios, or

sometimes log-intensity values corresponding to microarrays technologies. Then, LIMMA attempts to fit a linear model for each feature to the expression data. This is followed by empirical Bayes and other shrinkage methods to make statistical conclusions more reliable. Similar to the t-statistic, we select features using: 1)- a p-value cutoff at < 0.01 significance level implemented inside the LIMMA package, termed as LIMMA method, and 2)- fit a gamma distribution to BH statistics, obtained from LIMMA, and then trim features at < 0.01 significance level, which is referred to as LIMMA+Gamma workflow.

LIMMA has been a popular choice for feature detection through differential expression analyses of microarray, RNA-seq, and other types of data. However, LIMMA assumes that the data follows a normal distribution, which may not be true for large-scale population studies. Furthermore, the increase in sample size raises the presence of potential outliers, which may violate the normality assumption and impact the p-value estimation and the false discovery rate control. As such, LIMMA may not be suitable for large-scale population studies.

Dispersion based Highly Variable Features. To identify features with high variability across data, Satija and colleagues [28] developed a method based on the normalized dispersion analysis. The method computes dispersion measures (variance/mean) for each feature from all samples and then groups feature into a fixed number of bins (20 by default) according to their mean expression. Within each bin, the dispersion measures of features are standardized to detect features that have higher variability than the average expression values in the bin. Usually, several hyperparameters are required to be set to pick the top highly variable features. However, in our work, we collect the normalized dispersion values for all features and fit a gamma distribution to these dispersion values, and then pick features at < 0.01 significance level. We apply this approach to a combination of control and case samples, hence, it is called Dispersion (composite).

Moreover, we extend this process and introduce three diverse strategies to collect statistics or dispersion values: 1)- Dispersion (per condition) in which the dispersion values for features for each condition are computed separately, 2)- ΔDispersion where absolute differences of dispersion values between samples of control and case groups for all features are calculated, and 3)- ΔDispersion+ΔMean where the feature-wise sum of both the absolute difference of the dispersion value and the absolute differences of the mean value between samples of two groups is obtained. For all these methodologies, we fit a gamma distribution to collected values and then pick features at < 0.01 significance level.

InterQuartile Range (IQR). This measurement allows for the detection of a set of heterogeneous conserved features. We define four variants of IQR to accumulate statistics: 1)- IQR (composite) which estimates the interquartile range of composite of the control and case samples for each feature, 2)- IQR (per condition) where the sum of interquartile range of samples for each condition for each feature is computed, 3)- ΔIQR where the absolute differences of interquartile range between samples of two groups for all features are estimated, and 4)- ΔIQR+ΔMean where sums of the absolute differences of interquartile range and the absolute differences of mean between samples of two groups for all features are computed. The formula for ΔIQR+ΔMean for a feature j is defined as:

$$\text{statistic} = |\text{IQR}(\mathbf{X}_{:,j}) - \text{IQR}(\mathbf{Y}_{:,j})| + |\overline{\mathbf{X}}_{:,j} - \overline{\mathbf{Y}}_{:,j}| \tag{10}$$

where $\overline{\mathbf{X}}_{:,j}$ and $\overline{\mathbf{Y}}_{:,j}$ represent the mean expression values for control and case samples, respectively, for the feature j. For all IQR-based methods, a precondition to computing scores is that the expression values for each feature should be normalized to unit variance and zero means.

COPA. While the t-statistic is an effective way to determine the differentially expressed features, it is limited in discovering the heterogeneity of samples. This is because the t-statistic computes the mean difference of features between control and case samples without providing a way to discern outliers or subtypes. Typically, outlier samples have high expression values for a set of features, also known as highly variable features, within a group. This is true, in particular, for cancer studies where mutations can either amplify or down-regulate feature expression in only a minority of cancer (or case) samples. To address these extreme values of feature expression in the case group, Tomlins and colleagues [9] proposed Cancer Outlier Profile Analysis (COPA) which is defined for a feature j as:

$$\text{statistic} = \frac{q_r(\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) - \text{MED}_j}{\text{MAD}_j} \tag{11}$$

where $q_r(.)$ is the rth percentile (user-defined) of the expression data and MED_j is the median expression value for all samples defined as:

$$\text{MED}_j = \text{median}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \tag{12}$$

and MAD_j is the median absolute deviation of expression values in all samples given by:

$$\text{MAD}_j = 1.4826 \times \text{median}(\{(\mathbf{X}_{i,j} - \text{MED}_j) : 1 \leq i \leq n\}, \{(\mathbf{Y}_{i,j} - \text{MED}_j) : 1 \leq i \leq m\}) \tag{13}$$

The COPA statistic utilizes the median and median absolute deviation to define outliers, where samples with high outlier scores for many features are considered to be outliers. COPA has been used to detect outlier profiles in the Oncomine database [68] and was able to locate multiple outlier features.

OS. A major limitation in calculating the COPA statistic is the *rth* percentile, which is a tuning parameter that depends on the data. To overcome this problem, Tibshirani and colleagues [10] proposed the Outlier Sum (OS) statistic. For a feature j, it is defined to be the sum of expressions in case samples that have values beyond $q_{75}(.) + \text{IQR}(.)$ for that feature as:

$$\text{statistic} = \frac{\sum_{\mathbf{Y}_{i,j} \in R_j} (\mathbf{Y}_{i,j} - \text{MED}_j)}{\text{MAD}_j} \tag{14}$$

where,

$$\begin{aligned} R_j = &\{\mathbf{Y}_{i,j} : \mathbf{Y}_{i,j} > q_{75}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \\ &+ \text{IQR}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\})\} \end{aligned} \tag{15}$$

and IQR(.) is the interquartile range of the expression data:

$$\begin{aligned} \text{IQR}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) = &q_{75}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \\ &- q_{25}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \end{aligned} \tag{16}$$

ORT. Wu and colleagues [11] proposed another variant to the OS statistic where the median of expression values for a feature j is computed per group based as:

$$\text{statistic} = \frac{\sum_{\mathbf{Y}_{i,j} \in R_j} (\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{X})})}{\text{MAD}_j} \tag{17}$$

The definition of R_j is:

$$R_j = \{\mathbf{Y}_{i,j} : \mathbf{Y}_{i,j} > q_{75}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}) + \text{IQR}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\})\} \tag{18}$$

The set R_j is comprised of outliers in case samples that deviate from control samples based on the 75th percentile (q_{75}) and the IQR values of control samples for the feature j. The MAD_j is defined as:

$$\text{MAD}_j = 1.4826 \times \text{median}(\{|\mathbf{X}_{i,j} - \text{MED}_j^{(\mathbf{X})}| : 1 \leq i \leq n\}, \{|\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{Y})}| : 1 \leq i \leq m\}) \tag{19}$$

where $\text{MED}_j^{\mathbf{X}}$ and $\text{MED}_j^{\mathbf{Y}}$ are the median expression values for control and case samples, respectively. Here, the medians are computed for each group, instead of using all samples to estimate a global median value. This method, called Outlier Robust T-statistic (ORT), was observed to be competitive against OS and COPA using breast cancer microarray dataset [69].

MOST. Following the ORT statistic, Lian and colleagues [12] proposed the Maximum Ordered Subset T-statistics (MOST) to detect a subset of case samples having a set of aberrant feature expressions. The MOST statistic requires, for each feature, the case samples to be sorted in descending order according to their expression values. For a feature j, it is defined as:

$$\text{statistic} = \max_{1 \leq k \leq m} \left[\frac{\sum_{1 \leq i \leq k} (\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{X})})}{\text{MAD}_j} - \mu_k \right] / \delta_k \tag{20}$$

where μ_k and δ_k are obtained from the statistics of case samples using the standard normal distribution. A naive iterative approach over case samples is implemented to search for the k value which corresponds to the optimum MOST statistic for the feature j in this exhaustive search. The median absolute deviation of expressions(MAD_j) has the exact formula as in the ORT statistic:

$$\text{MAD}_j = 1.4826 \times \text{median}(\{|\mathbf{X}_{i,j} - \text{MED}_j^{(\mathbf{X})}| : 1 \leq i \leq n\}, \{|\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{Y})}| : 1 \leq i \leq m\}) \tag{21}$$

Results using the breast cancer microarray dataset [69] and an acute lymphoblastic leukemia dataset [70] have shown that MOST achieved superior performances than the previous methods.

LSOSS. Wang and colleagues [13] suggest that outliers in case samples have a different mode or peak of expression values than the remaining inactivated case samples. Therefore, they proposed the Least Sum of Ordered Subset Square t-statistic (LSOSS) method to detect the peak corresponding to outliers in case samples. The LSOSS method aims to find the peak of expression values that maximizes the difference between case and control samples. The LSOSS statistic for a feature j is defined as follows:

$$\text{statistic} = k \frac{\overline{\mathbf{Y}}_{:,j}^{(S)} - \overline{\mathbf{X}}_{:,j}}{s_j} \tag{22}$$

where $\overline{\mathbf{X}}_{:,j}$ represents the mean expression value for control samples for the feature j. $\overline{\mathbf{Y}}^{(S)}_{:,j}$ is the mean expression value up to k samples after sorting the case samples in descending order of their expression values for the feature j. To optimize the hyperparameter k for the feature j, case samples are divided randomly into two groups: $S = \{\mathbf{Y}_{i,j} : 1 \leq i \leq k\}$ and $T = \{\mathbf{Y}_{i,j} : k+1 \leq i \leq m\}$. Then, the mean and sum of squares for the feature j are calculated as:

$$\begin{aligned}
\overline{\mathbf{Y}}^{(S)}_{:,j} &= \text{mean}(\{\mathbf{Y}_{i,j} : 1 \leq i \leq k\}) \\
\overline{\mathbf{Y}}^{(T)}_{:,j} &= \text{mean}(\{\mathbf{Y}_{i,j} : k+1 \leq i \leq m\}) \\
SS^{(S)} &= \sum_{1 \leq i \leq k} (\mathbf{Y}_{i,j} - \overline{\mathbf{Y}}^{(S)}_{:,j})^2 \\
SS^{(T)} &= \sum_{k+1 \leq i \leq m} (\mathbf{Y}_{i,j} - \overline{\mathbf{Y}}^{(T)}_{:,j})^2
\end{aligned} \tag{23}$$

Afterward, the optimal k value is obtained by minimizing the pooled sum of squares from the case samples:

$$\arg \min_{1 \leq k \leq m-1} (SS^{(S)} + SS^{(T)}) \tag{24}$$

The s_j in the denominator of the LSOSS statistic is the squared root of the (pooled) standard error and has the following formula:

$$s_j^2 = \frac{e_j^2 + SS^{(S)} + SS^T}{n + m - 2} \tag{25}$$

where $e_j^2 = \sum_{1 \leq i \leq n} (\mathbf{X}_{i,j} - \overline{\mathbf{X}}_{:,j})^2$ represents the sum of squares for the control samples. The performance of LSOSS was compared with the previous methods using the breast cancer microarray data [69] and the human prostate cancer dataset [71]. Results have indicated that LSOSS is an effective model for detecting more outliers in the case samples than the previous approaches [13].

DIDS. de Ronde and colleagues [14] developed the Detection of Imbalanced Differential Signal (DIDS) algorithm that computes the absolute differences of expression values between the case samples and the maximal expression value from the control samples for a given feature. Then, the differences are summarized using one of the following functions: tanh, square, or square root to obtain the DIDS statistics:

$$\text{statistic} = \sum_{i=1}^{m} f(|\mathbf{Y}_{i,j} - \widehat{\mathbf{X}_{:,j}}|^+) \tag{26}$$

where $\widehat{\mathbf{X}_{:,j}} = \max_k(\{\mathbf{X}_{k,j} : 1 \leq i \leq n\})$ and $|.|$ is:

$$|w|^+ = \begin{cases} w & \text{if } w \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

The f in the DIDS statistics is a strictly increasing function of either:

$$\begin{aligned}
f_{tanh}(w) &= 1 + \tanh(3w - 3) \\
f_{quad}(w) &= w^2 \\
f_{sqrt}(w) &= \sqrt{w}
\end{aligned} \tag{28}$$

DIDS recorded an outstanding performance against t-statistic, COPA, OS, ORT, and MOST on the HER2 dataset [72].

DECO. Campos-Laborie and colleagues developed a method to address four types of heterogeneity in control and case samples (as depicted in Fig. 2). This approach is called DEcomposing heterogeneous Cohorts using Omic data profiling (DECO), and it consists of six major steps:

1. Recurrent Differential Analysis (RDA) step to identify, select and rank all significant changes (using LIMMA [27]) found for any feature between two subsets of control and case samples. A similar step was followed in PHet. The output of this step is the frequency matrix that counts the number of times that each feature was found to be significant in a given sample.
2. Non-Symmetrical Correspondence Analysis (NSCA) step to locate an asymmetric association among features and samples using the frequency matrix from the RDA step.
3. Heterogeneity statistic (h-statistic) computation step to score each feature using results from the NSCA step and the raw data matrix. The output from this step is a feature scoring matrix, called the h-statistic.

4. Hierarchical bi-clustering step to obtain the optimal number of subgroups. The clustering is applied to the h-statistic matrix. Pearson correlation is used as the distance metric for clustering.
5. Recover feature profile changes step to identify the four model types. This is achieved by calculating the overlap of signal between the sample's groups, using each distribution of expression values for each group. This allows mapping each feature belonging to each four mode types of change.
6. Rank all features using results from RDA, h-statistic, and feature profile changes steps.

DECO was shown to deliver promising results for the analysis of heterogeneity, biomarker identification, and subtype discovery [15]. Moreover, DECO can detect subpopulations without requiring labeled samples (control vs case samples), making it suitable to discover the biological heterogeneity of cells.

Parameter settings for benchmark algorithms

We implemented the standard Student t-test, Wilcoxon rank-sum test, Kolmogorov–Smirnov test, COPA, OS, ORT, MOST, LSOSS, DIDS, ΔIQR, and PHet in Python. The same settings were utilized on all benchmark problems. These are: two-tailed t-test for Student t-test, Wilcoxon rank-sum test, Kolmogorov–Smirnov test, and LIMMA; q_r was assigned to 75% for COPA and OS; 'tanh' was used as scoring function for DIDS; IQR range was set to [25,75] for OS, ORT, the four variants of IQR, and PHet; the vector of feature weight w was set to $[0.4, 0.3, 0.2, 0.1]$ for PHet; and up and down regulations were applied for OS, ORT, MOST, LSOSS, DIDS, all variants of IQR, and PHet. For LIMMA, we used the 'limma' package [27] written in R and used the default settings. For Dispersion-based methods, we used the SCANPY package with the default settings. To collect p-values and statistics from the Student t-test, Wilcoxon rank-sum test, and Kolmogorov–Smirnov test, we used the SciPy package [67]. For DECO, we employed the 'deco' package, written in R, and applied the 'decoRDA' module with the number of iterations assigned to 1000. All renaming parameters in 'deco' were set to their default configurations.

Evaluation metrics

Performances of algorithms on different datasets were compared using six evaluation metrics: F1, adjusted rand index (ARI), the adjusted mutual information (AMI), homogeneity, completeness, and V-measure.

F1. The F1 metric is a measure of the performance of a classifier. It is calculated as the harmonic mean of the precision and recall, which are two metrics that evaluate how well the classifier can identify the positive and negative classes as:

$$\mathrm{F}_1 = \frac{2tp}{2tp + fp + fn} \tag{29}$$

where tp, fp, and fn corresponds to the true positive features, false positive features, and false negative features, respectively. The F1 metric ranges in [0, 1], and higher values indicate higher classification accuracy. We used the F1 score to compare the performances of algorithms across simulated, HER2, microarray and single-cell RNA-seq datasets in Figs 3(a),(b), and (e), 4(a) and (d), and Supplementary Figs 1(a), (b), and (e), 2 and 14.

Adjusted rand index (ARI). This metric is used to quantify clustering accuracy by considering all pairs of samples or cells assigned in clusters in the current clustering results and previous clustering or pre-annotated labels adjusted by random permutation as:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E}[\mathrm{RI}]} \tag{30}$$

where E[.] is the expectation of random labeling and the rand index (RI) is defined as:

$$\mathrm{RI} = \frac{a + b}{\binom{n}{2}} \tag{31}$$

where a is the number of samples or cells correctly assigned to the same cluster as benchmark true subtypes; b is the number of samples or cells assigned to different clusters as benchmark true subtypes, and $\binom{n}{2}$ is the combination of selecting two cells from the total of n cells in the cluster. ARI ranges in [0, 1], and higher values imply higher similarities. ARI metric was used in Figs 4(c) and (f), Supplementary Figs 2 and 14.

Adjusted mutual information (AMI). This metric is based on the mutual information (MI) between true labeled subtypes and clustering labels of a dataset adjusted against chance. AMI has the following formula:

$$\mathrm{AMI} = \frac{\mathrm{MI} - \mathrm{E}[\mathrm{MI}]}{\mathrm{Average}(\mathcal{H}(t), \mathcal{H}(\hat{t})) - \mathrm{E}[\mathrm{MI}]} \tag{32}$$

where $t \in \mathbb{N}^{n+m}$ and $\hat{t} \in \mathbb{N}^{n+m}$ are the true subtypes and predicted clusters, respectively. E[.] is the expectation of random labeling and MI is defined as:

$$\mathrm{MI}(t, \hat{t}) = \sum_{c \in |t|} \sum_{k \in |\hat{t}|} \frac{|t_c \cap \hat{t}_k|}{n+m} \log \frac{(n+m)|t_c \cap \hat{t}_k|}{|t_c| \times |\hat{t}_k|} \tag{33}$$

where $\mathcal{H}(t)$ is the entropy of the truly labeled subtypes, which measures how much information is needed to identify the true labels of a sample (or cell) without any clustering information while the term $\mathcal{H}(\hat{t})$ is the entropy of the predicted clusters:

$$\begin{aligned} \mathcal{H}(t) &= -\sum_{c \in |t|} \frac{x_c}{n+m} \log(\frac{x_c}{n+m}) \\ \mathcal{H}(\hat{t}) &= -\sum_{k \in |\hat{t}|} \frac{x_k}{n+m} \log(\frac{x_k}{n+m}) \end{aligned} \tag{34}$$

where x_k represents the total number of samples (or cells) in the predicted cluster k and x_c is the number of samples (or cells) that have the true subtype c. AMI ranges in [0,1] where a higher score entails that both true labeled subtypes and clustering labels are identical. We used the AMI score to compare the clustering performances of algorithms for microarray and single-cell RNA-seq datasets in Supplementary Figs 2 and 14.

Homogeneity. This metric measures the subtype distribution over each predicted cluster:

$$\text{Homogeneity} = 1 - \frac{\mathcal{H}(t|\hat{t})}{\mathcal{H}(t)} \tag{35}$$

where t and $\hat{t}$ are the true subtypes and predicted clusters, respectively, and $\mathcal{H}(t)$ is defined in Eq. 34. The term $\mathcal{H}(t|\hat{t})$ is the conditional empirical entropy, which measures how much information is needed to identify the true labels of a sample (or cell) given its predicted clusters:

$$\mathcal{H}(t|\hat{t}) = -\sum_{c \in |t|} \sum_{k \in |\hat{t}|} \frac{x_{c,k}}{n+m} \log(\frac{x_{c,k}}{x_k}) \tag{36}$$

where $x_{c,k}$ corresponds to the number of samples (or cells) in cluster k that belongs to the true label c, x_k represents the total number of samples (or cells) in the predicted cluster k, and x_c is the number of samples (or cells) that have the true subtype c. The homogeneity score ranges in [0,1], where 1 denotes that each cluster contains only samples (or cells) from one true subtype (zero entropy), and 0 entails that each cluster contains samples (or cells) from all subtypes. A higher homogeneity value indicates a better clustering performance. This metric was applied to compare the clustering performances of algorithms for microarray and single-cell RNA-seq datasets in Supplementary Figs 2 and 14.

Completeness. This metric measures the clustering distribution over each true subtype. A clustering result satisfies completeness if all samples (or cells) that are members of a given subtype belong to the same cluster:

$$\text{Completeness} = 1 - \frac{\mathcal{H}(\hat{t}|t)}{\mathcal{H}(\hat{t})} \tag{37}$$

where t and $\hat{t}$ are the true subtypes and predicted clusters, respectively, and $\mathcal{H}(\hat{t})$ is defined in Eq. 34. $\mathcal{H}(\hat{t}|t)$ is the conditional entropy of the predicted cluster distribution given the true subtypes:

$$\mathcal{H}(\hat{t}|t) = -\sum_{k \in |\hat{t}|} \sum_{c \in |t|} \frac{x_{c,k}}{n+m} \log(\frac{x_{c,k}}{x_c}) \tag{38}$$

where $x_{c,k}$ corresponds to the number of samples (or cells) in cluster k that belongs to the true label c, x_k represents the total number of samples (or cells) in the predicted cluster k, and x_c is the number number of samples (or cells) that have the true subtype c. The completeness score ranges in [0,1] where a higher value indicates a better clustering completeness. This metric was applied to compare the clustering performances of algorithms for microarray and single-cell RNA-seq datasets in Supplementary Figs 2 and 14.

V-measure. This metric is the harmonic mean of two scores: homogeneity and completeness:

$$\text{V-measure} = (1+\beta) \times \frac{\text{Homogeneity} \times \text{Completeness}}{\beta \times \text{Homogeneity} + \text{Completeness}} \tag{39}$$

where Homogeneity and Completeness are defined in Eq. 35 and 37, respectively. $\beta \in \mathbb{R}$ is a tuning parameter to favor the contributions of homogeneity or completeness. By default, β is set to 1. The V-measure score ranges in [0,1] where a higher value quantifies the goodness of the clustering result. This metric was applied to compare the clustering performances of algorithms for microarray and single-cell RNA-seq datasets in Supplementary Figs 2 and 14.

UMAP visualization

To explore the heterogeneity of patient samples and cells expression matrix, we applied Uniform Manifold Approximation and Projection (UMAP) [73] on the significant features. We used the UMAP package [59] in Python and tuned some parameters to achieve better results. We set the minimum distance to 0 to achieve better subtyping (clustering) groupings, the number of neighbors to 5 to help locate smaller subtypes, and the number of iterations to 1000 to obtain adequate embeddings. The remaining UMAP hyperparameters were set to their default configurations.

Clustering

In this study, we compared different clustering methods on two types of data: microarray gene expression and single-cell RNA-seq. We used spectral and k-means clustering as representative algorithms for each data type (Supplementary Table 5). The number of clusters in both types of data was set to the number of original patient samples and cell types. We used the Scikit-learn package [74] to perform clustering using the default settings. We used SCANPY [41] to generate Fig. 23. Each cell was normalized to 10,000 read counts and the data was log-transformed. Then, we performed regression for each feature where the response variable was set to total counts for each cell. We scale residuals to unit variance and zero mean, and we truncated values with a maximum value set to 10. This is followed by applying UMAP at 2000 iterations and Leiden community detection at resolution ranges in [0.4, 0.8]. Empirically, we found these processing steps can ultimately identify clusters.

Differential feature expression

For Figs 4(a) and (d), we performed differential expression analysis using LIMMA [27], and the Benjamini-Hochberg [66] correcting method for the adjustment of p-values. We selected the top 100 DE features that have adjusted p-value below 0.01 to report the performances of algorithms.

Datasets

We used the following benchmark datasets: i)- two sets of simulated data, ii)- 1000 datasets generated using the HER2 breast cancer gene expression data (Supplementary Table 1), iii)- 11 real-world microarray expression datasets with known patient types (Supplementary Table 2), iv)- six single-cell transcriptomics datasets with annotated cell types (Supplementary Table 3), and v)- three single-cell transcriptomics datasets corresponding the lung airway epithelial cells admixed with known and unknown cell types (Supplementary Table 4).

Simulated datasets. A base artificial data was generated using the 'madsim' package [75] with the following configurations: the lower and upper bounds of $\log_2$ intensity values were set to 2 and 16, respectively, and the proportion of the DE features was set to 0.1. The remaining parameters in madsim package were set to their default values. The data consists of $p = 1100$ features, with 93 DE features, and 40 samples split evenly between control ($n = 20$) and case ($m = 20$) classes. The purpose of using this base data is to provide a prior distribution for the simulation of two additional sets of data. Each of these sets has 5 data that are drawn from the same underlying base artificial data in control-case outlier detection studies. We created the minority model type by randomly perturbing 100 features in the case group, including some DE features, and created outliers ranging from 5% to 45% of the case group, corresponding to 1 to 9 out of 20 samples. The first data set reflects this model type. For the mixed model scheme, we followed the same procedure, but we distributed the outliers equally among a subset of cases and controls. These two sets of simulated data correspond to the balanced situations, where the number of case and control samples is exactly the same ($n = 20$ and $m = 20$), which is ideal to evaluate the performances of ten algorithms in Table 1.

HER2 datasets. We applied the DIDS approach [14] to create 1000 synthetic batches of HER2 (human epidermal growth factor receptor 2) data, each with 27,506 features and 188 samples. The samples consisted of 178 fixed control examples from a breast cancer dataset (GEO accession ID GSE34138), which represented the HER2-negative group, and 10 case samples randomly chosen from 60 HER2-positive tumor samples (GEO accession ID GSE41656). The case samples were different for each batch, while the control samples were the same. We used this data to evaluate the ability and robustness of algorithms in Section 3 to identify 20 DE genes located on ch17q12 or ch17q21. It is important to note that the HER2-positive breast cancer in the case group may exhibit false positive tumors due to the limitations of the HER2 immunohistochemistry (IHC) test, which may result in the inaccuracy of ranking features by algorithms.

Microarray datasets. We downloaded 7 microarray expression data from Gene Expression Omnibus (GEO) databases with accession numbers of GSE412 (the GSE412 data); GSE3726 (the GSE3726 data); GSE89 (the GSE89 data); GSE68956 (the Braintumor data); GSE2685 (the GSE2685 data); GSE83227 (the Lung data); GSE1987 (the GSE1987 data). The Glioblastoma, Leukemia, MLL, and SRBCT data were downloaded

from Bioinformatics Laboratory at the University of Ljubljana from `https://file.biolab.si/biolab/supp/bi-cancer/projections/index.html`.

Single-cell transcriptomics datasets with known sub-types. We downloaded the six scRNA-seq datasets from the Gene Expression Omnibus (GEO) databases with accession numbers of GSE67835 (the Darmanis data); GSE36552 (the Yan data); GSE81252 and GSE96981 (the Camp data); GSE84133 (the Baron data); GSE81861 (the Li data); and GSE57872 (the Patel data).

Single-cell transcriptomics datasets with (un)known sub-types. We downloaded three scRNA-seq datasets corresponding the airway epithelial cells from Gene Expression Omnibus (GEO) databases with accession numbers of GSE102580 (the HBECs and MTECs data) and GSE103354 (the Pulse-seq data).

Software implementation

Tools and packages used in this paper include: Python version 3.7.6, numpy version $\geq$ 1.20, pandas version $\geq$ 2.0, scanpy version $\geq$ 1.9, scikit-learn version $\geq$ 1.2, SciPy version $\geq$ 1.10, matplotlib version $\geq$ 3.5, seaborn version $\geq$ 0.11, umap-learn version $\geq$ 0.5, statsmodels version $\geq$ 0.12, and R version $\geq$ 3.6.

Data availability

Available upon approval of this manuscript.

Code availability

Available upon request.

Author Contributions

A.R.M.A.B., C.H., and K.L. conceived the research. A.R.M.A.B. and C.H. wrote the PHet package, performed computational analysis, and experimental work, and analyzed and produced figures using the discussed data. A.R.M.A.B., C.H., and K.L. wrote the manuscript.

Acknowledgements

This work was supported by NIH, United States (Grant Number: R35GM133725).

Conflict of Interest

The authors declare no competing financial or non-financial interests.

Author Information

Correspondence and requests for material, data, and code should be addressed to K.L. (`kwonmoo.lee@childrens.harvard.edu`).

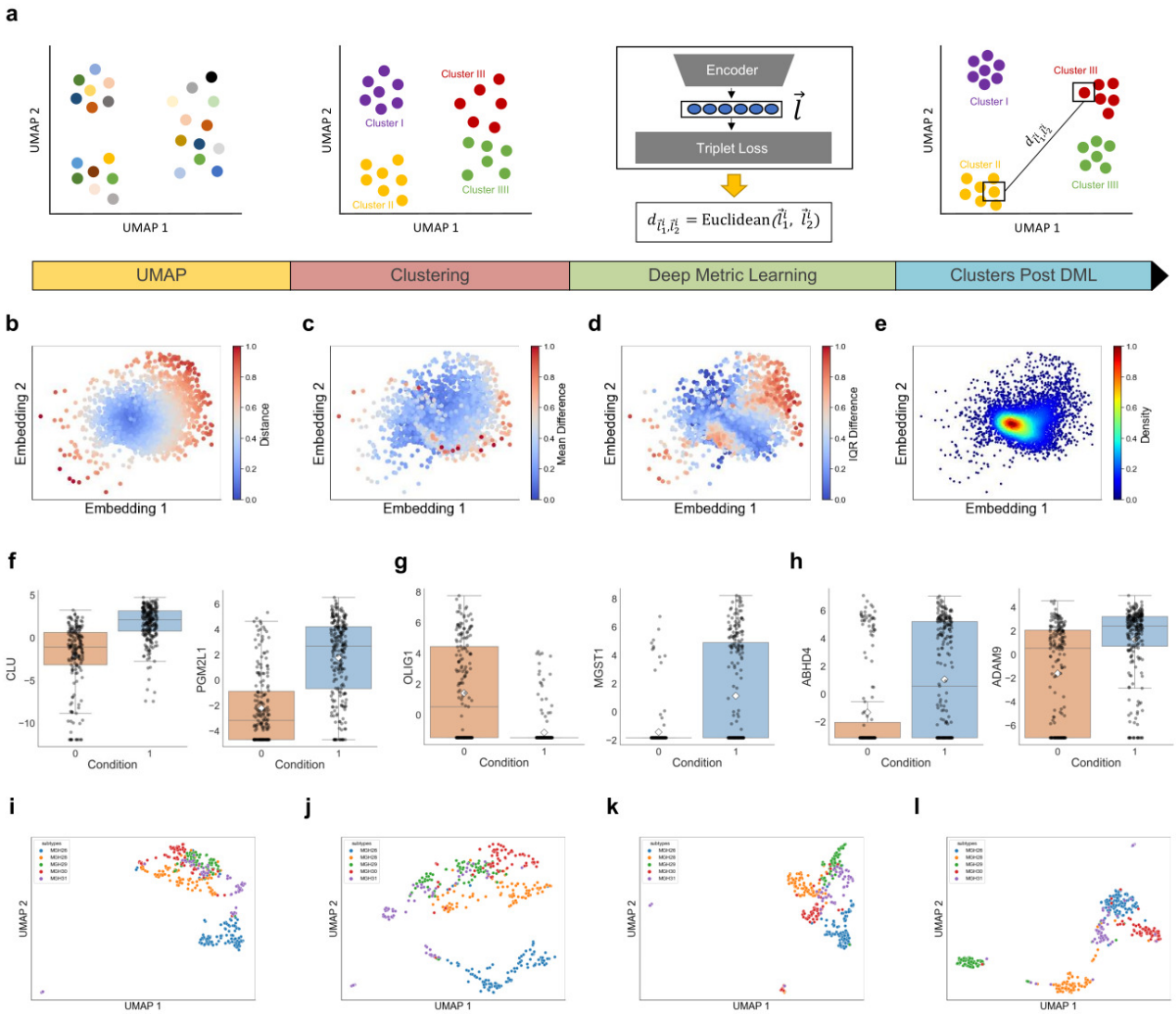

Figure 1: **Characterization of features extrapolated by deep-metric learning.** **a**, Schematic of Deep Metric Learning with triplet loss to analyze feature space. A Feature UMAP is generated where each point represents one feature from one condition. Feature Clustering is then performed on the UMAP space so that the cluster number corresponds to the y label and the original data will become input to the triplet loss autoencoder. Obtain embedding from autoencoder and perform Euclidean distance between the same feature of different classes, where $\vec{l}$ is the embedding, $\vec{l}_1^i$ is the embedding for gene i and condition 1, and $d_{\vec{l}_1^i, \vec{l}_2^i}$ is the Euclidean distance between embeddings of gene i of condition 1 and of condition 2. Feature embedding scatterplots with color representing distance between the same feature of different classes (**b**) mean difference between classes (**c**) and IQR difference between classes (**d**). **e**, A density plot of the feature embeddings. Boxplots of features with a high DML score that are differentially expressed features determined by a z-test (**f**), contain a large amount of variation when comparing classes (**g**), and contains a large mean/median difference as well as a large IQR difference (**h**). UMAPs of feature types of all 5948 features (**i**), 2302 DE features determined by a z-test (**j**), 1529 HV features using a threshold of 0.5 (**k**), and 146 HD features with a significance level of 0.01 (**l**).

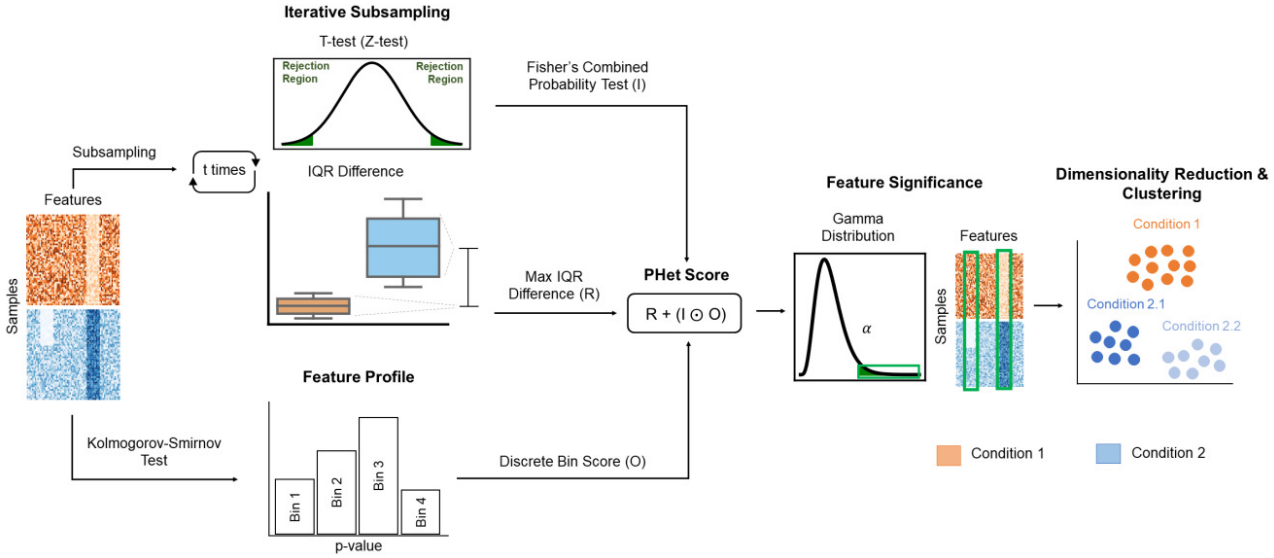

Figure 2: **Illustration of the PHet framework.** PHet pipeline is composed of six major steps: 1)- an iterative subsampling to calculate p-values, using t-test or z-test, and absolute interquartile range differences (ΔIQR) for each feature between two subsets of control and case samples, 2)- the Fisher's combined probability test to summarize the collected p-values, 3)- a weighted features profile to map each feature to its profile using the Kolmogorov–Smirnov test. 4)- feature statistics estimation using a combination of the ΔIQR values, feature profiles, and Fisher's combined probability scores, 5)- fitting features statistics using the gamma distribution, and features exceeding a user threshold are trimmed, and 6)- dimensionality reduction and clustering analysis on the reduced omics data to reconstruct data heterogeneity.

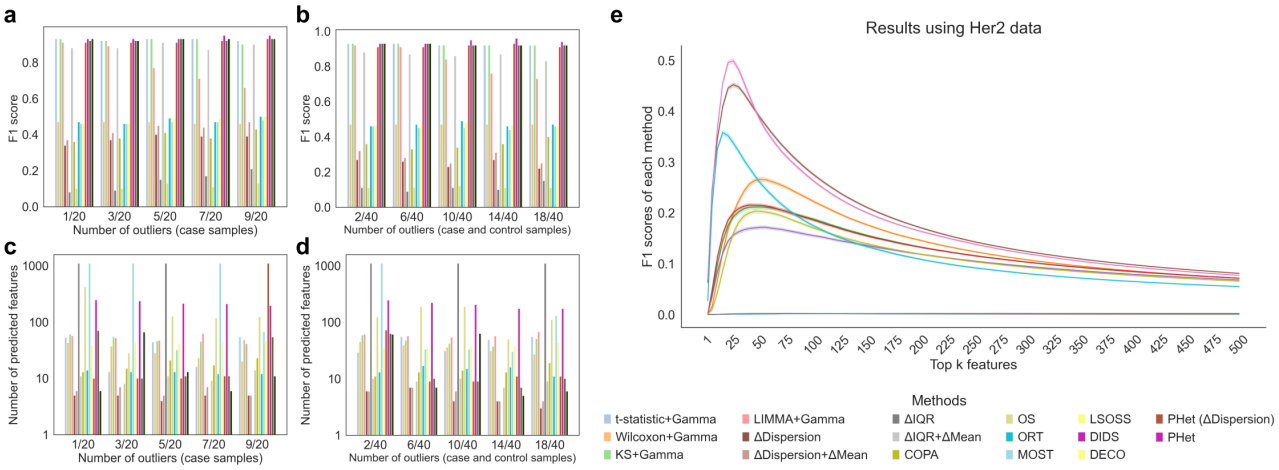

Figure 3: **Performance comparison of PHet and other 16 baseline methods using simulated and HER2 datasets.** **a**, **b**, F1 scores of 17 different methods to detecting the true expressed features for outliers ($< 10\%$). **c**, **d**, Number of predicted features for each method. **a**, **c**, Comparison between 10 methods under 5 conditions representing the number of outliers ($\{1, 3, 5, 7, 9\}$) in the case samples. **b**, **d** Comparison between 10 methods under 5 conditions corresponding to an increasing number of samples: from 2 to 18 in both case and control groups. Results (**a**, **b**, **c**, and **d**) are obtained using 10 simulated datasets (Supplementary Table 1). **e**, Comparison of the PHet's performance detecting top k true features on HER2 data with the baseline methods. Performance is measured as an F1 score. True features are ranked in the decreasing order of the p-value estimated using LIMMA.

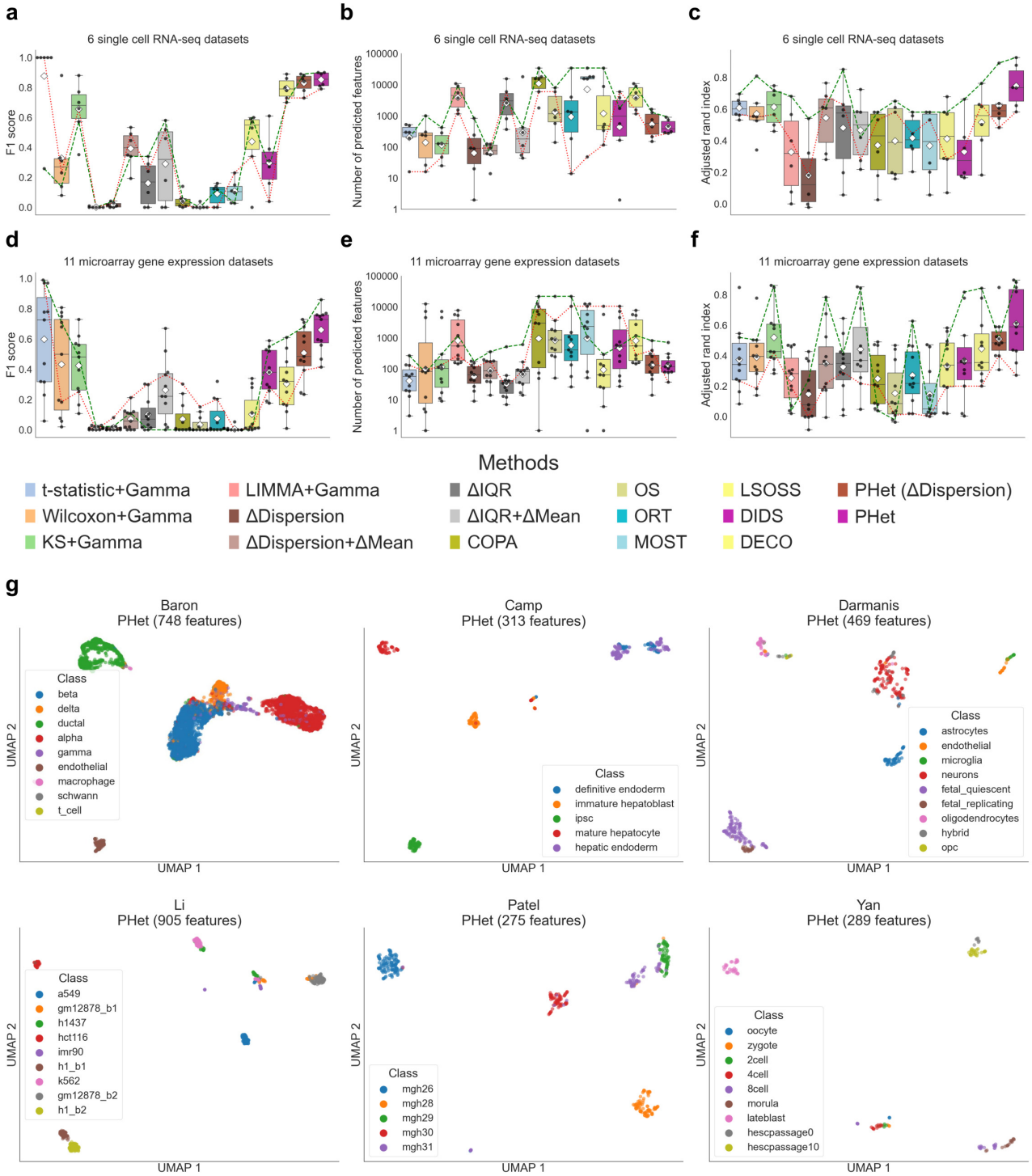

Figure 4: **Performance comparison of PHet and other baseline methods using 11 microarray and six single-cell transcriptomics datasets (Supplementary Tables 2 and 3).** **a**,**d**, F1 scores of each method for detecting the top 100 DE features that are obtained using LIMMA. **b**,**e**, Number of predicted important features by each method. **c**,**f**, Adjusted Rand index of each method for 11 microarray and six single-cell transcriptomics datasets, where 1 corresponds to high-quality clustering while 0 indicates random clustering. The box plots show the medians (centerlines), first and third quartiles (bounds of boxes), and 1.5× interquartile range (whiskers). A ◊ symbol represents a mean value. A green dashed line indicates the best-performing result on a dataset of PHet on each metric. Similarly, a red dashed line represents the worst-performing result on a dataset of PHet on each metric. **g**, UMAP visualizations of six single-cell transcriptomic datasets (Supplementary Table 3) using predicted PHet's features, colored by predefined original cell types.

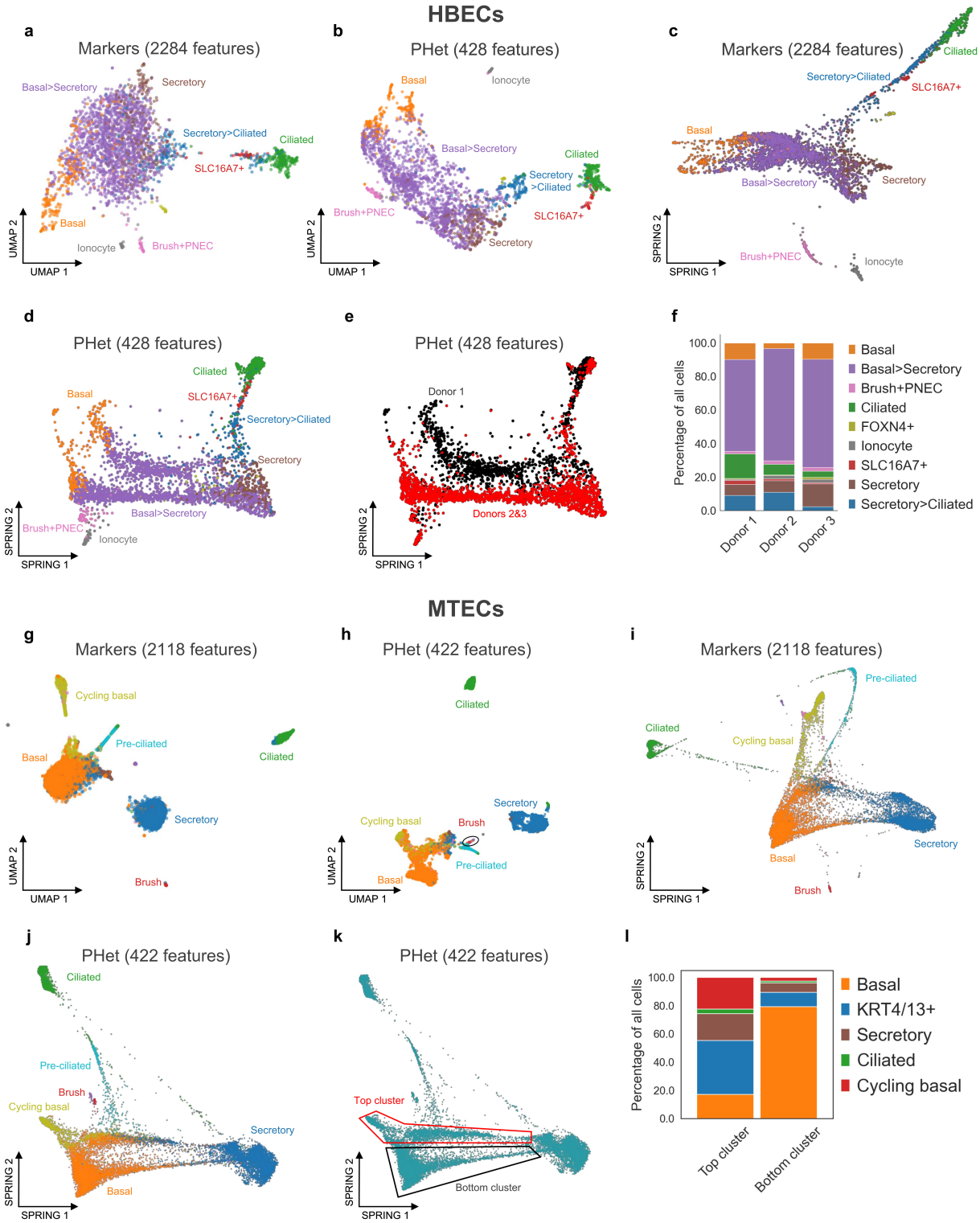

Figure 5: **Lineage reconstruction of HBECs and MTECs single-cell RNA datasets.** UMAP visualizations using **a** known markers and **b** PHet's predicted features and their corresponding SPRING visualizations **c** and **d**, respectively. Coloring represents previously annotated cell types. **e**, SPRING plot using PHet's features shows two trajectories: donor 1 (black) and donors 2 & 3 (red). **f**, Relative abundance of cell types grouped by donors. **g**, Top differentially expressed features for the two trajectories and three SPRING plots of selected top features are shown (Supplementary Fig. 21). **g**,**h**, UMAP visualizations corresponding markers and PHet's predicted features. The color indicates annotated cell types. **i**,**j**, SPRING visualization of markers and PHet features to explore trajectories of cells. **k**, Two distinct clusters of cells in the SPRING plot: one on the top and one on the bottom, and the relative abundance of cells between these clusters are shown as barplots (**l**), identified using markers for different cell types (Extended Data Fig. 22).