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

The discovery of subtypes is pivotal for disease diagnosis and targeted therapy, considering the diverse responses of different cells or patients to specific treatments. Exploring the heterogeneity within disease or cell states provides insights into disease progression mechanisms and cell differentiation. The advent of high-throughput technologies has enabled the generation and analysis of various molecular data types, such as single-cell RNA-seq, proteomic, and imaging datasets, at large scales. While presenting opportunities for subtype discovery, these datasets pose challenges in finding relevant signatures due to their high dimensionality. Feature selection, a crucial step in the analysis pipeline, involves choosing signatures that reduce the feature size for more efficient downstream computational analysis. Numerous existing methods focus on selecting signatures that differentiate known diseases or cell states, yet they often fall short in identifying features that preserve heterogeneity and reveal subtypes. To identify features that can capture the diversity within each class while also maintaining the discrimination of known disease states, we employed deep metric learning-based feature embedding to conduct a detailed exploration of the statistical properties of features essential in preserving heterogeneity. Our analysis revealed that features with a significant difference in interquartile range (IQR) between classes possess crucial subtype information. Guided by this insight, we developed a robust statistical method, termed PHet (Preserving Heterogeneity) that performs iterative subsampling differential analysis of IQR and Fisher's method between classes, identifying a minimal set of heterogeneity-preserving discriminative features to optimize subtype clustering quality. Validation using public single-cell RNA-seq and microarray datasets showcased PHet's effectiveness in preserving sample heterogeneity while maintaining discrimination of known disease/cell states, surpassing the performance of previous outlier-based methods. Furthermore, analysis of a single-cell RNA-seq dataset from mouse tracheal epithelial cells revealed, through PHet-based features, the presence of two distinct basal cell subtypes undergoing differentiation toward a luminal secretory phenotype. Notably, one of these subtypes exhibited high expression of BPIFA1. Interestingly, previous studies have linked BPIFA1 secretion to the emergence of secretory cells during mucociliary differentiation of airway epithelial cells. PHet successfully pinpointed the basal cell subtype associated with this phenomenon, a distinction that pre-annotated markers and dispersion-based features failed to make due to their admixed feature expression profiles. These findings underscore the potential of our method to deepen our understanding of the mechanisms underlying diseases and cell differentiation and contribute significantly to personalized medicine.

1 Introduction

Subtype discovery plays a pivotal role in disease diagnosis and targeted therapy, particularly due to the marked heterogeneity in how cells or patients react to specific treatments. For example, chromosomal translocations can lead to different tumor types[1,2], which can influence the choice of optimal and personalized treatments[3–5]. This underscores the utmost significance of uncovering extensive diversity and variability of tumor cells, both within individual cancer patients and across patient populations[6–11]. This holds profound implications in decerning the molecular mechanisms underlying disease initiation/progression, enhancing patient classification, and contributing to the ongoing advancement in tailoring medical interventions to individual characteristics and needs[12–16].

Molecular signatures, such as genes, mRNA transcripts, and proteins, play paramount roles in delineating diverse cell types and states[17]. The advancement of omics technologies in the last two decades enabled simultaneous analysis of thousands of molecular entities to facilitate the characterization of various cells and diseases[18]. This has ushered in an era conducive to characterizing intricate biological conditions, consequently augmenting the demand for in-depth subtyping of both cells and diseases. However, the complexity of omics data, which is inherently high dimensional, poses a challenge to downstream computational processes. Moreover, identifying the molecular signatures integral to data interpretation becomes a formidable task[19]. To circumvent this, a computational process, called feature selection should be employed to extract a subset of features that are the

most informative and pertinent from high-dimensional omics data to discern the specific molecular patterns distinctive to each cell or disease subtype. This process provides several benefits for omics data analysis, such as reducing noise, sparsity, and redundancy in the data, improving computational efficiency, and enhancing interpretability. Feature selection is proven by various theoretical and empirical studies to be effective for different tasks, such as classification, regression, clustering, and dimensionality reduction[20]. Nonetheless, conventional feature selection methods face limitations in the realm of subtype identification, primarily owing to the ensuing reasons.

Traditionally, discriminative feature selection techniques have been employed to examine the molecular patterns of cells or patient samples under predefined conditions. The prevalent methods for this approach fall within the area of differential feature expression analysis. This category of methods identifies molecules, such as genes that display contrasting expression levels or associations with conditions across two distinct groups, such as cancer and healthy individuals[21–36]. It is vital to recognize that features from these methods are discriminative and differentially expressed (DE). However, in numerous real-world scenarios, the complexity and heterogeneity of omics data are not readily captured by differential expression methods. Hence, while the discriminative feature selection paradigm centers on distinguishing established conditions and states, it can inadvertently stifle the exploration of diversity inherent within the data. This limitation can potentially simplify feature space, ultimately impeding the effectiveness of in-depth subtyping.

To address this issue, Tomlins and colleagues[37] introduced a novel statistical method called "cancer outlier profile analysis" (COPA). This method was designed to identify subtypes within cancer patients by focusing on the genes whose expression profiles exhibit minor outliers. This pioneering work inspired a series of studies that improved COPA with more sophisticated statistical and machine learning techniques for subgroup discovery within each experimental condition[38–43]. These methods involve ranking features based on statistical metrics that gauge the degree of abnormal expression across two experimental scenarios. Nonetheless, they are not well-suited for tackling intricate subtyping challenges, such as deciphering the population structure from single cells and detecting the cell state transitions across conditions such as development processes. The limitation lies in the fact that outlier genes do not encompass the full spectrum of sample heterogeneity and often lack the requisite discriminative potency. Conversely, the concept of selecting highly variable (HV) genes gained prominence, particularly in the context of single-cell RNA-seq analysis, for uncovering molecular signatures linked to diverse cell types[44–46]. Often, these methods seek a set of HV features that exhibit significant variation across samples, regardless of experimental conditions, employing these features for subsequent clustering analysis. The downstream differential analysis among these clusters can be conducted to establish associations with distinct molecular signatures. However, it is important to note that HV features are not intended for uncovering unknown subtypes within conditions. Moreover, they may lack the necessary discriminatory power and often incorporate substantial irrelevant heterogeneity for specific subtyping tasks. This phenomenon can significantly compromise the capability of subtype discovery.

Given the constraints inherent in current methods, the process of subtyping within each condition demands a considerable investment of time. Researchers must integrate domain expertise, differential expression analysis, and subsequent verification to achieve accurate results. Addressing this challenge requires the development of a comprehensive computational framework that adeptly identifies subtypes from omics data in a supervised setting involving two distinct experimental conditions. The algorithm should possess the capability to identify features that exhibit both differential expression and variability across latent subtypes. To surmount this obstacle, we introduce a novel category of features that has largely eluded the notice of existing methods, termed Heterogeneity-preserving Discriminative (HD) features, effectively possessing both DE and HV attributes. Through our deep metric learning approach, focused on gene embedding derived from single-cell RNA seq datasets, we discovered the existence of a significant proportion of HD features, characterized by the distinct interquartile range (IQR) differences present between the two pre-defined experimental conditions. Capitalizing on these HD features, we can achieve a more nuanced clustering of patient samples or cells, fostering a deeper comprehension of the factors that influence their diversity. Therefore, this approach unveils new molecular insights that might otherwise remain obscured, effectively transcending the challenges inherent in conventional methodologies.

Inspired by the significant advantage provided by HD features in discerning distinct subtypes, we developed a novel computational framework, named PHet (Preserving Heterogeneity), that leverages a iterative subsampling approach to determine an appropriate and adequate set of features for subtype discovery tasks from omics data. In each iteration, a small fraction of samples of fixed size from each biological condition is chosen to capture a balanced proportion of DE and HD features. Subsequently, we score each feature based on maximum absolute IQR differences[47] and the Fisher's method[48] between samples of two conditions. Then, PHet performs the Kolomorogov-Simirnov test between the two conditions to reduce scores for DE features. In this way, PHet attempts to minimize the number of features while maximizing the quality of subtype clustering. This min-max optimization of feature selection is seldom addressed in the literature, and PHet is the first method to solve this problem in the context of subtype detection. We demonstrated the power of PHet by conducting a com-

prehensive benchmarking on two simulated, 11 microarray, and six single-cell transcriptomics datasets. Given the dearth of consensus and comprehensive comparisons within the existing landscape, we diligently compared PHet against 24 computational methods, employing a range of performance metrics to ensure fairness and comprehensiveness. Our findings highlight the consistent superiority of PHet over competing methodologies in the realm of identifying subtypes under binary experimental conditions. Moreover, PHet exhibited a remarkable ability to unveil novel cell subtypes within both mouse and human airway epithelium single-cell transcriptomics datasets[49]. PHet also displayed superior scalability and robustness with high-dimensional data, setting it apart from the array of benchmarked algorithms.

2 Results

2.1 Identification of heterogeneity-preserving discriminative features in omics data by deep metric learning

To better understand the key statistical attributes that contribute to the preservation of subtype heterogeneity within features, we conducted a feature statistic embedding through the utilization of deep metric learning (DML)[50] (Section 3). Our methodology began with sorting feature measurements within each disease or cell condition. Following this, we employed UMAP (Uniform Manifold Approximation and Projection)[51] to group features within each condition based on the similarity of their respective statistical distributions. These clusters served as a foundation for our subsequent DML approach. Within the DML framework, we utilized the triplet loss to organize similar features by minimizing the distance that separates them, while simultaneously maximizing the distance between dissimilar features (Fig. 1(a)). Our DML encoder, in turn, performed the task of embedding feature statistics from each condition into a lower-dimensional space, facilitated by the acquisition of a meaningful distance between these feature embeddings. The final step of our process involved the calculation of the differences between the feature embeddings of the same feature under different conditions. This final subtraction step allows us to derive insights into the variations present across different disease states or cell conditions, thereby enhancing our understanding of the intrinsic heterogeneity preservation within the features.

Our implementation of DML was applied to the Patel data[52], a single-cell RNA-seq dataset encompassing heterogeneous subtypes of primary glioblastomas. We computed the distance between feature (i.e., gene) embeddings across two different conditions, enabling the assignment of distances to genes based on their informativeness. In the context of the Patel dataset, we employed DML to compare the embeddings of gene expression for two disease conditions—progenitor states (MGH26 and MGH30) and differentiated states (MGH28, MGH29, and MGH31). Upon observing the resulting embeddings, it becomes apparent that genes with high distances are notably situated at a considerable distance from the center (as depicted in Fig. 1(b)). While these embeddings effectively illustrate the areas occupied by genes with substantial mean differences between the two conditions (Fig. 1(c)), they fall short of providing a comprehensive explanation for the acquired distance metrics. Intriguingly, mapping the IQR differences between the two conditions onto the embedding demonstrates a clear pattern: genes with similar IQR differences cluster closely together, and a significant proportion of genes characterized by large IQR differences between the two conditions tend to cluster within regions of substantial distance (Fig. 1(d)). This underscores the potential of IQR differences as a valuable statistical attribute capable of reflecting the underlying heterogeneity between the two conditions.

Based on our observations, we introduce a novel concept known as Heterogeneity-preserving Discriminative (HD) features (Fig. 1(e)), which combine the properties of differentially expressed (DE) and differentially variable (DV)[53] traits. DE features highlight significant variations in expression or abundance across distinct groups, while HV features, as proposed by Satija and colleagues[45], display notable variability across samples, capturing inherent data heterogeneity without predefined conditions. It is important to note that DE features are a subset of discriminative features used in supervised machine learning models to separate different classes or conditions. Discriminative features, however, often exhibit redundancy, which intricates the process of characterizing samples under various conditions. By applying appropriate statistical methods, a smaller set of discriminative features can be obtained that show significant differences in their values or distributions across different classes or conditions. These selected features are the differentially expressed discriminative features[54] or DE. On the other hand, in our study, DV features are determined by IQR differences between two conditions. While previous studies have utilized DV features, incorporating variance differences, to identify disease-related genes[53], the specific application of DV features for subtyping within distinct conditions has not been considered. HD features constitute a distinct category, uniquely characterized by their combined representation of both mean expression differences (DE) and variability discrepancies (DV) (Fig. 1(e)). This intersection of attributes brings to light a previously underexplored feature class with potential implications.

To showcase the utility of HD features, we conducted a comparative subtyping analysis using full, DE (identified through mean differences), HV (located using dispersion-based selection[45]), and the novel HD features,

all within the context of the Patel dataset[52] (Fig. 1(f)-(i)). When using all features (a total of 5948), we observed admixed samples and lacked clear indications of subtypes (Supplementary Fig. 19). Interestingly, we found that DE features outperformed HV features in distinguishing the five subtypes in the Patel data (Fig. 1(f)-(g)), despite having a higher number of features (> 1900 features). This suggests that HV features, which consisted of 1529 features, were insufficient for accurately characterizing the subtypes. The IQR difference method reduced the feature size to 904, resulting in improved clustering (Fig. 1(g)). However, the MGH26 and MGH31 subtypes appear to be closely linked or attached in this analysis. This indicates that the Patel data may contain many irrelevant or redundant features for subtype discovery. The most promising outcomes were obtained when using the HD features, which comprise less than 15% (641 features) of the total features in the Patel data (Fig. 1(h)). These HD features combine the properties of both DE and IQR differences, offering a valuable resource for gaining a more refined understanding of the factors influencing experimental outcomes. This realization motivated the development of PHet, our computational framework aimed at optimizing the integration of DE, HD, and to a lesser extent HV features, facilitating effective subtyping in omics data under the supervised setting. The emergence of HD features and their incorporation into our methodology represents a significant step toward enhancing approaches for disease subtyping.

2.2 Overview of PHet

PHet is a method designed to detect informative features for subtype discovery from omics expression data. The initial step involves annotating the data with two distinct conditions: control and case (e.g., healthy and cancerous conditions). Subsequently, the data undergoes a series of preprocessing steps, including the removal of low-quality samples and features, as well as the normalization of expression values to have zero mean and unit variance. This preprocessed data is then fed to the PHet pipeline which is composed of six major stages (Fig. 1(j) and Section 3). (i) Iterative subsampling[55,56]: This step calculates both p-values, using t-test or z-test, and maximum absolute interquartile range differences (ΔIQR) for each feature between two subsets of control and case samples. The size of these subsets is determined by the closest integer to the square root of the minimum number of samples in either the control or case group, i.e., $\min(n, m)$, where n and m correspond to the number of samples in control and case groups, respectively. This approach ensures unbiased estimation of p-values by balancing both groups and has been applied to detect features containing intrinsic heterogeneity[43]. 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 them. To ensure capturing sufficient features that help in subtype identification, the subsampling procedure is repeated for a predefined number of iterations (default is 1000).

(ii) Fisher's combined probability test: Following subsampling, the collected p-values are summarized by the Fisher's combined probability test. The results from this test serve as prior information to calculating feature statistics and ranking. (iii) Incorporating discriminatory power of features: During the iterative subsampling process, a challenge arises in obtaining samples contributing to the multimodal distributions or multiple peaks, which represent distinct subtypes. This challenge becomes particularly pronounced when peaks are represented by a small proportion of samples under a specific set of features, risking exclusion or underrepresentation of these peak-associated subsamples. Inaccurate estimations of a key statistical measure, Fisher's scores, can occur for features linked to these subtypes. A remedy involves the nonparametric two-sample Kolmogorov-Smirnov (KS) test between control and case samples. This test compares the cumulative distribution functions of the two sample groups, identifying features significantly contributing to observed differences. By incorporating the KS test, we adjust Fisher's scores to capture peak-associated features. The p-values from the KS test are organized into predefined intervals (default is four), each with a predefined weight assigned to specific features within the interval. These weights reflect the discriminatory power of respective features, allowing the regularization of the scores obtained from Fisher's method. Importantly, these weights (default is $[0.4, 0.3, 0.2, 0.1]$) are empirically determined and may vary based on the specific analysis conducted. (iv) Feature Statistics and Thresholding: Fisher's scores ($\mathbf{f}$) multiplied by the features discriminatory power ($\mathbf{o}$) are then combined with the absolute maximum values of ΔIQR ($\mathbf{r}$) to estimate feature statistics, i.e., $\mathbf{r}+(\mathbf{f}\odot\mathbf{o})$, where $\odot$ corresponds to the Hadamard (element-wise) product. This step identifies DE, HD, and, to a lesser extent HV features that are important for subtype detection. Then, the 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. (v) Dimensionality Reduction and Clustering: In the final step, dimensionality reduction and clustering methods are performed on the reduced omics data with the selected features (from the gamma distribution) to reveal the data's underlying heterogeneity.

2.3 Evaluation of PHet's performance in identifying subtypes of single cells and patients

To assess the performance of PHet in cell subtyping tasks, we subjected PHet to systematic benchmarking using six publicly available single-cell transcriptomic (scRNA-seq) datasets (Section 3; Supplementary Table

3). Cells in the scRNA-seq datasets were divided into binary experimental conditions: control and case. An effective algorithm should accurately discern cell subtypes while preserving biological integrity and heterogeneity using a minimum number of features. Prior to algorithm execution, we visualized each scRNA-seq dataset with all features using UMAP and observed that cells in Baron[57], Darmanis[58], and Patel[52] datasets displayed random distribution without the potential existence of cell subclusters, indicating the presence of challenging cell subpopulations in real-world scenarios (Supplementary Figs 15, 17, and 19). Conversely, Camp[59], Li[60], and Yan[61] scRNA-seq data already exhibited discernible cell subpopulations using all the features (Supplementary Figs 16, 18, and 20), allowing for evaluation of subtype detection methods to select a minimal meaningful feature set to retain the true cell heterogeneity. In both scenarios, we performed a comprehensive comparison of PHet against 24 different methods (Table 1) for cell subpopulation identification across multiple performance metrics including F1 score for identifying DE features, adjusted Rand index, adjusted mutual information, homogeneity, completeness, and V-measure for discovered subtypes. The evaluated methods encompassed four DE feature analysis tools and their variants: the standard Student t-test[62], Wilcoxon rank-sum test[63], Kolmogorov–Smirnov test[64], and LIMMA[65]. Additionally, we considered a dispersion-based method and its variants[45], as well as an IQR-based approach and its variations[47]. Furthermore, seven outlier detection algorithms were included in the evaluation: COPA[66], OS[38], ORT[39], MOST[40], LSOSS[41], DIDS[42], and DECO[43]. We also assessed the performance of PHet using the ΔDispersion metric instead of ΔIQR. For more details about these methods, please refer to Section 3.

When it comes to discriminative performance, PHet outperformed established outliers detection methods (e.g., DECO) across all scRNA-seq datasets with a mean F1 score exceeding 85% (Fig. 2(a)). These F1 scores were determined by comparing the top 100 DE features identified from LIMMA[65] with the top 100 features selected by each method for each scRNA-seq dataset (see Section 3 for details on metrics). Notably, PHet was able to strike a balance between minimizing the number of selected features with an average of fewer than 450 features and maximizing the discriminative performance for all single-cell transcriptomic datasets (Fig. 2(b)). Since the DE features were derived from LIMMA, the average F1 score for LIMMA, is 1. A basic t-statistic method obtained comparable results to LIMMA which is a more advanced statistical model. This indicates that the top 100 differentially expressed features are consistent across these two methods. By fitting feature statistics using the gamma distribution, most DE-based methods including LIMMA struggled to match PHet's performance. This is likely due to the improper estimation of feature statistics, leading to highly skewed feature scores. PHet reduces the skewness of feature distributions by computing three statistics for each feature: the ΔIQR value, the Fisher's score, and the feature discriminatory power.

In terms of average adjusted Rand index (ARI) which measures the agreement between ground truth and clustering results, PHet outperformed existing methods and achieved $> 10\%$ gain over the second-best performing algorithm, KS+Gamma (Fig. 2(c)), and over 15% and 20% increase from PHet (ΔDispersion) and DECO, respectively. This suggests that the IQR-based approach is more effective than dispersion-based feature selection, leading to improved clustering quality and better recovery of cell types. The KS test is useful for detecting DE features because this test does not require any assumptions about the shape or parameters of data distributions. Instead, the KS test compares the cumulative distributions of features, which makes it sensitive to any changes in the data[67]. This sensitivity allows the KS to effectively detect variations in expression levels that other tests (e.g., t-statistic and Wilcoxon rank-sum test) may not be able to capture. PHet leverages the benefits from both ΔIQR and KS, leading to better clustering results as indicated by multiple performance metrics, such as adjusted mutual information, homogeneity, completeness, and V-measure, across six scRNA-seq datasets (Supplementary Fig. 14).

Validating against ground truth cell types for scRNA-seq datasets, UMAP visualizations (using only selected features for each method) demonstrated that PHet attained a high level of biological integrity across all transcriptomics datasets. Notably, among the 14 cell types in the Baron dataset (Supplementary Table 3), PHet selected 748 features associated with at least 7 clusters of cells (Fig. 2(g); Fig. 3(a)). This was accomplished by considering endocrine cells as the control group and non-endocrine cells as the case group. Previous work by Baron and colleagues[57] demonstrated that four cell types—alpha, beta, gamma, and delta—representing the endocrine cells, exhibit inherent hierarchical relationships due to their shared expression profiles. Specifically, the alpha and gamma cells were observed to form a closely related group. This implies a plausible functional interrelation in the regulation of glucose metabolism. While alpha cells release glucagon, elevating blood sugar levels, gamma cells secrete a pancreatic polypeptide that concurrently inhibits glucagon and insulin secretion. Their co-localization may allow them to coordinate their actions to maintain glucose homeostasis[68]. Furthermore, delta cells were positioned as a group connected with the alpha/gamma cell group in the subsequent hierarchy, followed by the integration of beta cells into this configuration. Remarkably, among compared methods, PHet exclusively managed to preserve this hierarchical relationship among the endocrine cells. In addition, UMAP visualization using PHet's features revealed that ductal and endothelial cells were distantly localized, consistent with the hierarchical model suggested by Baron and colleagues[57]. No other algorithms were able to accurately delineate these distinct cell populations in this dataset (Fig. 3(b)-(c); Supplementary Fig. 15).

PHet algorithm also demonstrated its capability to deliver high-quality clustering results for Darmanis[58] and Patel[52] datasets (Fig. 2(i) and (k); Fig. 3(d) and (g); Supplementary Figs 17 and 19). Comparatively, the second and third-best methods exhibited less distinct clusters (Fig. 3(e)-(f) and (h)-(i)). In the analysis of the Camp dataset[59], several DE methods (e.g., t-statistic, Wilcoxon, KS, LIMMA), ΔIQR, and DECO were able to recover the five distinct cell subsets (Supplementary Fig. 16). However, these methods required the inclusion of nearly all features to achieve this outcome (Supplementary Fig. 14). In contrast, the PHet method was able to reach the same conclusion using a significantly smaller number of selected features (~ 313) (Fig. 2(h); Supplementary Fig. 16). Similar trends were also noted in the analysis of Li[60] and Yan[61] scRNA-seq datasets (Supplementary Figs 18 and 20).

We extended our evaluation to include 11 well-known microarray gene expression datasets to determine PHet's performance in discovering patient subtypes (Section 3). For each dataset, patient samples were divided into control and case groups based on their clinical diagnosis (Supplementary Table 2). PHet significantly outperformed other baseline methods with an average ARI score exceeding 60% while selecting fewer features (> 130 on average; Fig. 2(d)-(f)). For example, in the analysis of the MLL dataset[69] (Supplementary Table 2), a total of 72 patient samples were examined. These samples were categorized into three distinct types of leukemia: acute lymphoblastic leukemia (ALL), mixed-lineage leukemia (MLL), and acute myeloid leukemia (AML). We assigned the ALL patient category in the control group while MLL and AML patients in the case group. PHet recognized three clusters of patients with an ARI of 88% that closely matched their true sample types. In contrast, none of the other algorithms, including the KS method, that we compared were capable of distinctly identifying these three patient subtypes (Fig. 3(j)-(l); Supplementary Fig. 12). This outcome confirms that PHet is a robust and reliable algorithm for detecting sample heterogeneity and subtypes. Similar observations were made for GSE1987[70] data (Fig. 3(m)-(o); Supplementary Fig. 10). On the remaining benchmark microarray datasets, PHet exhibited competitive to best results across multiple metrics (Fig. 3(p)-(r); Supplementary Figs 2-9, 11, and 13). In both sets of scRNA-seq and microarray datasets, PHet consistently demonstrated superior clustering quality compared to other baseline methods and across various metrics, such as F1 and the number of predicted features.

To further explore the impact of PHet's components on feature selection for subtype detection, ablation experiments were conducted on the same microarray and single-cell transcriptomics datasets (Supplementary Figs 30-31). These experiments involved systematically removing and reintegrating three key statistical components: Fisher's scores, feature discriminatory power, and the absolute maximum values of ΔIQR (Section 2.2). The outcomes of these experiments revealed that utilizing only ΔIQR in PHet led to suboptimal ARI scores. This suggests that the exclusive reliance on ΔIQR limits the discriminatory power in extracting DE features (Section 3), thereby hindering the accurate delineation of subtypes within conditions. In contrast, when solely utilizing the Fisher's method, notable improvements were observed, with gains of 10% and $< 30\%$ on microarray and scRNA datasets, as measured by the ARI metric. These improvements can be attributed to the subsampling process, which was able to capture variations in condition and case samples. Importantly, when all three components were incorporated, the most favorable outcomes were achieved, with ARI scores of more than 60% and 75% on microarray and scRNA datasets, respectively. These results underscore the significance of feature scores derived from discriminatory and IQR components, highlighting their valuable contribution to subtype detection.

2.4 Evaluation of PHet's discriminative performance on simulated data

While PHet exhibited competitive discriminative performance in our previous benchmark test, the evaluation was confined to comparing the selected features with LIMMA-based DE features. To bolster the validation of PHet's discriminative capabilities, we conducted a comprehensive evaluation of 25 algorithms (outlined in Section 3 and Table 1) using two sets of simulated data, specifically tailored to test DE capability of each algorithm. These datasets correspond to the 'minority' and 'mixed' model schemes, as proposed by Campos and colleagues[43], and are designed to capture sample heterogeneity under the supervised settings. In the 'minority' model, a small fraction of case samples exhibited changes in specific features, while the 'mixed' model displayed substantial intra-group variation in both case and control samples for those features. Each dataset consists of 40 samples, evenly split between control and case groups with 1100 features, including 93 differentially expressed (DE) features between the two groups. We generated 5 datasets based on the minority model, varying the proportions of perturbed samples in the case group from 5% (1 in 20) to 45% (9 in 20). Those perturbed samples were introduced by modifying the expressions of 100 randomly selected features, including DE ones. Similarly, we constructed five datasets following the same procedure for the mixed model, with perturbed samples evenly distributed between a subset of both case and control groups. Further details about these datasets are provided in Section 3. These simulated datasets are ideal to compare the performance of 25 algorithms under balanced conditions, where the numbers of case and control samples are the same. For quantitative analysis, the F1 score was used to compare the top 100 predicted features of each method with the top 100 true features. Furthermore, we meticulously recorded the number of informative features

yielded by each method. We ensured a fair and consistent experimental process by following the model-specific parameters recommended by the respective authors of each method, thus avoiding any bias or error stemming from inappropriate parameter choices. A best performing algorithm should attain high F1 scores across both model schemes, while also exhibiting the ability to predict a minimal set of important features.

Both PHet and PHet (ΔDispersion) retained a minimum set of significant features (at $\alpha < 0.01$) for both minority and mixed changes under all settings (Fig. 4(c) and (d); Supplementary Fig. 1(c) and (d)). This indicates that both methods have a robust ability to detect important features. The Dispersion (composite) method, which operates in an unsupervised approach, performed on par with PHet (F1 score of 0.92 on average) for both model schemes, achieving an F1 score of 0.88 on average. However, the supervised approach of the ΔDispersion+ΔMean method resulted in lower F1 scores (averaging 0.36) compared to the ΔIQR+ΔMean method, which achieved an average F1 score of 0.88 (Fig. 4(a) and (b); Supplementary Fig. 1(a) and (b)). This indicates that the IQR statistic is a better measure for capturing DE features than dispersion under binary experimental conditions. Interestingly, when comparing PHet (ΔDispersion) with PHet, both methods achieved similar F1 scores across both model schemes. This can be attributed to the fact that PHet integrates Fisher's scores and discriminative power for identifying DE features, explaining why PHet (ΔDispersion) obtained a high F1 score. A majority of DE-based methods, DIDS, and DECO demonstrated comparable performance with PHet, suggesting that PHet has a sufficient discriminative performance (Supplementary Fig. 1).

2.5 Analysis of PHet's ability to identify markers with low statistical power

The presence of outlier DE genes with low statistical power creates unique challenges in cancer studies, as they contribute to the observed heterogeneity in tumor samples. Specific algorithms such as COPA[42], categorized as "outlier detection", have been proposed to address this issue. In this study, we further assessed PHet's capability to identify DE features with low statistical power, referred to as outlier features. To achieve this, we evaluated the effectiveness of 25 algorithms, including seven outlier detection algorithms (Table 1), in identifying biomarkers under imbalanced settings using 1000 batches of HER2 (human epidermal growth factor receptor 2) data (Section 3; Supplementary Table 1). Each batch consisted of 188 samples, with 178 fixed HER2 non-amplified samples in the control group and 10 randomly drawn samples from 60 HER2-amplified samples in the case group, following the methodology described by de Ronde and colleagues[42]. The objective is to identify the true 20 biomarkers, located on ch17q12 or ch17q21 chromosome regions, from a pool of 27,506 features in each batch. It is worth noting that these 20 specific features have limited statistical power. The performance of 25 algorithms was evaluated using the F1 score, which measures the algorithm's ability to accurately identify the true 20 markers among its top predicted features. The study involved varying the number of top features from 1 to 50.

Despite not specifically designed to detect outlier features, PHet displayed a competitive performance with a mean F1 score ranging from 4% to 39% across batches (Fig. 4(e)). PHet was able to identify 10 true biomarkers within its top 20 features. This ability to detect true biomarkers sets PHet apart from dispersion based features selection and several DE and outlier detection algorithms, such as KS, Wilcoxon, OS, ORT, and COPA (Supplementary Fig. 1(e)). The performance of the DIDS method in terms of mean F1 score over batches was impressive, surpassing all other models with a score greater than 44%. However, DIDS demonstrated suboptimal performance in extracting features for subtype detection, especially when dealing with single-cell RNA-seq datasets (Section 2.3). This is not surprising considering that DIDS was primarily designed to extract features contributing to outliers in tumor samples, rather than for subtype detection. Although DECO has proven its ability to identify DE features under various complex scenarios, and outperformed DIDS in subtype detection (Section 2.3), its performance fell short and achieved a mean F1 score of less than 1% over 1000 batches. This underperformance can be attributed to low statistical power of 20 true markers, which impedes DECO's ability to accurately identify these features. Consequently, DECO prioritizes other strong features, which have distinct expression profiles that differ significantly from the features of interest.

2.6 PHet uncovers distinct differentiation lineages in airway epithelium

Preserving inherent complex relationships among cells is a fundamental challenge in the analysis of cell differentiation. To gain insights the mechanisms of cell differentiation, it is essential to capture the interactions and dependencies that occur among various cell types in terms of changes in gene expression. In order to assess the capability of PHet in detecting subpopulations constituting differentiation trajectories, we utilized two established scRNA-seq data of the respiratory airway epithelium[49]: 14,163 mouse tracheal epithelial cells (MTECs) and 2,970 primary human bronchial epithelial cells (HBECs). The MTECs dataset comprises cells collected from injured and uninjured mice at different time points (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 consisting of secretory, ciliated, and tuft cells, as well as other rare cell populations, such as PNECs, brush cells, and pulmonary ionocytes[49,71,72]. For both datasets, basal

cells were considered as the control group while the remaining cell types were grouped under the case category (Supplementary Table 4). Similar to the results obtained using pre-annotated markers in the previous study[49] (Fig. 5(a) and (h); Supplementary Figs 21(a) and 22(a)), the UMAP visualizations of PHet's features revealed distinct cell clusters including basal and secretory cells, within both HBECs and MTECs data (Fig. 5(b) and (i); Supplementary Figs 21(b) and 22(b)). These visualizations also identified clusters corresponding to rare cell populations encompassing brush and PNECs, pulmonary ionocytes, and SLC16A7.

To assess the role of PHet's HD features in delineating cell differentiation trajectories, we performed SPRING visualization[73]. SPRING is a graph-based method that constructs a cell similarity network based on their expression profiles. This aids in uncovering the structures underlying cellular differentiation trajectories. We employed the known pre-annotated markers and PHet features to generate the SPRING plots from HBECs and MTECs datasets. This comparative analysis allowed us to examine which approach more effectively captures the differentiation trajectories of cell populations. While the pre-annotated markers and PHet features displayed similar cell differentiation processes, PHet exclusively revealed the presence of two distinct trajectories for the HBECs spanning basal-to-luminal differentiation, including rare cells (Fig. 5(c)-(d); Supplementary Fig. 21(c)-(d)). By examining the metadata of HBECs, we found that each trajectory was closely linked to a subset of donors (Fig. 5(e); Supplementary Fig. 21(e)). For instance, cells in the black cluster were specifically associated with Donor 1 that exhibited enrichment in cytokeratin genes KRT4 and KRT13[74] (Fig. 5(g); Supplementary Fig. 21(h)). These cells could potentially serve as progenitors initiating luminal differentiation. Furthermore, an assessment of the relative abundance of cell types revealed that the secretory and basal to secretory cells were more prevalent in the red cluster (Donors 2 and 3; Fig. 5(f); Supplementary Fig. 21(f)). These cells have a significantly higher expression of the CYP2F1 gene compared to those in the black cluster (Fig. 5(g), Supplementary Fig. 21 (g)-(h)). The CYP2F1 gene encodes a cytochrome P450 enzyme crucial for the metabolism of xenobiotics and endogenous compounds, aligning with the primary roles of the secretory club cells[75]. Moreover, we also pinpointed several genes with diverse functions that were differentially expressed between the clusters (Fig. 5(g)). These include a member of a BPI fold protein (BPIFA1) that plays a role in the innate immune responses of the conducting airways[76] (Fig. 5(g); Supplementary Fig. 21 (g)-(h)).

UMAP visualizations of the MTECs data using both pre-annotated markers and PHet features yielded comparable observations regarding cell types (Fig. 5(h)-(i); Supplementary Fig. 22 (a)-(b)). Similar to the HBECs analysis, when SPRING was applied to both known pre-annotated markers and PHet features, two identified distinguishable cell trajectories that emerged within the MTECs dataset (Fig. 5(j)-(k); Supplementary Fig. 22 (c)-(d)). Interestingly, the SPRING plot based on PHet's features exhibited more clear clustering outcomes compared to the one based on the pre-annotated markers. This suggests that PHet's features encapsulate more comprehensive information concerning cell states and transitions. Subsequently, we characterized cell populations within these two trajectories (Fig. 5(l); Supplementary Fig. 22 (e)) using cell-specific gene signatures. The lower cluster of cells was enriched with basal cells while the upper cluster contained a significant proportion of cycling basal cells ($> 20\%$) (Fig. 5(m); Supplementary Fig. 22 (f)). This implies a regenerative function for these cells, responding to injury by undergoing proliferation and differentiation into other cell types (Supplementary Figs 22(g)-(j) and 23). Moreover, the upper cluster contains abundant transitional cells uniquely expressing KRT4 and KRT13 genes. These cells may represent an intermediate population positioned between tracheal basal stem cells and differentiated secretory cells as suggested by previous studies[49,71,74]. Taken together, these results suggest that PHet's feature selection approach was much more effective in identifying HD features that helped discovering cell subpopulations for HBECs and MTECs in comparison to the manual marker-based analysis.

2.7 PHet effectively identifies subpopulations of basal cells in the MTECs dataset

Cell subtype identification is one of the most fundamental applications within the realm of single-cell data analysis. This endeavor involves assigning each cell to a specific group based on its feature expression profile, thereby shedding light on the heterogeneity and diversity inherent within cell populations in complex biological contexts, such as tissues, organs, or tumors. Moreover, the process of cell subtype identification holds great potential to uncover new biomarkers, while enhancing the understanding of cellular functions and interactions. Hence, it becomes paramount to investigate PHet's features for the discovery of cell subtypes. In this specific case study, we re-examined the heterogeneity of basal cells in the mouse tracheal epithelial cells (MTECs)[49]. The research findings reported by Carraro and colleagues[77] served as a catalyst for our exploration into the existence of potential basal cell subtypes in MTECs data. We used PHet's features to detect basal cell clusters and contrasted them with dispersion based HV features and pre-annotated basal cell markers[49]. To perform the clustering analysis, the Leiden community detection algorithm from the SCANPY package[44] was utilized, and the resolution hyperparameter was fixed to 0.5. The cluster quality was evaluated using the silhouette score, which measures how well each cell belongs to its assigned cluster. A higher silhouette score implies a better quality of clustering.

PHet based features revealed four distinct clusters of basal cells (Fig. 6(a); Supplementary Figs 24 and

27(a)), thereby achieving the highest silhouette score (47%) (Fig. 6(i); Supplementary Fig. 27(c)). Each basal cluster is characterized by distinct gene expression profiles and biological functions, as observed by Carraro and colleagues[77]. Basal-1 and Basal-3 clusters for PHet exhibited elevated expression of the canonical basal cell markers, including TPR63 (tumor protein P63) and the cytokeratin 5 (KRT5) (Fig. 6(m) and (o); Supplementary Fig. 27(d)-(f)). Conversely, cells in Basal-2 and Basal-4 clusters exhibited the reduced expression of the basal cell markers and showed enrichment for SCGB3A2/BPIFA1 for Basal-2 and MSLN/AGR2 and members of the serpin family (e.g. TSPAN1) for Basal-4, respectively (Fig. 6(m) and (o); Supplementary Fig. 27(d)-(f)). This finding suggests that these clusters represent two distinct basal cell subtypes undergoing transitions toward a luminal secretory phenotype (Supplementary Figs 24 and 27(b)-(c)). Of note, BPIFA1 was expressed in Basal-2 predominantly and was also differentially expressed between the two donor groups in the HBECs dataset (Section 2.6; Fig. 6(g)), providing further evidence that these cells are undergoing differentiation into secretory cells. A previous study has established a connection between BPIFA1 secretion and the appearance of secretory cells during mucociliary differentiation of airway epithelial cells[78]. PHet effectively identified the specific basal cell subtype linked to this phenomenon. It also underscores that PHet's features can accurately reflect the developmental trajectories of cell differentiation.

In contrast to the results obtained from PHet, the pre-annotated markers displayed a less distinct four basal subtypes, with a relatively low silhouette score of 41% (Fig. 6(b) and (i); Supplementary Fig. 28(a)). This suggests that the pre-annotated markers may not accurately capture the unique characteristics associated with basal subtypes. One particular challenge was observed in identifying the Basal-3 subpopulation, as it appeared to be situated between the Basal-1 and Basal-4 subtypes (Fig. 6(b)). This arrangement made it difficult to identify unique characteristics associated with the Basal-3 cells (Supplementary Fig. 28(d)-(f)). Moreover, cells in the Basal-2 cluster exhibited a complex expression pattern, where both basal and secretory cell markers are being enriched in these cells, while simultaneously losing expression of BPIFA1 (Fig. 6(h), (n), and (p); Supplementary Figs 25 and 28(d)-(f)). This presents a challenge in accurately interpreting biologically meaningful signals related to cells in this cluster. These findings highlight the necessity for updating the annotated markers to characterize basal cell subtypes. The current markers may not adequately capture the diversity and distinctiveness of these subpopulations. The dispersion based HV features displayed three basal clusters with a silhouette score below 40% (Fig. 6(c) and (i); Supplementary Fig. 29(a)). A subpopulation of basal cells is observed to have a mixture of signatures from both Basal-2 and Basal-3, so it is referred to as Basal-2/3 (Supplementary Figs 26 and 29(d)-(f)). Overall, these HV features were not adequate in delineating basal subpopulations.

In an attempt to gain insights into the occurrence of basal subtypes in both injured and uninjured conditions of MTECs data (Fig. 6(d)-(f)), we analyzed the distribution of basal cell clusters. Leveraging the PHet's features, the Basal-4 population was observed to be almost exclusively present in the injured condition (Fig. 6(j)). In contrast, when examining the pre-annotated markers, we found that a small fraction (5%) of the Basal-4 population was present in the uninjured case (Fig. 6(k)), reassuring that these markers are insufficient for uncovering accurate subtypes at a higher resolution. The HV features constituted three clusters of basal, but their significance and biological relevance are unclear (Fig. 6(l)).

3 Discussion

Subtype discovery within ever-increasing omics expression datasets, encompassing various experimental or biological conditions, is crucial for exploring tissue heterogeneity, unraveling cellular differentiation pathways, and uncovering the molecular signatures associated with diverse biological states or complex diseases such as 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 imposes several challenges associated with noise, sparseness, heterogeneity, or confounding factors. Therefore, a preliminary and essential step is to select features that are linked to a specific subtype. Many existing methods have primarily concentrated on identifying DE features, which are features showing distinct expression levels across specified biological conditions. In recent single-cell RNA-seq analyses, a novel category called HV features has been introduced, representing genes with high variability irrespective of conditions. In this paper, we propose a new feature set termed HD features, which play a crucial role in capturing the heterogeneity and diversity of subtypes while remaining discriminative between known biological conditions. Through our deep metric learning, HD features are characterized by their comprehensive representation, encompassing both mean expression differences and variability (IQR) discrepancies between two conditions.

Several approaches, such as DIDS[42] and DECO[43], have been developed to extract relevant features from expression data and to identify subtypes. However, these methods have substantial limitations in capturing subtype-related features because they focus on a specific type of outlier features, which means they may miss important information that could be useful for subtyping. PHet overcomes these drawbacks by using delta IQR,

iterative 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 heterogeneity. Based on our benchmark studies, PHet significantly outperformed the existing methods by retaining a minimum set of features while ensuring optimal clustering quality for subtypes detection.

It is important to note that PHet goes beyond mere integration with downstream analysis tools, such as Milo[33], SCANPY[44], and scVI[79]. This framework stands as a comprehensive solution for tracking cellular changes across diverse experimental conditions. While it excels in handling binary conditions, its flexibility allows for extension 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[80–83].

The establishment of comprehensive atlas datasets for various organisms and tissues, exemplified by initiatives like the Human Cell Atlas[84] and the Mouse Cell Atlas[85], has paved the way for discovering and characterizing novel cell types[86–88]. These endeavors provide invaluable insights into the diversity and function of cellular phenotypes across diverse biological contexts and conditions, illuminating how cell populations change in disease conditions[89,90]. PHet stands poised to make a substantial contribution to this initiative, amplifying the discovery of unknown cell subtypes and offering profound insights into the cellular diversity within healthy and diseased tissues. This is achievable through its in-depth analysis of large-scale omics expression data at single-cell resolution, aligning seamlessly with the objectives of comprehensive atlas initiatives.

References

1. Rabbitts, T. H. Chromosomal translocations in human cancer. *Nature* **372**, 143–149 (1994).

2. Hogenbirk, M. A. *et al.* Defining chromosomal translocation risks in cancer. *Proceedings of the National Academy of Sciences* **113**, E3649–E3656 (2016).

3. Saria, S. & Goldenberg, A. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems* **30**, 70–75 (2015).

4. Lyons, Y. A., Wu, S. Y., Overwijk, W. W., Baggerly, K. A. & Sood, A. K. Immune cell profiling in cancer: molecular approaches to cell-specific identification. *NPJ precision oncology* **1**, 26 (2017).

5. Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer* **108**, 479–485 (2013).

6. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology* **15**, 81–94 (2018).

7. Collisson, E. A., Bailey, P., Chang, D. K. & Biankin, A. V. Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology* **16**, 207–220 (2019).

8. Lomberk, G. *et al.* Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nature communications* **9**, 1978 (2018).

9. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences* **109**, 2724–2729 (2012).

10. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research* **12**, 1–18 (2010).

11. Liang, C., Shang, M. & Luo, J. Cancer subtype identification by consensus guided graph autoencoders. *Bioinformatics* **37**, 4779–4786 (2021).

12. Alizadeh, A. A. *et al.* Toward understanding and exploiting tumor heterogeneity. *Nature medicine* **21**, 846–853 (2015).

13. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).

14. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature reviews cancer* **12**, 323–334 (2012).

15. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

16. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science* **286**, 531–537 (1999).

17. Mirnezami, R., Nicholson, J. & Darzi, A. Preparing for precision medicine. *N Engl J Med* **366**, 489–491 (2012).

18. Yu, K.-H. & Snyder, M. Omics profiling in precision oncology. *Molecular & Cellular Proteomics* **15**, 2525–2536 (2016).

19. Li, Y., Mansmann, U., Du, S. & Hornung, R. Benchmark study of feature selection strategies for multi-omics data. *BMC bioinformatics* **23**, 1–18 (2022).

20. Li, J. *et al.* Feature selection: A data perspective. *ACM computing surveys (CSUR)* **50**, 1–45 (2017).

21. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nature Precedings* 1–1 (2010).

22. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The nbp negative binomial model for assessing differential gene expression from rna-seq. *Statistical applications in genetics and molecular biology* **10** (2011).

23. Hardcastle, T. J. & Kelly, K. A. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics* **11**, 1–14 (2010).

24. Auer, P. L. & Doerge, R. W. A two-stage poisson model for testing rna-seq data. *Statistical applications in genetics and molecular biology* **10** (2011).

25. Leng, N. *et al.* Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).

26. Van De Wiel, M. A. *et al.* Bayesian analysis of rna sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14**, 113–128 (2013).

27. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in rna-seq: a matter of depth. *Genome research* **21**, 2213–2223 (2011).

28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).

29. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **26**, 139–140 (2010).

30. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3** (2004).

31. Gao, X., Hu, D., Gogol, M. & Li, H. Clustermap: compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics* **35**, 3038–3045 (2019).

32. Skinnider, M. A. *et al.* Cell type prioritization in single-cell data. *Nature biotechnology* **39**, 30–34 (2021).

33. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology* **40**, 245–253 (2022).

34. Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scrna-seq data. *Proceedings of the National Academy of Sciences* **118**, e2100293118 (2021).

35. Vandenbon, A. & Diez, D. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications* **11**, 4318 (2020).

36. Kim, C., Lee, H., Jeong, J., Jung, K. & Han, B. Marcopolo: a method to discover differentially expressed genes in single-cell rna-seq data without depending on prior clustering. *Nucleic acids research* **50**, e71–e71 (2022).

37. Tomlins, S. A. *et al.* Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *science* **310**, 644–648 (2005).

38. Tibshirani, R. & Hastie, T. Outlier sums for differential gene expression analysis. *Biostatistics* **8**, 2–8 (2007).

39. Wu, B. Cancer outlier differential gene expression detection. *Biostatistics* **8**, 566–575 (2007).

40. Lian, H. Most: detecting cancer differential gene expression. *Biostatistics* **9**, 411–418 (2008).

41. Wang, Y. & Rekaya, R. Lsoss: detection of cancer outlier differential gene expression. *Biomarker insights* **5**, BMI–S5175 (2010).

42. de Ronde, J. J., Rigaill, G., Rottenberg, S., Rodenhuis, S. & Wessels, L. F. Identifying subgroup markers in heterogeneous populations. *Nucleic acids research* **41**, e200–e200 (2013).

43. Campos-Laborie, F. J. *et al.* Deco: decompose heterogeneous population cohorts for patient stratification and discovery of sample biomarkers using omic data profiling. *Bioinformatics* **35**, 3651–3662 (2019).

44. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

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

46. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053–1058 (2018).

47. Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. *A Modern Introduction to Probability and Statistics: Understanding why and how*, vol. 488 (Springer, 2005).

48. Fisher, R. A. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, vol. 22, 700–725 (Cambridge University Press, 1925).

49. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

50. Hoffer, E. & Ailon, N. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, 84–92 (Springer, 2015).

51. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology* **37**, 38–44 (2019).

52. Patel, A. P. *et al.* Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

53. Ho, J. W., Stefani, M., Dos Remedios, C. G. & Charleston, M. A. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* **24**, i390–i398 (2008).

54. Hasan, M. A. M., Maniruzzaman, M. & Shin, J. Differentially expressed discriminative genes and significant meta-hub genes based key genes identification for hepatocellular carcinoma using statistical machine learning. *Scientific Reports* **13**, 3771 (2023).

55. Babu, G. J. Subsample and half-sample methods. *Annals of the Institute of Statistical Mathematics* **44**, 703–720 (1992).

56. Hall, P. & Martin, M. A. On bootstrap resampling and iteration. *Biometrika* **75**, 661–671 (1988).

57. Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems* **3**, 346–360 (2016).

58. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112**, 7285–7290 (2015).

59. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).

60. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics* **49**, 708–718 (2017).

61. Yan, L. *et al.* Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* **20**, 1131–1139 (2013).

62. Cramér, H. *Mathematical methods of statistics*, vol. 26 (Princeton university press, 1999).

63. Datta, S. & Satten, G. A. Rank-sum tests for clustered data. *Journal of the American Statistical Association* **100**, 908–915 (2005).

64. Berger, V. W. & Zhou, Y. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online* (2014).

65. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**, e47–e47 (2015).

66. MacDonald, J. W. & Ghosh, D. Copa-cancer outlier profile analysis. *Bioinformatics* **22**, 2950–2951 (2006).

67. Khan, A. A. *et al.* Transfection of small rnas globally perturbs gene regulation by endogenous micrornas. *Nature biotechnology* **27**, 549–555 (2009).

68. Röder, P. V., Wu, B., Liu, Y. & Han, W. Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine* **48**, e219–e219 (2016).

69. Armstrong, S. A. *et al.* Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics* **30**, 41–47 (2002).

70. Dehan, E. *et al.* Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. *Lung cancer* **56**, 175–184 (2007).

71. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes cftr-expressing ionocytes. *Nature* **560**, 319–324 (2018).

72. Deprez, M. *et al.* A single-cell atlas of the human healthy airways. *American journal of respiratory and critical care medicine* **202**, 1636–1645 (2020).

73. Weinreb, C., Wolock, S. & Klein, A. M. Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).

74. Hewitt, R. J. & Lloyd, C. M. Regulation of immune responses by the airway epithelial cell landscape. *Nature Reviews Immunology* **21**, 347–362 (2021).

75. Davis, J. D. & Wypych, T. P. Cellular and functional heterogeneity of the airway epithelium. *Mucosal immunology* **14**, 978–990 (2021).

76. Liu, Y. *et al.* Splunc1/bpifa1 contributes to pulmonary host defense against klebsiella pneumoniae respiratory infection. *The American journal of pathology* **182**, 1519–1531 (2013).

77. Carraro, G. *et al.* Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. *Nature medicine* **27**, 806–814 (2021).

78. Clifton, C. *et al.* Bpifa1 is a secreted biomarker of differentiating human airway epithelium. *Frontiers in Cellular and Infection Microbiology* **12**, 1035566 (2022).

79. Gayoso, A. *et al.* A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology* **40**, 163–166 (2022).

80. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

81. Huang, M. *et al.* Saver: gene expression recovery for single-cell rna sequencing. *Nature methods* **15**, 539–542 (2018).

82. Li, W. V. & Li, J. J. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications* **9**, 997 (2018).

83. Finak, G. *et al.* Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology* **16**, 1–13 (2015).

84. Regev, A. *et al.* The human cell atlas. *elife* **6**, e27041 (2017).

85. Han, X. *et al.* Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107 (2018).

86. Brbić, M. *et al.* Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods* **17**, 1200–1206 (2020).

87. Wang, S. *et al.* Leveraging the cell ontology to classify unseen cell types. *Nature communications* **12**, 5556 (2021).

88. Dohmen, J. *et al.* Identifying tumor cells at the single-cell level using machine learning. *Genome Biology* **23**, 123 (2022).

89. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865–868 (2017).

90. Datlinger, P. *et al.* Pooled crispr screening with single-cell transcriptome readout. *Nature methods* **14**, 297–301 (2017).

91. Weinberger, K. Q. & Saul, L. K. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research* **10** (2009).

92. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* **18**, 559–563 (2017).

93. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018).

94. Abadi, M. *et al.* {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283 (2016).

95. Wan, X., Wang, W., Liu, J. & Tong, T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology* **14**, 1–13 (2014).

96. Etikan, I. & Bala, K. Sampling and sampling methods. *Biometrics & Biostatistics International Journal* **5**, 00149 (2017).

97. Lehmann, E. L. & D'Abrera, H. J. *Nonparametrics: statistical methods based on ranks.* (Holden-day, 1975).

98. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **14** (2001).

99. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**, 433–459 (2010).

100. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

101. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

102. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods* **17**, 261–272 (2020).

103. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467 (2001).

104. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

105. Dembélé, D. A flexible microarray data simulation model. *Microarrays* **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. An uppercase calligraphy character (e.g., $\mathcal{X}$) indicates a set.

Data preprocessing

Omics data, denoted by $\mathbf{M} \in \mathbb{R}^{n+m,p}$, takes expression profile (e.g., gene) as the input and is 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. It is important to note that throughout the manuscript, the terms "samples" and "cells" were utilized interchangeably.

Deep metric learning for feature embedding

Deep metric learning (DML) aims to learn a distance metric that can measure the similarity or dissimilarity between data samples[91]. In the context of triplet loss that considers three types of sample types: positive, negative, and an anchor, the goal of DML is to maximize the distance between the anchor-negative samples while minimizing the distance between the anchor-positive samples by a predetermined margin. This way, DML can generate low-dimensional embeddings that effectively represent the original high-dimensional features. In our study, we applied DML to our preprocessed data $\mathbf{M}$ to learn feature embeddings within each disease or cell condition. To achieve this, we first transposed the input data, and then sorted the values of each sample in ascending order. This has resulted in a modified omics data that encompasses both control and case conditions. However, due to the uneven size of control and case samples in $\mathbf{M}$, we have utilized the "RandomUnderSampler" function from the `imblearn` package[92] to downsample from a condition that had a larger sample size. This allowed us to balance the dataset and ensure that our results were not biased towards one condition.

Then, we employed UMAP (Uniform Manifold Approximation and Projection)[93] to obtain two-dimensional embeddings for features. The minimum distance parameter and the number of neighbors for UMAP were set to 0 and 15, respectively. After we acquired the two-dimensional embeddings, we applied clustering using the k-means algorithm to partition features into a predetermined number of clusters (default is 200). Subsequently, clustering was applied to these two-dimensional embeddings using the K-means algorithm with the number of clusters being set to 200. The resulting clustering labels were then used to construct triplets for deep metric learning. Each triplet consisted of an anchor (a), a positive (p), and a negative (n) features. The process of constructing triples is as follows. For each anchor feature, we randomly selected a positive feature p from a set of features that have the same cluster label as a and randomly chose a negative feature n from features that have a different cluster label from a. These triplets are fed to DML to learn feature embeddings using the triplet loss function:

$$L(a, p, n) = \max(d(a, p) - d(a, n) + m, 0) \tag{1}$$

This function utilizes the Euclidean distance d between anchor, positive, and negative samples, along with a margin hyperparameter m (default is 1) to control the separation between similar and dissimilar pairs. To optimize the triplet loss function, the semi-hard triplet loss is used. This method selects triplets in which the negative feature is farther from the anchor than the positive feature, but still within a margin, i.e. $d(a, p) < d(a, n) < d(a, p)+m$. By doing so, meaningful embeddings can be generated that accurately distinguish between similar and dissimilar pairs. The semi-hard triplet loss ensures that the negative feature is neither too easy nor too hard, which further enhances the quality of the embeddings.

The architecture of our DML is a simple fully connected three-layer neural network. The input layer has a dimension of 25, followed by a hidden layer with a dimension of 10, and an output or embedding layer with a dimension of 2. To optimize the learning process for DML with triplet loss, we employed a mini-batch strategy with a batch size of 128. The Adam optimizer is utilized, and the activation function used is ReLU. We trained the network for 200 epochs to ensure convergence and optimal performance. The resulted embeddings, which have a dimension of 2, were used to calculate the Euclidean distance between any two features belonging to the same class (Fig. 1(a)). The entire process starting from sorting value was repeated 30 rounds to account for uneven sample size between case and control groups and the Euclidean distances were averaged. The implementation of our DML system is based on the TensorFlow framework[94].

PHet framework

PHet is a framework that identifies features that can separate different groups of samples in a control-case study. PHet accomplishes this goal in a pipeline procedure that is composed of six steps (see Fig. 1(j)): (1)-

iterative subsampling, (2)- Fisher's combined probability test, (3)- measuring features discriminatory power, (4)- computing feature statistics, (5)- features ranking and selection, and (6)- dimensionality reduction and clustering.

Iterative subsampling

Given an annotated omics dataset ($\mathbf{M}$), PHet computes p-values of mean differences and 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 suggests that the observed difference is unlikely to be explained by chance, indicating that the feature is differentially expressed[21]. Conversely, a p-value close to 1 indicates that the observed difference is likely due to chance and is not statistically significant.

The estimation of p-values is influenced by two main factors: the sample size of the groups and the spread of samples within each group. Generally, having an uneven sample size between the two groups can introduce bias and affect the accuracy of the p-value calculation. To overcome this problem, 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. These p-values are stored in a matrix, denoted as $\mathbf{P}$, which has dimensions $p \times t$, where p represents the number of features and t represents the number of iterations. 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 extract features by their biological relevance, as they do not fully capture the heterogeneity of changes in expression for subtype detection. To address this issue, PHet utilizes the absolute difference of interquartile range (ΔIQR) between each condition as a measure of data variability. The IQR is defined as the difference between the first and third quartiles of the data, denoted as q_1 and q_3 respectively[95]. While a high IQR, for a feature, indicates a large variation of expression values within each condition, a high ΔIQR indicates a large dispersion difference of expression values between control and case samples. It is important to note that a feature may have a near-zero ΔIQR value even if the IQR values for each condition are high. This can impact the identification of important features for subtype detection. The iterative subsampling process employed by PHet ensures that every sample in the omics data has an equal opportunity to be included in the subset[96]. This allows the selection of important features, which may have high ΔIQR values after subsampling. By considering both p-values and ΔIQR, PHet aims to overcome the limitations of p-values alone and provide a more comprehensive approach for feature extraction in subtype detection. The formula for ΔIQR of a subset s is:

$$\Delta\text{IQR}^s = |\text{IQR}(\mathbf{X}^s) - \text{IQR}(\mathbf{Y}^s)| \tag{2}$$

The ΔIQR values are stored in $\mathbf{R} \in \mathbb{R}^{p \times t}$, and then the max operation for each feature is applied to obtain a vector of size p, i.e., $\mathbf{r} \in \mathbb{R}^p$. This vector and the corresponding p-values ($\mathbf{P}$) serve as inputs in the PHet pipeline.

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[48]:

$$\mathbf{f} = -2 \times \text{diag}(\log \mathbf{P} \log \mathbf{P}^\intercal) \tag{3}$$

where `diag` corresponds the matrix-to-vector diagonal operation. The Fisher's scores, represented by $\mathbf{f} \in \mathbb{R}^p$, provide a measure of differential consistency of probabilities across t number of independent tests or iterations. When p-values for a feature, across t iterations, are small, the corresponding Fisher's score will be large, implying that this feature is likely to be differentially expressed, and vice versa. By ranking and filtering out the features with low Fisher scores, we can reduce the dimensionality of the omics data and obtain a smaller set of DE features. The output of this step is a set of Fisher's combined probability scores ($\mathbf{f}$).

Determining features discriminatory power

The iterative subsampling may introduce bias in Fisher's scores for some features, leading to unreliable feature statistics, as alluded in Section 2.2. Therefore, PHet uses the nonparametric two-sample Kolmogorov–Smirnov (KS) test between control and case samples to readjust the Fisher's values. The KS test identifies a set of features by measuring the maximum difference for each feature between the two cumulative distributions

without considering the type of distribution of control and case samples[97]. To address the potential disparity in sample sizes, PHet randomly permutes both sample groups and employs a fixed sample size, which is determined as the nearest integer to the square root of the smaller sample size between the control and case groups. Then, PHet compartments (aka binning) p-values into a fixed number of intervals $b \in \mathbb{N}$ and assigns each interval $i \in b$ a weight $w_i \in [0, 1]$. These weights are then associated with a specific subset of features falling within that particular interval. This weighting mechanism serves as an indicator of the discriminatory strength of each feature, enabling PHet to identify and prioritize the most relevant features for subtype detection:

$$\mathbf{w} = \{w_i : 0 \leq w_i \leq 1\}_{i \in \{1,\ldots,b\}} \text{ and } \mathbf{w}^\intercal 1 = 1 \tag{4}$$

By default, the number of bins is set to four and $\mathbf{w} = [0.4, 0.3, 0.2, 0.1]$. Finally, PHet obtains the feature profile matrix as:

$$\mathbf{o} = \widehat{\mathbf{O}}\mathbf{w} \tag{5}$$

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 with only one element equal to 1 and the rest equal to 0. This element indicates the discriminatory strength that the corresponding feature displays. The output of this step is a vector representing the discriminatory power of each feature ($\mathbf{o}$).

Estimating feature statistics

In this step, PHet computes overall feature scores, referred to as the $\mathbf{s}$ statistics, using the outputs from previous steps: (1)- the maximum ΔIQR scores ($\mathbf{r}$), which measures the difference in interquartile range between conditions, (2)- the Fisher's scores ($\mathbf{f}$), which indicates how discriminant each feature is, and (3)- the features discriminatory power ($\mathbf{o}$), which links features to its discriminatory weight:

$$\mathbf{s} = \mathbf{r} + (\mathbf{f} \odot \mathbf{o}) \tag{6}$$

where $\odot$ corresponds to the Hadamard (element-wise) product. As can be seen in Eq. 6, $\mathbf{o}$ also adjusts Fisher's scores to prevent the overestimation of DE features. The matrix $\mathbf{s} \in \mathbb{R}^p$ holds the final statistics, where a high feature score indicates its importance for subtype detection.

Features selection

The statistics, $\mathbf{s}$, from the previous step are used to rank and select features. By using a histogram, one can show that the $\mathbf{s}$ statistics follow a gamma distribution with shape γ, scale β, and location μ parameters, i.e., gamma(μ, γ, β). As such, PHet fits a gamma distribution to estimate these parameters after scaling feature 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 reduced set of features, i.e., $\mathcal{P}' \subseteq \{1, .., p\}$.

Clustering and Dimensionality reduction

In this step, PHet uses the identified features ($\mathcal{P}'$) to detect subtypes in sample groups using a clustering algorithm, such as spectral clustering[98]. Also, 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 (Uniform Manifold Approximation and Projection)[51] which is currently considered to be the superior method to collapse high-dimensional features in omics data. Other alternative methods are PCA[99] and t-SNE[100]. The UMAP algorithm uses the stochastic gradient descent approach to minimize the difference between distances among samples in a higher-dimensional space and their projected lower-dimensional space. PHet also provides outputs in a tabular format that can be used in SPRING application[73] to explore cell differentiation trajectories.

Benchmark evaluation compared to existing tools

The performance of the PHet algorithm was evaluated in comparison with four DE feature analysis tools and their variants: the standard Student t-test[62], Wilcoxon rank-sum test[63], Kolmogorov–Smirnov test[64], and LIMMA[65]; a dispersion-based method and its variants[45]; an IQR-based approach and its variations[47]; and seven outlier detection algorithms: COPA[66], OS[38], ORT[39], MOST[40], LSOSS[41], DIDS[42], and DECO[43]. In addition, we also assessed the performance of PHet using the ΔDispersion metric instead of ΔIQR. 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 t-statistic, also known as the Student t-test, is a conventional statistical method that measures the difference in mean values 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 can be used to find DE features between the two groups, which can help to understand the biological processes that are associated with the disease or the treatment outcome. The two-class t-statistic for a feature j is defined as:

$$\text{statistic} = \frac{\overline{\mathbf{Y}}_{:,j} - \overline{\mathbf{X}}_{:,j}}{s_j} \tag{7}$$

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:

$$s_j^2 = \frac{\sum_{1\leq i\leq n}(\mathbf{X}_{i,j} - \overline{\mathbf{X}}_{:,j})^2 + \sum_{1\leq i\leq m}(\mathbf{Y}_{i,j} - \overline{\mathbf{Y}}_{:,j})^2}{n+m-2} \tag{8}$$

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[101] and 2)- fit a gamma distribution to statistics and then trim features at < 0.01 significance level. The first approach is termed *t-statistic* and the later method is called *t-statistic+Gamma*. For both approaches, we applied a two-sided test using the SciPy package[102].

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:

$$\begin{aligned} \text{statistic} &= \frac{W_j - \mu}{\sigma} \\ W_j &= \sum_{i\in n} \text{rank}(\mathbf{X}_{i,j}) \\ \mu &= \frac{n(n+m+1)}{2} \\ \sigma &= \sqrt{\frac{nm(n+m+1)}{12}} \end{aligned} \tag{9}$$

For a large sample size and two-sided test, please see[63]. Having obtained p-values for all features, the important features can be determined based on FDR at < 0.01 significance level using the Benjamini & Hochberg method[101]. Alternatively, we can fit a gamma distribution to statistics and then select features at < 0.01 significance level. These two approaches are referred to as *Wilcoxon* and *Wilcoxon+Gamma*, respectively. For both approaches, we applied a two-sided test using the SciPy package[102]. It is worth mentioning that the Wilcoxon rank-sum test is robust to outliers and does not assume any specific distribution of the data; however, this test is less effective when the sample size per condition is smaller than 8.

Kolmogorov–Smirnov (KS) test. This is a nonparametric two-sample test that is used to compare the distributions of two groups of samples, such as control and case, without making any assumptions about their shapes or parameters[64]. The test statistic for the KS test, for a given feature j, is the maximum absolute difference between the empirical cumulative distributions of control and case samples[97]:

$$\text{statistic} = \sup_j |F(\mathbf{X}_{:,j}) - G(\mathbf{Y}_{:,j})| \tag{10}$$

where $F(.)$ and $G(.)$ are the empirical distribution functions of control and case samples, respectively, and sup is the supremum function. To identify important features, we employed two approaches: *KS* and *KS+Gamma*. The KS method involves setting a p-value threshold based on false discovery rate (FDR) at a significance level of less than 0.01 using the Benjamini & Hochberg method[101]. In contrast, the KS+Gamma method involves fitting a gamma distribution to the test statistics and trimming features at a significance level of less than 0.01. The two-sided test was performed using the SciPy package[102] for these two strategies.

LIMMA. LIMMA (Linear Models for Microarray Data)[65] 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 package takes expression data as input, which can be log ratios or log-intensity values from microarray technologies. Then, LIMMA attempts to fit a linear model for each feature to the expression data. LIMMA then fits a linear model for each feature to the expression data and applies empirical Bayes and other shrinkage methods to improve the statistical inference. Similar to the t-statistic, features are

determined using two approaches: 1)- a p-value cutoff at < 0.01 significance level implemented by LIMMA, termed as *LIMMA* method, and 2)- fit a gamma distribution to statistics, obtained from LIMMA, and then trim features at < 0.01 significance level, which is named as *LIMMA+Gamma*.

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 affect the p-value estimation and the false discovery rate control.

Dispersion based HV Features. Satija and colleagues[45] developed a method based on the normalized dispersion analysis to identify HV features. The method computes dispersion measures (variance/mean) for each feature from all samples and then groups features, based on their mean expression, into a fixed number of bins (20 by default). Within each bin, the dispersion measures of features are standardized to identify features that have higher variability than the average expression values in the bin. Typically, selecting the top HV features requires the configuration of various hyperparameters. In our methodology, we adopt an alternative approach. Rather than specifying hyperparameters, we collect the normalized dispersion values for all features and fit a gamma distribution to these values. We then select features at a significance level of less than 0.01. This approach allows us to determine HV features without the need for manual parameter tuning. We apply this method to a combination of control and case samples, which is why we refer to it as *Dispersion (composite)*.

Furthermore, we extend this process by introducing three additional strategies to collect statistics or dispersion values. The first strategy, *Dispersion (per condition)*, computes dispersion values for features separately for each condition. The second strategy, Δ*Dispersion*, calculates the absolute differences of dispersion values between samples of control and case groups for all features. The third strategy, Δ*Dispersion*+Δ*Mean*, takes the feature-wise sum of the absolute differences of both dispersion and mean values between samples of the two conditions. For all these methodologies, we fit a gamma distribution to the collected values and select features at a significance level of less than 0.01.

InterQuartile Range (IQR). This measurement allows to detection of a set of features that are conserved across different samples. We propose four methods to calculate the IQR for each feature: 1)- *IQR (composite)* which involves combining the control and case samples and computing the IQR of their composite distribution. This approach takes into account the overall variability in the data; 2)- *IQR (per condition)* which calculates the IQRs of samples for each condition separately. By considering the IQRs within each condition, this method provides insights into the variability specific to each condition; 3)- Δ*IQR*, which finds the absolute differences of the IQRs between samples of the two conditions. This method highlights the features that exhibit the greatest variation between the control and case samples; and 4)- Δ*IQR*+Δ*Mean*, which combines the absolute differences in IQRs with the absolute differences in means between samples of the two conditions. By considering both IQR and mean differences, this method captures both variability and shifts in the data. The formula for ΔIQR+ΔMean for a feature j is defined as:

$$\text{statistic} = |\text{IQR}(\mathbf{X}_{:,j}) - \text{IQR}(\mathbf{Y}_{:,j})| + |\overline{\mathbf{X}}_{:,j} - \overline{\mathbf{Y}}_{:,j}| \tag{11}$$

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 a unit variance and zero mean.

COPA. While the t-statistic is an effective way to determine the DE features, this method does not account for the heterogeneity of samples within each group. This is because the t-statistic computes the mean difference of features between control and case samples without providing a way to discern outliers. Typically, outlier samples have high expression values for a set of features within a group. This phenomenon is common 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[37] proposed Cancer Outlier Profile Analysis (COPA) which is defined for a feature j as:

$$\text{statistic} = \frac{q_r(\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) - \text{MED}_j}{\text{MAD}_j} \tag{12}$$

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

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

MAD_j is the median absolute deviation of expression values in all samples and is defined as:

$$\text{MAD}_j = 1.4826 \times \text{median}(\{(\mathbf{X}_{i,j} - \text{MED}_j) : 1 \leq i \leq n\}, \{(\mathbf{Y}_{i,j} - \text{MED}_j) : 1 \leq i \leq m\}) \tag{14}$$

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.

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[38] 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:

$$\text{statistic} = \frac{\sum_{\mathbf{Y}_{i,j} \in R_j} (\mathbf{Y}_{i,j} - \text{MED}_j)}{\text{MAD}_j} \tag{15}$$

where,

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

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

$$\begin{aligned} \text{IQR}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) = &q_{75}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \\ &- q_{25}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}, \{\mathbf{Y}_{i,j} : 1 \leq i \leq m\}) \end{aligned} \tag{17}$$

ORT. Wu and colleagues[39] proposed another variant to the OS statistic where the median of expression values for a feature j is computed per group based as:

$$\text{statistic} = \frac{\sum_{\mathbf{Y}_{i,j} \in R_j} (\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{X})})}{\text{MAD}_j} \tag{18}$$

The definition of R_j is:

$$R_j = \{\mathbf{Y}_{i,j} : \mathbf{Y}_{i,j} > q_{75}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\}) + \text{IQR}(\{\mathbf{X}_{i,j} : 1 \leq i \leq n\})\} \tag{19}$$

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

$$\text{MAD}_j = 1.4826 \times \text{median}(\{|\mathbf{X}_{i,j} - \text{MED}_j^{(\mathbf{X})}| : 1 \leq i \leq n\}, \{|\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{Y})}| : 1 \leq i \leq m\}) \tag{20}$$

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 separately for each condition, rather than using all the samples to get a global median value. This method, called Outlier Robust T-statistic (ORT), showed good performance compared to OS and COPA on a breast cancer microarray dataset[103].

MOST. Following the ORT statistic, Lian and colleagues[40] 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:

$$\text{statistic} = \max_{1 \leq k \leq m} \left[\frac{\sum_{1 \leq i \leq k} (\mathbf{Y}_{i,j} - \text{MED}_j^{(\mathbf{X})})}{\text{MAD}_j} - \mu_k \right] / \delta_k \tag{21}$$

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 represents the optimum MOST statistic for the feature j. The median absolute deviation of expressions(MAD_j) has the exact formula as in the ORT statistic (Eq. 20).

LSOSS. Wang and colleagues[41] proposed an interesting approach to identifying outliers in case samples. Their research revealed that these outliers exhibit a distinct mode or peak of expression values compared to the rest of the inactivated case samples. This observation led them to introduce a novel statistical method, the Least Sum of Ordered Subset Square (LSOSS) method. The primary objective of this method is to identify the specific peak in expression values that corresponds to the outliers in case samples. By doing so, the LSOSS method endeavors to maximize the disparity 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. $\overline{\mathbf{Y}}_{:,j}^{(S)}$ 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 corresponds 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 for the feature j.

DIDS. de Ronde and colleagues[42] developed the Detection of Imbalanced Differential Signal (DIDS) algorithm that calculates the absolute differences in expression values by comparing the case samples with the maximum expression value observed in the control samples. These differences are then aggregated using one of three statistical 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:

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

DECO. Campos-Laborie and colleagues[43] developed a method to address feature profile changes in control and case samples. This approach is called DEcomposing heterogeneous Cohorts using Omic data profiling (DECO), and it consists of six steps:

1. Recurrent Differential Analysis (RDA) step to identify, select, and rank all significant changes found for any feature between two subsets of control and case samples. 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 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. Classify each feature into four model-type scheme. The categories include complete, majority, minority, and mixed. The classification is achieved by calculating the overlap of signal between samples in each condition. A feature is considered complete when it shows a significant change in all sample cases compared to controls. A feature is classified as majority when it displays a change in the majority of cases. A minority feature only exhibits changes in a minority of cases. Lastly, a feature is considered mixed when it displays significant changes in both cases and controls.

6. Rank all features using results from RDA, h-statistic, and feature classification steps.

DECO was shown to deliver promising results for the analysis of heterogeneity, biomarker identification, and subtype discovery[43]. Moreover, DECO can detect subpopulations without requiring labeled samples in unsupervised approach.

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. For the Student t-test, Wilcoxon rank-sum test, Kolmogorov–Smirnov test, and LIMMA, we used a two-tailed test with default settings. For COPA and OS, we assigned q_r to 75%. We used the 'tanh' scoring function for DIDS and set the IQR range to $[25, 75]$ for OS, ORT, the four variants of IQR, and PHet. Additionally, we set the vector of feature weight w for discriminatory power to $[0.4, 0.3, 0.2, 0.1]$ for PHet. To ensure a comprehensive analysis, we applied up and down regulations for OS, ORT, MOST, LSOSS, DIDS, all variants of IQR, and PHet. For LIMMA and Dispersion-based methods, we used the LIMMA[65] and the SCANPY[44] packages with default settings, respectively. To collect p-values and statistics from the Student t-test, Wilcoxon rank-sum test, and Kolmogorov–Smirnov test, we utilized the SciPy package[102]. Finally, for DECO analysis, we employed the deco package written in R and applied the `decoRDA` module with 1000 iterations while keeping the remaining parameters at their default configurations.

Evaluation metrics

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

F1. The F1 metric is a widely used measure of the effectiveness 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. Precision is defined as the ratio of true positive predictions to the total number of positive predictions, while recall is the ratio of true positive predictions to the total number of actual positive instances. The F1 metric is defined as:

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

where tp, fp, and fn correspond to the true positive, false positive, 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 on the microarray, single-cell RNA-seq, simulated, and HER2 datasets in Fig. 2(a)-(d), Fig. 4(a)-(b), and (e), Supplementary Fig. 1(a)-(b), and (e), and Supplementary Figs 2, 14, and 30-31.

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

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

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

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

where a is the number of pairs that are assigned to the same cluster in both the current and benchmark true subtypes and b is the number of pairs that are assigned to different clusters in the current and benchmark true subtypes. The denominator of the RI formula $\binom{n}{2}$ is the total number of pairs that can be formed from n samples. The ARI ranges in [0,1] with higher values indicating greater similarity between the current clustering results and either previous clustering results or pre-annotated labels. The ARI metric was used in Fig. 2(c) and (f), and Supplementary Figs 2, 14, and 30-31.

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. The AMI is calculated using the following formula:

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

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. 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 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 in the predicted cluster k and x_c is the number of samples that have the true subtype c. The AMI ranges in [0,1] where a higher score entails that both true labeled subtypes and clustering labels are near identical. We used the AMI metric to compare the clustering performances of algorithms for microarray and single-cell RNA-seq datasets in Supplementary Figs 2, 14, and 30-31.

Silhouette. This metric aims to determine whether a clustering solution has successfully minimized the within-cluster dissimilarity while maximizing the inter-cluster dissimilarity. The formula for calculating the silhouette score is as follows:

$$\text{Silhouette} = \frac{b-a}{\max(a,b)} \tag{35}$$

where a is the average intra-cluster distance, i.e., the average distance between each sample within a cluster and b is the average inter-cluster distance, i.e., the average distance between all clusters. The silhouette score ranges from -1 to 1, where a high silhouette value indicates that samples are close to their own cluster and far from other clusters. We used this metric to compare the clustering performances of PHet, pre-annotated markers, and HV features for MTECs single-cell RNA-seq data in Fig. 6(i).

Homogeneity. This metric measures the distribution of subtypes within each predicted cluster:

$$\text{Homogeneity} = 1 - \frac{\mathcal{H}(t|\hat{t})}{\mathcal{H}(t)} \tag{36}$$

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 quantifies how much information is required to determine the true labels of a sample based on its predicted cluster assignment:

$$\mathcal{H}(t|\hat{t}) = -\sum_{c \in |t|} \sum_{k \in |\hat{t}|} \frac{x_{c,k}}{n+m} \log(\frac{x_{c,k}}{x_k}) \tag{37}$$

where $x_{c,k}$ corresponds to the number of samples in cluster k that belongs to the true label c, x_k represents the total number of samples in the predicted cluster k, and x_c is the number of samples that have the true subtype c. The homogeneity score ranges in $[0, 1]$, where 1 denotes that each cluster contains only samples from a single true subtype (i.e., zero entropy), and 0 entails that each cluster contains samples 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, 14, and 30-31.

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

$$\text{Completeness} = 1 - \frac{\mathcal{H}(\hat{t}|t)}{\mathcal{H}(\hat{t})} \tag{38}$$

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:

$$\mathcal{H}(\hat{t}|t) = -\sum_{k \in |\hat{t}|} \sum_{c \in |t|} \frac{x_{c,k}}{n+m} \log(\frac{x_{c,k}}{x_c}) \tag{39}$$

where $x_{c,k}$ corresponds to the number of samples in cluster k that belongs to the true label c, x_k represents the total number of samples in the predicted cluster k, and x_c is the number number of samples 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, 14, and 30-31.
V-measure. This metric is the harmonic mean of two scores, homogeneity and completeness:

$$\text{V-measure} = (1 + \beta) \times \frac{\text{Homogeneity} \times \text{Completeness}}{\beta \times \text{Homogeneity} + \text{Completeness}} \tag{40}$$

where Homogeneity and Completeness are defined in Eq. 36 and 38, 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, 14, and 30-31.

Clustering

We applied spectral and K-means clustering algorithms to two types of data: microarray gene expression and single-cell RNA-seq (Supplementary Table 5). We performed clustering using the Scikit-learn package[104] with the number of clusters equal to the number of original patient samples and cell types in both types of data and the other hyperparameters at their default settings. We reported the quality of clustering results using various metrics, such as ARI, AMI, silhouette, homogeneity, completeness, and V-measure (Section 3).

UMAP settings

We utilized the UMAP package[93] 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 fixed to their default settings. We applied UMAP to generate several figures, such as Figs 1(f)-(i), 2(g)-(l), 3, 5(a)-(b) and (h)-(i), 6(a)-(h), and Supplementary Figs 3-13, 15-20, 21(a)-(b), 22(a)-(b), 24-26, 27(a), 28(a), and 29(a).

SCANPY settings

We used SCANPY[44] to normalize cells, select HV features based on the minimum dispersion criterion of 0.5, perform clustering using the Leiden algorithm at resolution ranges in $[0.4, 0.8]$, apply differential expression analysis between cell clusters using Wilcoxon rank-sum test, and visualize results using various plots, such as heatmaps. We employed SCANPY to obtain the results for Dispersion (composite), Dispersion (per condition), ΔDispersion, ΔDispersion+ΔMean, and PHet (ΔDispersion) methods in Sections 2.3, 2.4, 2.5, and 2.7. Additionally, we created several figures using this library in Fig. 6(a)-(h) and (m)-(p) and Supplementary Figs 22(g)-(j), 24-26, 27(a)-(b) and (d)-(f), 28(a)-(b) and (d)-(f), and 29(a)-(b) and (d)-(f).

SPRING settings

We utilized SPRING[73] to uncover complex high-dimensional structures within single-cell gene expression data. Specifically, we set the number of PCA dimensions to 50, the minimum number of cells for each gene to 3, the gene variability percentile to 0.0, the number of nearest neighbors to 5, and the minimum number of genes for each cell to 0.0. We applied SPRING to generate several figures, such as Fig. 5(c)-(d), (g), and (j)-(l), and Supplementary Figs 21(c)-(e) and (h), 22(c)-(e), and 23(a)-(b) and (e).

Differential feature expression

For Fig. 4(a)-(b) and Supplementary Fig. 1(a)-(b), we performed differential expression analysis using LIMMA[65], and the Benjamini-Hochberg[101] 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 (Supplementary Table 1), ii)- 1000 subsets of the HER2 breast cancer gene expression data with varying sample sizes and proportions (Supplementary Table 1), iii)- 11 gene expression datasets from microarray platforms with known clinical labels (Supplementary Table 2), iv)- six single-cell RNA-seq datasets with annotated cell identities (Supplementary

Table 3), and v)- two single-cell RNA-seq datasets from lung airway epithelial cells with mixed known and unknown cell subtypes (Supplementary Table 4).

Simulated datasets. A base synthetic data was generated using the 'madsim' package[105] with specific configurations. The lower and upper bounds of $\log_2$ intensity values were set to 2 and 16, respectively. The proportion of DE features was set to 0.1. The remaining parameters in the madsim package were set to their default values. The resulting dataset consisted of 1100 features, with 93 of them being DE features. There were a total of 40 samples, evenly distributed between control and case conditions. The purpose of creating this base dataset was to establish a prior distribution for simulating two additional sets of data. Each of these sets consisted of five datasets drawn from the same underlying base synthetic data. The first set represented the minority model type, where random perturbations were applied to 100 features in the case condition, including some DE features. The number of perturbed case samples ranged from 5% to 45%, corresponding to 1 to 9 out of the 20 samples. The second set represented the mixed model scheme, which followed a similar procedure as the minority model. However, in this case, the perturbed samples were distributed equally among a subset of cases and controls. These two sets of simulated data were designed to represent balanced situations, where the number of case and control samples was exactly the same. We used these datasets to evaluate the performance of 25 algorithms in Table 1 as discussed in Section 2.4.

HER2 datasets. We followed the approach described by de Ronde and colleagues[42] to produce 1000 synthetic batches of HER2 (human epidermal growth factor receptor 2) data. Each batch consisted of 27,506 features and 188 samples. To create these batches, we incorporated 178 fixed control samples from a breast cancer dataset (GEO accession ID GSE34138) representing the HER2-negative group, and 10 case samples randomly selected from 60 HER2-positive tumor samples (GEO accession ID GSE41656). It's important to note that the case samples varied for each batch, while the control samples remained constant. The main objective behind generating this dataset was to assess the effectiveness of algorithms in predicting 20 true outlier genes. These genes are specifically located on either ch17q12 or ch17q21, as discussed in Section 2.5. These regions were chosen due to their well-established association with HER2 amplification and breast cancer.

Microarray datasets. We downloaded seven microarray expression datasets from Gene Expression Omnibus (GEO) databases with accession numbers of GSE412; GSE3726; GSE89; GSE68956 (the Braintumor data); GSE2685; GSE83227 (the Lung data); GSE1987. 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 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 from airway epithelium. We downloaded two scRNA-seq datasets corresponding the airway epithelial cells from Gene Expression Omnibus (GEO) databases with accession numbers of GSE102580 (the HBECs and MTECs datasets).

Data availability

All datasets analyzed in this manuscript are publicly available and were obtained from public data repositories. See Section 3 and Supplementary Tables 1-4 for detailed information on microarray and single-cell omics datasets used in this study. The manuscript includes the source data necessary for generating key figures.

Code availability

The PHet framework in implemented in Python, which is freely available at `https://github.com/kleelab-bch/phet`. PHet is licensed under the MIT License.

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 Numbers: R35GM133725 and R01HL163513).

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).

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	✓	×	×	×	[65]
LIMMA+Gamma	Feature-wise linear models, empirical Bayes to obtain posterior variance estimators, and the gamma distribution	✓	×	×	×	[65] and this work
COPA	Percentile MAD	×	×	×	×	[37]
OS	Quantile ordered & cut-off	✓	×	×	×	[38]
ORT	Robust t-statistic	×	×	×	×	[39]
MOST	Maximum ordered	✓	×	×	×	[40]
LSOSS	Least sum of ordered subset square t-statistic	✓	×	×	×	[41]
DIDS	Maximum value from control group	✓	✓	×	×	[42]
DECO	RDA and NSCA	✓	✓	✓	✓	[43]
Dispersion (composite)	Normalized dispersion of all samples	–	–	✓	–	[45]
Dispersion (per condition)	Normalized dispersion of samples based on each condition	–	–	✓	✓	[45] and this work
ΔDispersion	Differences of normalized dispersion between samples of each condition	–	–	✓	✓	[45] and this work
ΔDispersion+ΔMean	Differences of normalized dispersion and mean between samples of each condition	–	–	✓	✓	[45] and this work
IQR (composite)	Z-score and IQR of all samples	–	–	✓	–	[47] and this work
IQR (per condition)	Z-score and IQR of samples based on each condition	–	–	✓	✓	[47] and this work
ΔIQR	Z-score and ΔIQR	–	–	✓	✓	[47] and this work
ΔIQR+ΔMean	Z-score, ΔIQR, and ΔMean	–	–	✓	✓	[47] and this work
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.

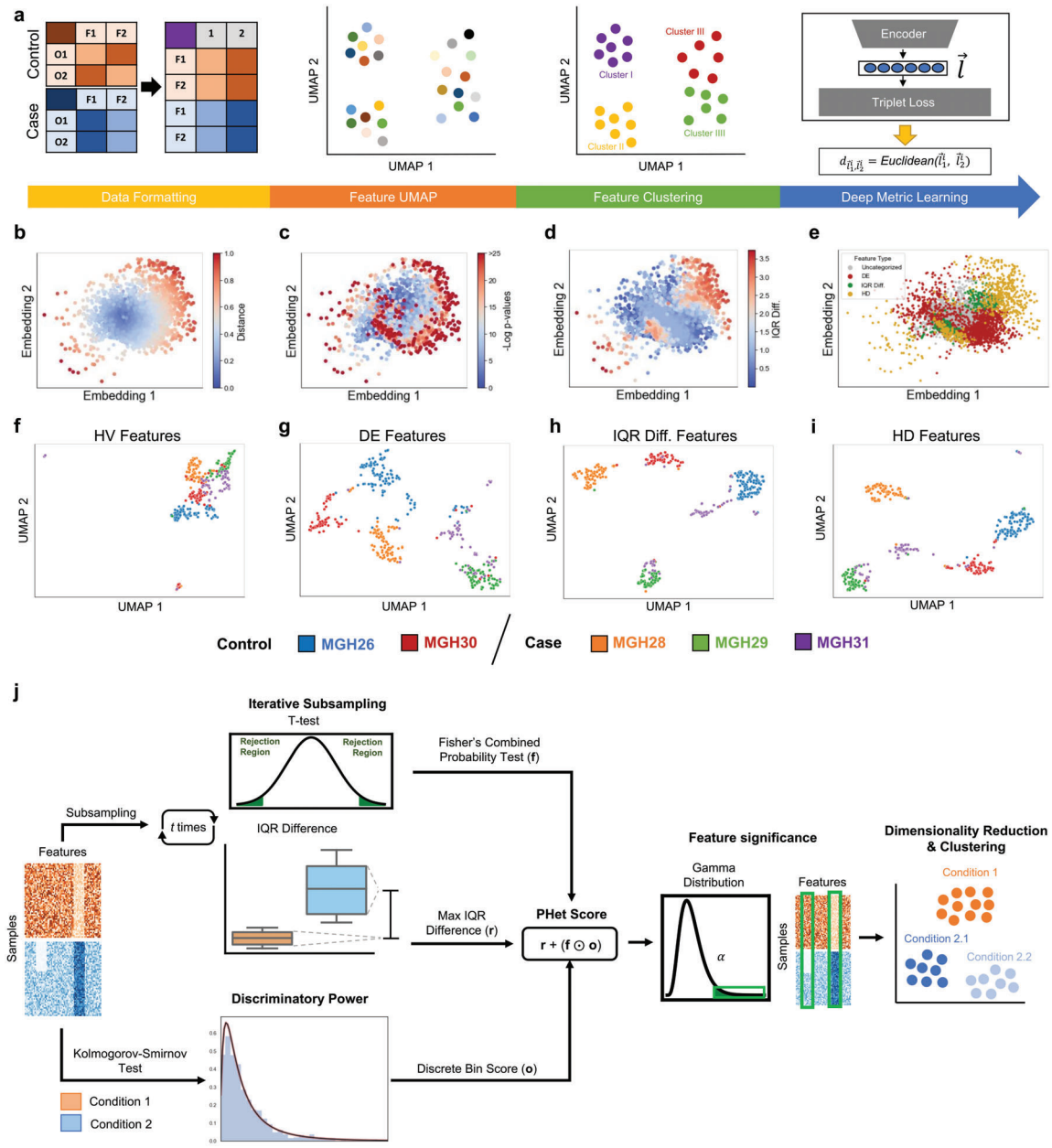

Figure 1: **Characterization of features extracted by deep-metric learning (DML) and the overview of PHet framework. a**, A schematic of DML approach with triplet loss, which is used to analyze the feature space. In this method, a UMAP is generated, where each point represents a feature from a specific condition. Clustering is then applied to the UMAP space. The resulting clusters and the original data are used as input for the triplet loss in DML. The embeddings from the encoder are used to calculate the Euclidean distance between the same feature of different classes, where $\vec{l}_1^{\,i}$ and $\vec{l}_2^{\,i}$ are the embeddings for the gene i in case and control conditions, respectively, and $d_{\vec{l}_1^{\,i},\vec{l}_2^{\,i}}$ is the Euclidean distance between these embeddings. **b**,**c**,**d**,**e**, Scatter plots of feature embeddings with color representing the distance between the same feature of different conditions (**b**), logged p-values from mean differences (using z-test) between conditions of features (**c**), IQR difference between conditions (**d**), and location of respective feature types on the plot (**e**). **f**,**g**,**h**,**i**, UMAP visualizations of 1529 HV features (**f**) (based on a dispersion threshold of > 0.5), 1936 DE features (**g**) (based on the z-test at significance level of 0.01), 904 ΔIQR features (**h**) (based on a threshold of > 0.4), and 641 HD features (**i**) that are intersection of both DE and ΔIQR features. **j**, PHet pipeline is composed of six major steps: 1)- an iterative subsampling to calculate p-values, using t-test or z-test, and maximum 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)- the Kolmogorov-Smirnov test to adjust the Fisher's scores. 4)- feature statistics estimation using a combination of the ΔIQR values, the Fisher's combined probability scores, and the weighted features representing their discriminatory power, 5)- fitting feature statistics using the gamma distribution, and features exceeding a user threshold are trimmed (< 0.01), and 6)- dimensionality reduction and clustering analysis on the reduced omics data to reconstruct data heterogeneity.

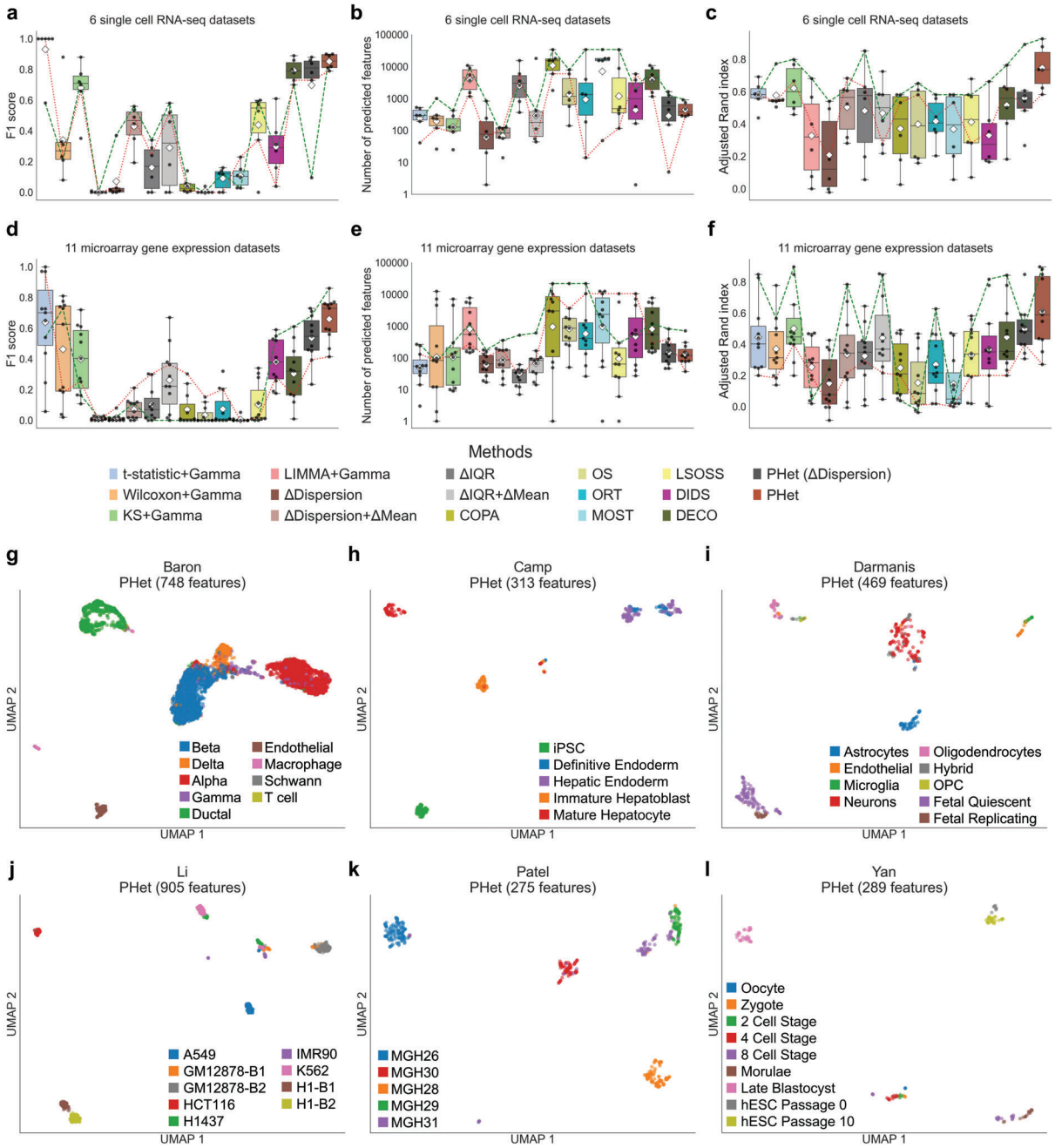

Figure 2: **Performance comparison of PHet against 16 baseline methods across 6 single-cell transcriptomics and 11 microarray datasets.** **a**,**d**, F1 scores of each method for detecting the top 100 DE features that are obtained using LIMMA. **b**,**e**, The number of selected features by each method. **c**,**f**, The adjusted Rand index of each method for 11 microarray and six single-cell transcriptomics datasets (Supplementary Tables 2 and 3). 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, while a red dashed line represents the worst-performing result on a dataset of PHet on each metric. **g**-**l**, UMAP visualizations of six single-cell transcriptomic datasets using PHet's selected features, colored by predefined original cell types (Supplementary Table 3).

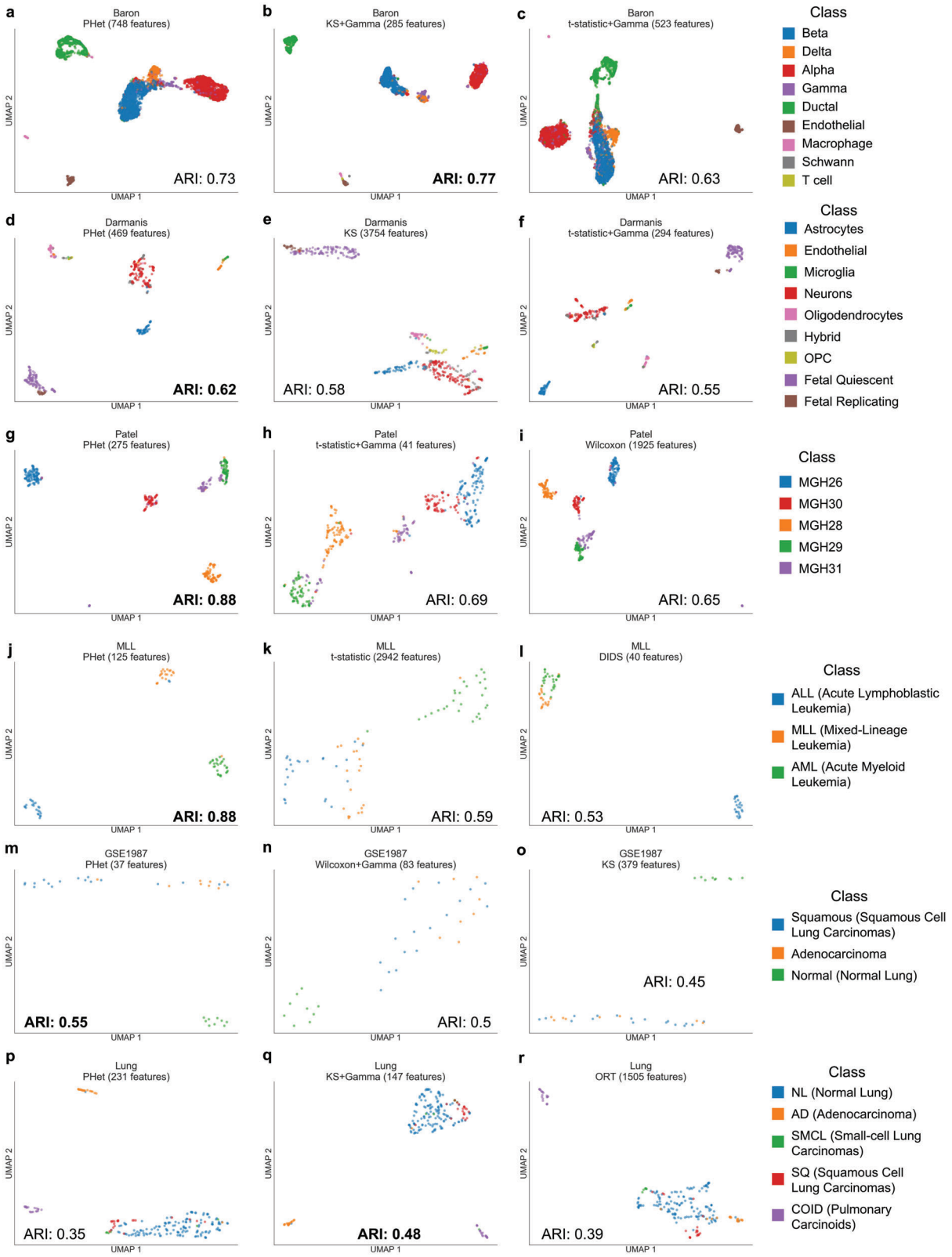

Figure 3: **UMAP visualizations of PHet and the top two existing methods. UMAP plots of Baron (a)-(c), Darmanis (d)-(f), Patel (g)-(i), MLL (j)-(l), GSE1987 (m)-(o), and Lung (p)-(r) datasets are overlaid with predefined original cell/patient types (Supplementary Tables 2 and 3).** For each method, the selected features are used. The bold font ARI score indicates the best performing method for the corresponding data.

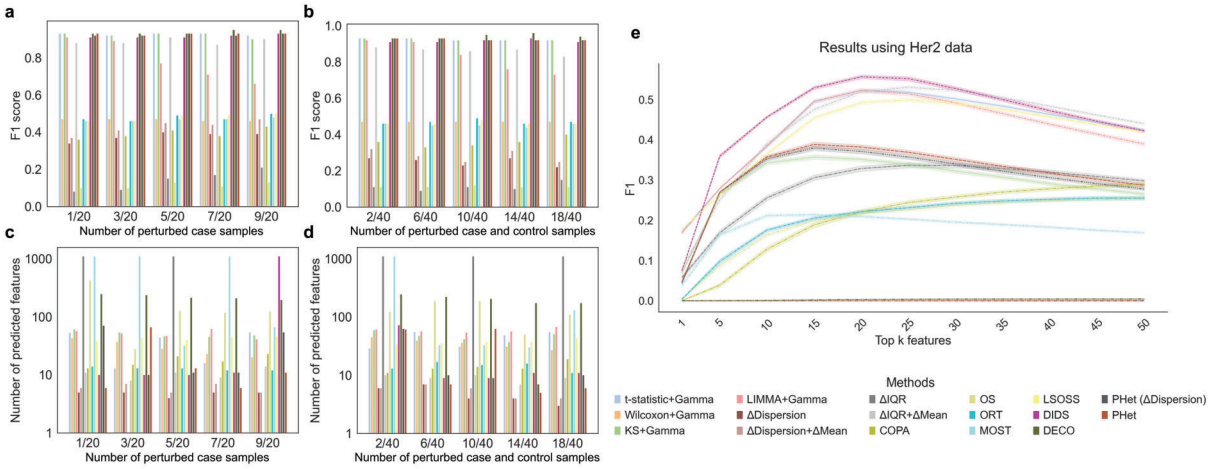

Figure 4: **Performance comparison of PHet against 16 baseline methods using simulated and HER2 datasets. a**, **b**, F1 scores of 17 different methods to detect the true DE features in perturbed samples. **c**, **d**, The number of selected features for each method. **a**, **c**, The performance of the 17 methods is compared under five scenarios, where the number of perturbed case samples varies from 1 to 9 ($\{1, 3, 5, 7, 9\}$). **b**, **d** The performance of the 17 methods is compared under five scenarios, which correspond to an increasing number of perturbed 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**, The performance of the 17 algorithms, using the F1 metric, in identifying the true 20 biomarkers on HER2 data (Supplementary Table 1). The algorithms were assessed based on their top k predicted features, ranging from 1 to 50.

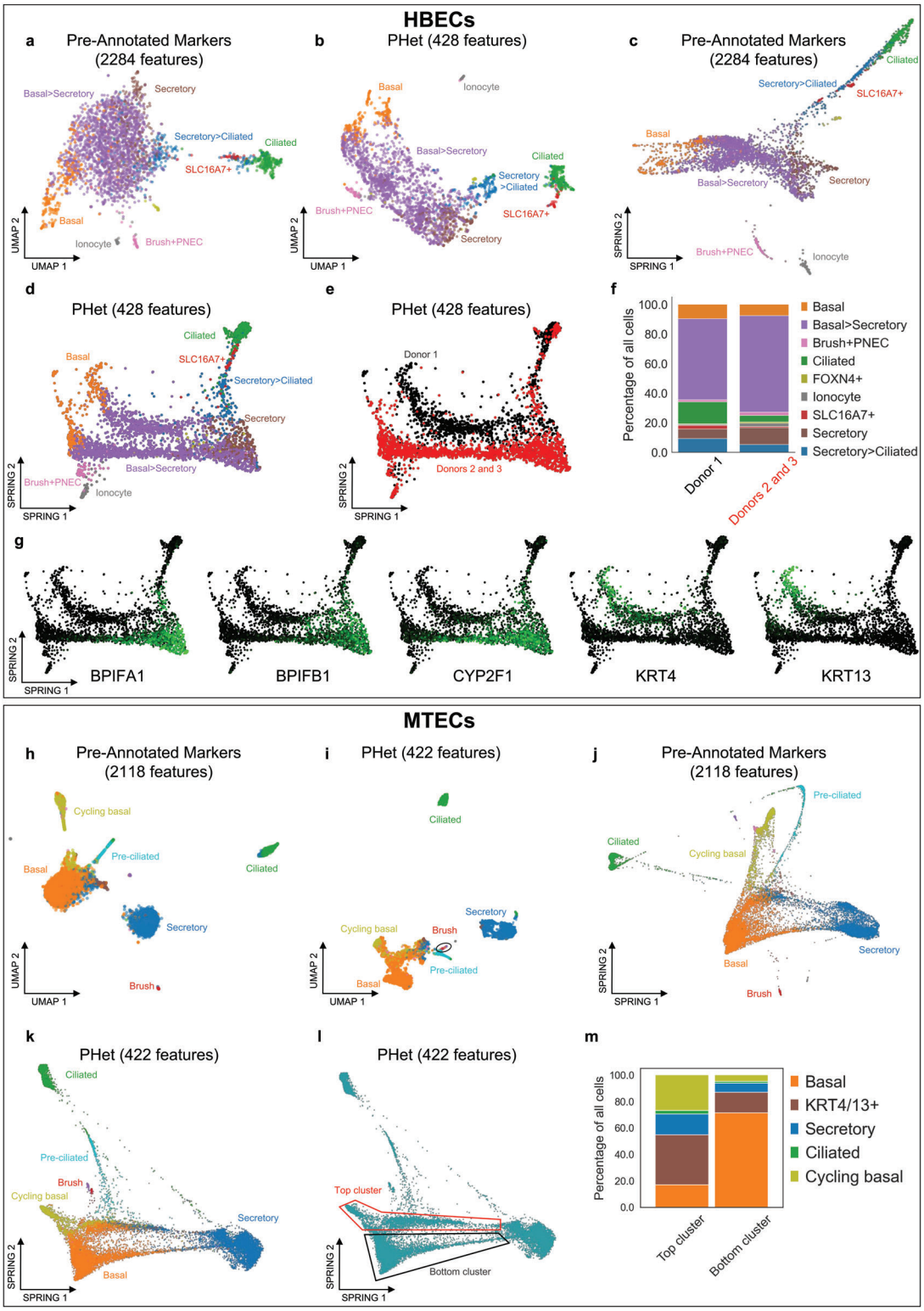

Figure 5: **Differentiation trajectory reconstruction of HBECs and MTECs single-cell RNA datasets.** **a**,**b**,**c**,**d**, UMAP visualizations using pre-annotated markers and PHet's selected features and their corresponding SPRING plots, respectively. Coloring represents the previously annotated cell types. **e**, A SPRING plot using PHet's features displaying two distinct trajectories: Donor 1 (black) and Donors 2 and 3 (red). **f**, A bar plot representing the relative abundance of cell types grouped by donors. **g**, SPRING plots of the selected top features predicted by PHet. The color gradient from black to green indicated cells enriched with the corresponding feature. **h**,**i**,**j**,**k**, UMAP visualizations using pre-annotated markers and PHet's selected features and their corresponding SPRING plots, respectively. Coloring represents the previously annotated cell types. **l**, Two distinct clusters are displayed in the SPRING plot, with one cluster (injured) located at the top and another (uninjured) at the bottom. **m**, The relative abundance of cells between these clusters are shown as a bar plot.

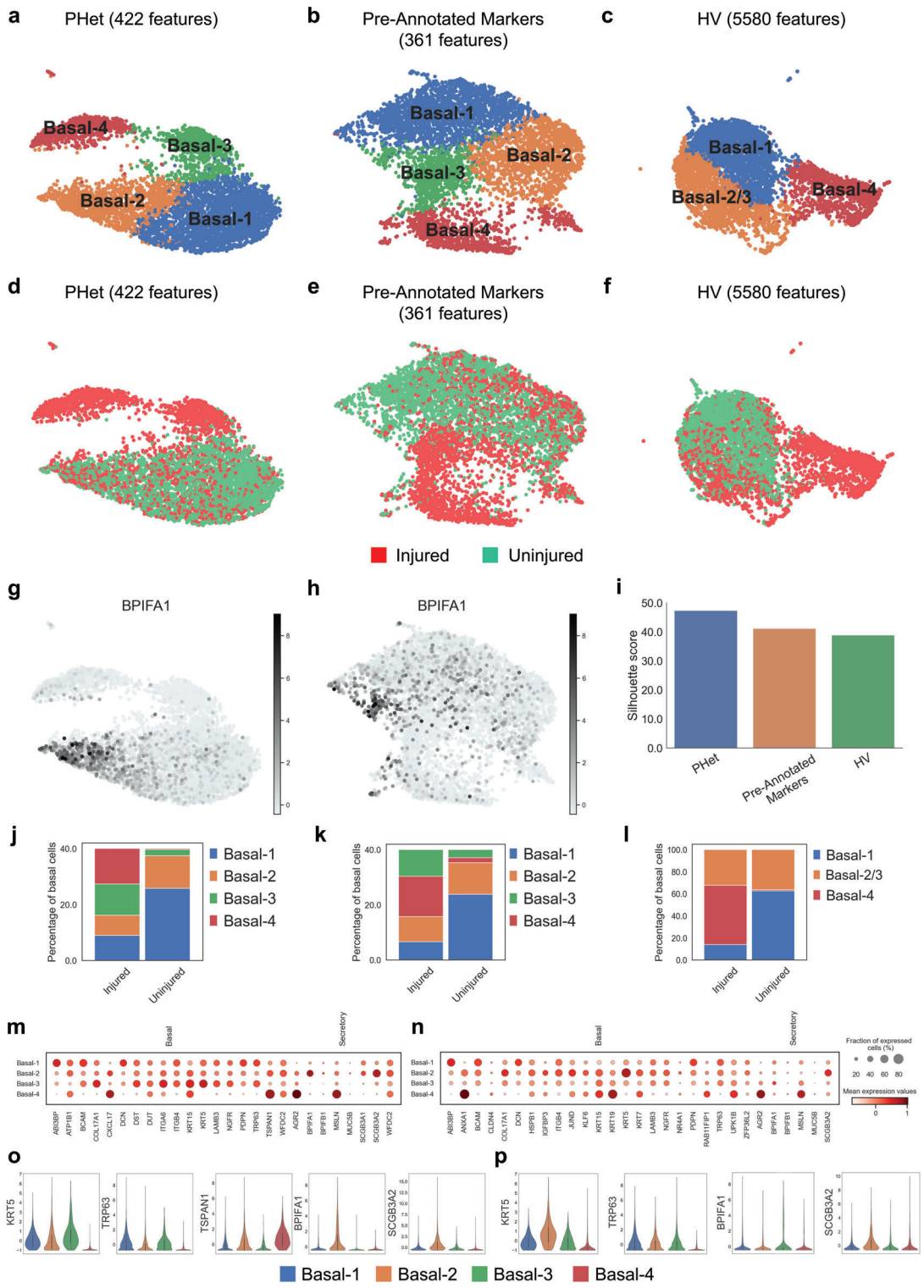

Figure 6: **Basal cell subtype discovery in MTECs single-cell RNA data.** **a**,**b**,**c**, UMAP visualizations using PHet's selected features (**a**), pre-annotated markers (**b**), and HV features (**c**). Coloring represents the detected basal cell types. **d**,**e**,**f**, UMAP visualizations of injured vs uninjured conditions using PHet's selected features (**d**), pre-annotated markers (**e**), and HV features (**f**). **g**, **h**, UMAP plots of basal cells for PHet (**g**) and pre-annotated markers (**h**). These plots provide insights into the expression patterns of the secretory cell marker "BRIFA1". **i**, Silhouette scores for clustering results based on PHet's features, pre-annotated markers, and HV features. **j**,**k**, **l**, The proportion of each basal subtype in the injured and uninjured cell groups given PHet (**j**), pre-annotated markers (**k**), and HV features (**l**). **m**, **n**, dot plots of the expression levels of selected basal and secretory cell markers among the four basal subtypes for PHet (**m**) and pre-annotated markers (**n**). **o**, **p**, Violin plots of the distribution of feature expressions of selected basal and secretory cell markers among the four basal cell subpopulations for PHet (**o**) and pre-annotated markers (**p**).