METHODOLOGY

BMC Methods





Discovery of optimal cell type classification marker genes from single cell RNA sequencing data

Angela Liu¹, Beverly Peng¹, Ajith V. Pankajam², Thu Elizabeth Duong³, Gloria Pryhuber⁴, Richard H. Scheuermann² and Yun Zhang^{1*}

Abstract

Background The use of single cell/nucleus RNA sequencing (scRNA-seq) technologies that quantitively describe cell transcriptional phenotypes is revolutionizing our understanding of cell biology, leading to new insights in cell type identification, disease mechanisms, and drug development. The tremendous growth in scRNA-seq data has posed new challenges in efficiently characterizing data-driven cell types and identifying quantifiable marker genes for cell type classification. The use of machine learning and explainable artificial intelligence has emerged as an effective approach to study large-scale scRNA-seq data.

Methods NS-Forest is a random forest machine learning-based algorithm that aims to provide a scalable data-driven solution to identify minimum combinations of necessary and sufficient marker genes that capture cell type identity with maximum classification accuracy. Here, we describe the latest version, NS-Forest version 4.0 and its companion Python package (https://github.com/JCVenterInstitute/NSForest), with several enhancements to select marker gene combinations that exhibit highly selective expression patterns among closely related cell types and more efficiently perform marker gene selection for large-scale scRNA-seq data atlases with millions of cells.

Results By modularizing the final decision tree step, NS-Forest v4.0 can be used to compare the performance of userdefined marker genes with the NS-Forest computationally-derived marker genes based on the decision tree classifiers. To quantify how well the identified markers exhibit the desired pattern of being exclusively expressed at high levels within their target cell types, we introduce the On-Target Fraction metric that ranges from 0 to 1, with a metric of 1 assigned to markers that are only expressed within their target cell types and not in cells of any other cell types. NS-Forest v4.0 outperforms previous versions in simulation studies and on its ability to identify markers with higher On-Target Fraction values for closely related cell types in real data, and outperforms other marker gene selection approaches for cell type classification with significantly higher F-beta scores when applied to datasets from three human organs—brain, kidney, and lung.

Discussion Finally, we discuss potential use cases of the NS-Forest marker genes, including for designing spatial transcriptomics gene panels and semantic representation of cell types in biomedical ontologies, for the broad user community.

Keywords Single cell RNA-seq, Marker gene, Machine learning, Random forest, Cell type classification

*Correspondence: Yun Zhang zhangy@jcvi.org Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Single-cell/single-nucleus RNA sequencing (scRNA-seq) methods have become an established approach for measuring cell transcriptional phenotypes and better understanding distinct cell types and their states based on gene expression patterns. Cell types can be defined as distinct cell phenotypes that include both canonical cell types and discrete cell states [1]. Efforts to define and categorize these cell types using advanced single cell technologies have been ongoing over the past decade, including the Human Cell Atlas (HCA) [2], the NIH Human Bio-Molecular Atlas Program (HuBMAP) [3], and the NIH BRAIN Initiative [4]. These efforts have led to consortium-scale datasets from multiple tissues/organs across the human body. For example, an early BRAIN Initiative study of the middle temporal gyrus (MTG) region in the human brain identified 75 distinct brain cell types with a dataset of approximately 16,000 nuclei [5]. A recent study on the transcriptomic diversity across the whole human brain revealed 461 cell type clusters and 3313 subclusters, with a final dataset comprised of more than three million cells [6]. The HuBMAP consortium covers other major human organs, including the kidney [7] and lung [8], resulting in a collection of 898 cell types with approximately 280 million cells across multi-omics assays [9].

While the number of cell types being identified in these scRNA-seq data atlases is increasing rapidly, there is a lack of a scalable and generalizable marker gene selection method that can systematically characterize these newly identified cell types for downstream use cases, such as for designing spatial transcriptomics gene panels and semantic representation of cell types in biomedical ontologies. Historically, two main approaches have been used to identify cell type-specific marker genes from scRNA-seq data: differential expression (DE) analysis and manual curation of gene lists using prior domain knowledge. For example, the anatomical structures (AS), cell types (CT), and biomarkers (B) ASCT + B tables [10] provided by the HuBMAP consortium use markers found in the scientific literature curated by domain experts for most of the organs in HuBMAP [11]. This approach is not only infeasible for large-scale datasets, but also leads to potentially incomplete (missing markers) or redundant (markers of parent cell type being used for child cell type) information for the most granular cell types. Alternatively, DE genes selected by modified Wilcoxon rank sum (performed using the "presto" R package) or related statistical tests are used in the popular Azimuth [12] web application for those cell types represented in the references. The DE approach selects genes based on the gene expression distributions and the adjusted *p*-values produced by a chosen testing method, which does not directly test the ability to classify cell types. Therefore, we formally introduce the notion of "cell type classification marker gene combinations" for scRNA-seq data, which must meet the following criteria: 1) each gene is expressed in the majority of cells of a given type, 2) each gene displays a "binary expression pattern" (i.e., highly expressed in the target cell type and little to no expression in other cell types), and 3) gene combinations are optimized for cell type classification using metrics that quantify classification confidence. By meeting these criteria, a generalizable method that can produce reproducible "cell type classification marker gene combinations" would emerge.

For the above-described challenge, we have proposed to use a machine learning approach to identify marker genes for cell type classification from scRNA-seq data and developed the NS-Forest method [13, 14]. NS-Forest uses the random forest machine learning algorithm to select informative gene features (or markers) that are optimized for cell type classification. Random forest is a machine learning classification model that is well-known for retaining high explainability, which is preferable for biomedical use cases.

NS-Forest was first introduced in 2018 as an algorithm that takes in scRNA-seq data and outputs the minimum combination of necessary and sufficient features that capture cell type identity and uniquely characterize a discrete cell phenotype [1]. In NS-Forest v1.3 [14] (the first publicly released version), the method first produces a list of top gene features (marker candidates) for each cell type ranked by Gini index calculated in the random forest model. (Fig. 1 summarizes major steps of the NS-Forest workflow compared across all versions.) The median gene expression value of each potential marker within the target cell type is calculated as the expression threshold to determine the number of true/false positives/negatives for each marker candidate in each cell type. Finally, the minimum set of markers for each cell type is determined by evaluating the unweighted F1-score following the stepwise addition of each of the ranked genes for each cell type.

NS-Forest v2.0 [13] was developed in 2021, and introduced the concept of the Binary Expression Score, a metric used to quantify how well a marker gene exhibits a "binary expression pattern" in which the marker gene is expressed at high levels in the majority of cells of the target cell type and not in cells of other cell types. Version 2.0 uses Binary Expression Score as a post random forest ranking step to preferentially select genes with the desired binary expression pattern, in addition to filtering out genes with negative expression levels, after the initial feature selection process from the random forest classifier. Instead of simply using each gene's median expression value within the target cluster to determine its expression threshold, version 2.0 builds a one-versus-all

Workflow step	NS-Forest v1.3	NS-Forest v2.0	NS-Forest v3.9	NS-Forest v4.0
Input data format	Cell-by-gene expression matrix (log-transformed/normalized) and cell cluster labels		Annotated data (anndata) object in h5ad format	
Gene pre-selection	None	Select positive genes (i.e., me least one cluster)	dian gene expression > 0 in at	a. Select positive genes b. Select by Binary Scoring Criterion ("BinaryFirst")
Random Forest feature	Rank gene features by Gini	a. Rank gene features by Gini Impurity		
selection	Impurity	b. Rank gene features by Binary Score		
Feature evaluation	Calculate median gene expression in the target cluster for each top-ranked gene feature	Use each of the top-ranked gene features to build one-vs-all decision trees for the target cluster to determine the optimal expression threshold		
Combinatorial marker gene determination	Calculate the F1 scores for stepwise addition of the top- ranked gene features and find the maximum F1 score	Calculate the F-beta scores based on the decision trees for all combinations of any length of the top-ranked gene features and determine the best marker gene combination with the maximum F-beta score		
Performance metrics	F1 score	F-beta score	a. F-beta scoreb. Positive Predictive Value (PPV)	a. F-beta scoreb. PPV (precision)c. Recalld. On-Target Fraction

Fig. 1 Major steps of the NS-Forest workflow compared across all versions

decision tree for each marker candidate to derive the optimal expression level for classification. Finally, the F-beta score is calculated for all possible combinations of the top-ranked, most binary-expressed marker candidates in order to identify the best combination of markers with the maximum F-beta score. The F-beta score is different from the F1 score in that it is the weighted harmonic mean of the precision and recall (instead of just the harmonic mean), with the beta parameter weight adjustment allowing for emphasis of either precision or recall. In version 2.0 and all following versions, beta is set to 0.5 by default to weight precision higher than recall, to control for excess false negative values introduced by the dropout technical artifact in scRNA-seq experiments.

NS-Forest v3.9 is algorithmically very similar to version 2.0, and mainly differs by the format of the data used as input to the algorithm. Instead of the simple cell-by-gene expression matrix, where each entry contains the log-transformed or normalized expression level of each gene in each cell along with the cluster labels, version 3.9 takes the annotated data (anndata) [15] object in the.h5ad file format as input. Version 3.9 also provides calculation of the Positive Predictive Value (PPV) metric (precision) for quantifying the classification performance of the algorithm in addition to the F-beta score, emphasizing the pragmatic importance of the predicted positives in many of the applications.

One of the observed lingering weaknesses in versions 2.0 and 3.9 was the lower performance of NS-Forest marker genes in distinguishing between closely-related cell types with similar transcriptional profiles. In the human brain middle temporal gyrus (MTG) dataset [5] that was used to develop previous versions of

NS-Forest, there exist several of these closely-related cell type groups, especially within the VIP, PVALB, and L4 neuronal cell subclasses (see Results section). Here, we describe NS-Forest v4.0, which adds an enhanced feature selection step to improve discrimination between similar cell types without sacrificing the overall classification performance. This new "BinaryFirst" step enriches for candidate genes that exhibit the binary expression pattern as a feature selection approach prior to the random forest classification step. The BinaryFirst strategy effectively reduces the complexity of the input feature set for the random forest classifier, decreasing the runtime and allowing for the preferential selection of informative binary markers during the iterative random forest process and thus resulting in a more distinct and concise set of marker genes.

Results

Informative gene selection prior to random forest in NS-Forest version 4.0

The most significant change to the workflow of NS-Forest in version 4.0 is the introduction of the BinaryFirst module that is implemented in the gene pre-selection step of the workflow (Fig. 1). The BinaryFirst strategy is designed to enrich for candidate genes that exhibit the desired gene expression pattern prior to the random forest feature ranking. This step pre-selects gene candidates that have a Binary Expression Score value that is greater than or equal to a dataset-specific threshold based on the distribution of the Binary Expression Scores of all genes in the dataset (Fig. 2). In version 4.0, four threshold options used in the BinaryFirst step were implemented: 'none', 'BinaryFirst_mild', 'BinaryFirst_moderate',



Fig. 2 NS-Forest version 4.0 workflow. The algorithm uses an anndata object in.h5ad format, containing the cell-by-gene expression matrix and cluster labels for each cell, as data input (step 1). The median gene expression for each gene in each cluster (i.e., a cluster-by-gene median matrix) is calculated and genes that have positive median expression in at least one cluster are pre-selected (not shown). The Binary Expression Score (see Methods for explaination of notations) is then calculated for each cluster-gene pair (step 2) producing a cluster-by-gene Binary Score matrix (note that a gene may have different Binary Score values in different clusters), and a dataset-specific threshold is calculated based on the Binary Score distribution and user-selected mild, moderate, or high criterion. This threshold value is used to select candidate genes for each cluster with a Binary Expression Score greater than or equal to the threshold (step 3). These candidate genes are passed to build binary classification models for each cluster using the random forest (RF) machine learning method. Features (genes) are extracted from the RF model and ranked by the Gini Impurity index, and the top RF features are then reranked by their pre-calculated Binary Scores (step 4). A short list of the top-ranked candidate genes that are not only ranked high in the RF classification models but also have high Binary Scores are passed for decision tree feature evaluation and determining the best marker gene combination. A single-split decision tree is built for each evaluated gene for determining the optimal expression threshold for classification. All combinations of any length of these genes are considered using 'AND' logic to combine the decision trees, and the best combination is determined by the highest F-beta score as an objective function for optimizing the overall classification performance (step 5). The F-beta score, Positive Predictive Value (PPV) (a.k.a. precision), recall, On-Target Fraction, as well as true/false positive/negative classification values are reported for each cluster, serving as metrics for evaluating the performance of the final maker gene combinations (step 6)

or 'BinaryFirst high' (see Methods). If the threshold is 'none', the algorithm is the same as version 3.9. The other thresholds are calculated based on the distribution of Binary Expression Scores from all genes to account for the dataset-specific gene expression variabilities arising from many factors, including the organ/tissue type, sample pre-processing, sequencing platform, etc. As the thresholding value increases, the number of selected genes decrease. Thus, the BinaryFirst strategy effectively reduces the feature space that the random forest classifier must search over in the subsequent workflow step of NS-Forest and serves as an informative dimensionality reduction step. A scRNA-seq dataset typically contains tens of thousands of genes, but the majority will not be useful as marker genes for any given cell type clusters. Random forest is an ensemble machine learning method that uses the bagging technique to train a large number of decision trees. The bagging technique is an iterative procedure of random selection of features and, therefore, is time-consuming for classifying cell types from scRNA-seq data [16]. Each of the decision trees in the forest is constructed of nodes, at which the data is split into groups that optimize class purity using randomly selected features. Because performing an exhaustive search of all possible combinations of features at each split in the decision trees is computationally intractable, random forest classifiers can produce sub-optimal collections of decisions trees due to the random selection of available features [17]. In previous versions of NS-Forest, the large number of genes in the input datasets reduced the likelihood that the optimal genes for classification would be adequately sampled. Our hypothesis was that by reducing the size of the set of gene candidates, the BinaryFirst strategy would be able to adequately sample all of the candidate input genes with a reasonable number of decision trees, thereby simplifying the task of the random forest classifier while simultaneously reducing the overall runtime of NS-Forest. In summary, NS-Forest v4.0 utilizes the BinaryFirst strategy to enhance the stability and classification performance of the random forest classifier by pre-selecting informative features from scRNA-seq data.

Improved marker gene selection on the human brain dataset

The performance of NS-Forest v4.0 was assessed on the same human middle temporal gyrus (MTG) brain dataset that was used to evaluate previous versions of NS-Forest [5]. To determine the best thresholding criterion, the mild, moderate, and high BinaryFirst configurations were compared for the version 4.0 algorithm. The evaluation is based on the On-Target Fraction metric (see Methods), which is specifically designed to quantify how much of the marker gene expression is restricted to the target cluster. Comparing these three configurations, the On-Target Fractions significantly increase as the stringency of BinaryFirst thresholding increases (Fig. 3A), which results in fewer candidate genes for random forest construction (Fig. 3B). This suggests that the gene preselection step helps the NS-Forest algorithm better select on-target marker genes while effectively reducing the dimensionality of the input gene space. Hereinafter, the 'BinaryFirst_high' configuration will be the default setting for NS-Forest v4.0, unless specified otherwise.

The performance of all versions of NS-Forest were compared on the same brain dataset. The overall performances of different versions are shown in Fig. 3C, where a noticeable decrease in off-target expression is observed across the three heatmaps from versions 1.3 to 2.0/3.9 to 4.0. In the most ideal scenario where there exist marker genes that are exclusively expressed for each cluster, marker gene expression would only be observed in a stair-step pattern along the diagonal axis in such expression heatmaps. By this standard, NS-Forest v4.0 shows the cleanest diagonal pattern in its heatmap. Improvement is also observed when comparing the performance metrics, as the median PPV and the median On-Target Fraction increased from v2.0/v3.9 to v4.0 across all 75 clusters (Fig. 3D). The improvement in PPV (precision) means that version 4.0 better identifies marker genes that, when used as features by decision tree classifiers, lead to improved performance at identifying cells that belong to each unique cell type by reducing the number of false positives in this classification task. The increase in On-Target Fraction further supports this claim that version 4.0 is able to identify marker genes that are exclusively expressed at high levels in their target clusters. This improvement in the PPV and On-Target Fraction is slightly offset by a small decrease in the median F-beta score between the two versions. The tradeoff in the F-beta score was previously discussed in Aevermann et al. [13] when comparing the off-the-shelf random forest marker candidates strictly ranked by feature importance (Gini Impurity) to the top candidates re-ranked by their Binary Expression Score values. Aevermann et al. demonstrated that the marker genes selected with significantly higher Binary Expression Scores are more useful for many downstream assays such as RT-PCR and spatial transcriptomics. A similar trend was observed in this current comparison of versions 2.0/3.9 and version 4.0: the Binary Expression Scores are higher for v4.0 than for v2.0/3.9, with an average of 0.971 compared to 0.936. This is consistent with the observations of a cleaner diagonal pattern in the heatmap and higher On-Target Fraction values for version 4.0.

In total, NS-Forest versions 2.0/3.9 and version 4.0 identified 168 and 167 total marker genes, and 154 and 147 unique markers, respectively, to optimally distinguish between the 75 cell type clusters in the human MTG dataset. 85 of the total markers (~51% of the markers identified by v2.0/3.9) were identified for the same cluster for both sets. The full list of NS-Forest v4.0 marker genes on the human MTG dataset are available in Supplementary Table 1.

Localized improvement in marker gene specificity for closely related cell types

One of the motivations for developing NS-Forest v4.0 was to improve the previously sub-optimal performance on specific subclades of closely related cell types in the MTG dataset that may be more difficult to distinguish compared to all the other cell types. The MTG study found that the inhibitory neuron types are highly diverse but mostly sparse (45 types and 4,297 nuclei), and the excitatory neuron types span multiple brain layers and are most similar to types in the same or adjacent layers (24 types and 10,708 nuclei) [5]. These highly similar yet distinct cell types are usually grouped as subclades in the hierarchical dendrogram (Supplementary Fig. 1A). The performance of version 4.0 on the VIP (vasoactive intestinal polypeptide-expressing inhibitory neurons), PVALB (parvalbumin-expressing inhibitory neurons) and L4 (layer 4 excitatory neurons) subclades was examined using the three different BinaryFirst thresholds to determine if NS-Forest v4.0 would produce higher On-Target Fractions. Visually, there appears to be a clear improvement in the VIP subclade, as the amount of off-target expression (represented by the number of yellow squares not on the diagonal axis in the highlighted VIP box) looks to be substantially fewer in the heatmap for the 'BinaryFirst_high' configuration (Supplementary Fig. 1B). The amount of on-target expression (represented by the amount of red and orange squares on the diagonal axis) appears to be greatest in this heatmap as well. While the pattern of the PVALB and L4 subclades is less obvious, the pattern of increased on-target expression and decreased off-target expression is still observable.



	NA	0.857	0.000	
Median On-Target Fraction	NA	0.280	0.399	
Fig. 3 NS-Forest performance evaluated on the human MTG dataset. A Boxplots displaying the distribution of On-Target Fraction values across the 75 clusters in the human MTG dataset from running NS-Forest v4.0 with mild, moderate, and high BinaryFirst configurations. Paired test results: mild vs. moderate <i>p</i> -value = 0.01, moderate vs. high <i>p</i> -value = 1.79e-04, mild vs. high <i>p</i> -value = 1.68e-06. B Boxplots displaying the distribution of the number of genes retained after the BinaryFirst thresholding step from (A). C Heatmaps of NS-Forest v1.3, v2.0/v3.9.				

and v4.0 marker genes for the 75 cell type clusters from the human MTG dataset. The colors correspond to the normalized median expression level (log2-transformed counts per million) for the marker gene (rows) in a given cell type cluster (columns), with high expression in red/yellow, and low expression in blue/white. The clusters are ordered according to the hierarchical dendrogram provided in the original study (Fig. 1c in [5]). **D** Comparison of the performance metrics of the corresponding versions of NS-Forest results shown in (C)

*Note that the unweighted F1 score is used in v1.3. The On-Target Fraction difference between v2.0/3.9 and v4.0 corresponds to the mild and high BinaryFirst threshold comparison in (A)

These trends are also observed in the median On-Target Fraction values, as this value is the highest for all three subclades using the 'BinaryFirst_high' configuration (Supplementary Fig. 1C).

The distribution of On-Target Fractions for each of the three subclades is consistent with these visual patterns. The boxplot showing the On-Target Fraction values for the 'BinaryFirst_high' threshold is clearly higher than that for the 'BinaryFirst_mild' and 'BinaryFirst_moderate' thresholds in all three subclades (Supplementary Fig. 2). The *p*-value for comparing the mild and high BinaryFirst thresholds for the VIP subclade is statistically significant (*p*-value=0.01), but the *p*-values for the L4 subclade and the PVALB subclade are not significant (*p*-value=0.37 and 0.07, respectively); this is likely due to the small sample size for the paired t-test (the L4 subclade has 5 distinct cell types and the PVALB subclade has 6, whereas the VIP subclade has 21).

In addition, an extremely stringent threshold of mean+3 standard deviations was evaluated to investigate the further impact of the BinaryFirst strategy on these three subclades and globally (Supplementary Fig. 3). Overall, a cleaner heatmap was observed (Supplementary Fig. 3A), where 159 marker genes (145 unique genes) were identified for the human MTG dataset. Less off-diagonal expression was also observed in the VIP subclade (Supplementary Fig. 3B). At this higher threshold, quantiative markers that could be useful in refining the cell type classification in the random forest model would be filtered out at the cost of lower median F-beta score and lower median PPV (Supplementary Fig. 3C), even though the median On-Target Fraction was higher as reflected in the cleaner heatmap. The median On-Target Fractions of the three subclades showed that the local performance in the L4 subclade was worse than the 'BinaryFirst high' threshold, and the performance in the PVALB and VIP subclades were further improved (Supplementary Fig. 3D). By design, the higher thresholds result in fewer number of genes with higher Binary Scores being inputted to the random forest step. However, there is not a linear trend between the On-Target Fractions and the number of input genes to the random forest step using different thresholds (Supplementary Fig. 3E), as the linear fitted lines had R² values less than 0.01 for 'BinaryFirst_moderate' and 'BinaryFirst_high' ('BinaryFirst_mild' does not apply as it takes in all the genes for the random forest step) and the linear fitted line was mainly driven by the two outliners for the mean +3standard deviations threshold. Based on this evaluation, the 'BinaryFirst_high' is set as the default threshold in the NS-Forest v4.0 algorithm.

Validation on additional datasets of human kidney and lung

Datasets from two other human organs—the human kidney dataset from the Kidney Precision Medicine Project (KPMP) [7] and the human lung dataset from the Lung Airways and Parenchymal Map (LAPMAP) [8], both contributing to the HuBMAP consortium – were used to validate the performance of NS-Forest v4.0. In the lung dataset, three annotation levels were evaluated: level 3 (L3), level 4 (L4), and level 5 (L5) subclasses. The kidney dataset has 75 distinct cell types and the lung dataset has 61 cell types at the L5 subclass. For both the kidney and lung datasets, the heatmaps show more specific expression along the main diagonal with version 4.0 (Fig. 4A), which complement the observed high On-Target Fraction values for these datasets (Fig. 4B). These human kidney and lung metric values are higher than the metrics for the brain dataset because the human brain is a more complex organ in terms of the diversity of related cell types. Version 2.0/3.9 already performs quite well at selecting marker genes for these two organs compared to the brain, and hence the improvement of version 4.0 is less obvious.

Comparing versions 2.0/3.9 and 4.0 (Fig. 4B), the median F-beta scores are very similar in both NS-Forest versions for each organ. In the kidney dataset, a slight improvement in PPV and a slight decrease in On-Target Fraction were observed going from version 2.0/3.9 to version 4.0. It is interesting to note that although the On-Target Fraction is slightly lower, the number of false positive classifications is much fewer in version 4.0 (dropped from an average of 312.5 cells in version 2.0/3.9 to 198 cells in version 4.0), confirming that the version 4.0 marker genes are selected for optimal classification. In the lung L5 subclass dataset, increases in both PPV and On-Target Fraction were observed with version 4.0.

NS-Forest v2.0/3.9 and v4.0 identified 157 and 169 total marker genes for the kidney dataset, and 144 and 151 unique markers, respectively, to optimally classify the 75 cell types (Supplementary Tables 2-3). 117 of the total markers (75% of the markers identified by v2.0/3.9) were identified for the same cluster for both sets. NS-Forest v2.0/3.9 and v4.0 identified 131 and 126 total marker genes, respectively, for the lung L5 dataset, and 125 and 121 unique markers, respectively, to optimally distinguish between the 61 cell types. 107 of the total markers (82% of the markers identified by v2.0/3.9) were identified for the same cluster for both sets. Similar results were obtained running NS-Forest on the L4 and L3 subclasses (49 and 44 types, respectively) of the same lung dataset (Supplementary Fig. 4 and Supplementary Tables 4-5), suggesting that it is easier to select marker genes at less granular levels. The L4 and L3 subclasses results (Supplementary Fig. 4B) showed that while the median F-beta and median PPV are very similar between the two versions, these metric values are slightly lower in v4.0 (although all differences are less than 0.03). However, there is a substantial gain in the median On-Target Fraction. As previously explained, the reason for the difference in F-beta is the trade-off of the major gain in Binary Score as intended by the algorithm's design. After a careful review of the clusters that showed lower PPV



Fig. 4 NS-Forest performance evaluated on other human organs (kidney and lung). **A** Heatmaps of NS-Forest v2.0/v3.9 and v4.0 marker genes for the 75 cell type clusters from human kidney dataset and 61 cell types from human lung L5 subclass dataset. The colors correspond to the normalized median expression level (log2-transformed counts per million) for the marker gene (rows) in a given cell type cluster (columns), with high expression in red/yellow, and low expression in blue/white. The clusters are ordered according to the hierarchical ordering in the dendrogram generated by the scanpy package (*scanpy.tl.dendrogram*) using default settings. **B** Comparison of the performance metrics corresponding to the NS-Forest results shown in (A)

in v4.0, it was identified that more positives were classified in those clusters in v4.0 than v2.0/3.9, and the gain of the increase in true positivies came at the cost of a larger increase in false positives for those clusters.

Improvement in runtime in version 4.0

In addition to the improvements in the classification performance, the overall runtime of NS-Forest v4.0 (by default, 'BinaryFirst_high' is used) is much lower than that of v2.0/3.9 in all three human organ datasets (Table 1), with the ratio of runtime (v4.0 to v2.0/3.9) ranging from 0.10 to 0.26 across the datasets. In all three datasets, the 'BinaryFirst_mild' configuration did not filter out any genes, indicating that more than half of the genes have a value of 0 for their Binary Expression Score and thus, a value of 0 for their median expression per cluster. We note that for all three datasets, there is a large decrease in runtime in v4.0 with the 'BinaryFirst_moderate' configuration, and a less substantial decrease going from the moderate to 'BinaryFirst_high' configuration. These differences in runtime correspond to the decreases in the average number of genes left per distinct cell type after the BinaryFirst step. When the 'BinaryFirst_high' threshold was used, 1–7% of the total original genes passed the BinaryFirst threshold in the three datasets. This indicates that the majority of genes in these datasets have low Binary Expression Scores, and that the distribution of Binary Expression Scores in these datasets is heavily right-skewed (Supplementary Fig. 5), which is generally true for all scRNA-seq data. Overall, the BinaryFirst step can dramatically reduce the number of candidate genes that are considered as potential markers as input to the random forest step and simultaneously provide improvement in important measures of classification performance.

Marker gene comparison for the human lung cell atlas

While the goal of NS-Forest is to define the minimum set of necessary and sufficient marker genes to classify cell types from scRNA-seq data, other popular marker gene approaches aim to define marker genes by identifying genes that are differentially expressed between cell types (e.g., Azimuth [12]), or by manually curating knowledge historically reported in the scientific literature (e.g., ASCT + B from the HuBMAP consortium [10]). Although Azimuth and ASCT + B provide pan-organ marker

Dataset	BinaryFirst configuration	Time to run (hrs:min:secs)	Ratio to run time of v2.0/3.9	Average fraction of genes left per cluster after BinaryFirst filtering (n/total)
Human MTG	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	00:55:33	1	1 (13,945/13945)
	BinaryFirst: moderate (mean + 1 std. dev.)	00:18:10	0.325	0.164 (2292.47/13945)
	BinaryFirst: high (mean + 2 std. dev.)	00:10:58	0.196	0.073 (1022.99/13945)
Kidney	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	68:17:37	1	1 (33,920/33920)
	BinaryFirst: moderate (mean + 1 std. dev.)	06:49:36	0.100	0.013 (447.79/33920)
	BinaryFirst: high (mean + 2 std. dev.)	06:46:40	0.100	0.013 (446.39/33920)
Lung: L5	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	02:37:58	1	1 (29,800/29800)
	BinaryFirst: moderate (mean + 1 std. dev.)	00:44:56	0.284	0.021 (639.61/29800)
	BinaryFirst: high (mean + 2 std. dev.)	00:41:16	0.261	0.020 (609.87/29800)

 Table 1
 Comparison of runtime and number of genes that passed the BinaryFirst filtering criterion between different BinaryFirst configurations

gene lists as centralized resources, it is more often that individual studies provide their own marker gene lists derived for each specific dataset. To compare the performance of different marker gene selection approaches, the Human Lung Cell Atlas (HLCA) core dataset was used as a common comprehensive data resource, consisting of~0.5 million cells clustered into 61 cell types from healthy lung tissues from 107 individuals [18]. The HLCA authors provided cell type-specific marker genes by iteratively subsetting the atlas into sequentially granular classifications and filtering for unique genes within the compartments (hereinafter, HLCA markers). Meanwhile, the lung single cell community also constructed the LungMAP single-cell reference (CellRef) to provide integrated information for both human and mouse lungs [19]. For this comparison, we consider five marker gene lists for healthy human lung cell types: NS-Forest markers, HLCA markers, CellRef markers, ASCT + B markers, and Azimuth markers. The HLCA markers are available in Supplementary Table 6 from the publication [18]. The CellRef markers were extracted from the human Lung-MAP CellCards, which are available from Supplementary Data 2 from the publication [19]. The ASCT + B markers are available from the "lung v1.4" table at the Human Reference Atlas portal (https://humanatlas.io/asctb-tables). The Azimuth markers derived from the HLCA core dataset are pre-calculated and available under "Human-Lung v2 (HLCA)" at the Azimuth portal (https://azimuth. hubmapconsortium.org/). Because CellRef and ASCT + B markers are not directly derived for the HLCA cell types, the Simple Standard for Sharing Ontological Mappings (SSSOM) guideline [20] and Cell Ontology [21] IDs were used to map the CellRef and ASCT + B cell types to the HLCA cell types, resulting in 33 and 18 exact matches, respectively.

To derive NS-Forest markers, NS-Forest v4.0 was applied to the HLCA core dataset and 122 marker genes (1-4 marker genes per type) for the 61 finest level cell types were identified (Supplementary Table 6). Figure 5A shows the NS-Forest marker gene expression in the 61 HLCA cell types ordered according to the dendrogram in Fig. 5B. In this dendrogram, similar cell types are grouped according to the hierarchical clustering of the transcriptome profiles of these cell types; three major branches consisting of immune cells, endothelial and stromal cells, and epithelial cells are observed. With the dendrogram ordering, the NS-Forest marker genes show a strong and clean expression pattern along the main diagonal in the expression dotplot (Fig. 5A). Similar dotplots were produced for the other four marker gene lists (Supplementary Fig. 6). The HLCA marker list contains 162 marker genes (1-5 markers per type) for the 61 cell types, and the dotplot has an expected diagonal pattern in the expression dotplot but with more off-diagonal expressions (Supplementary Fig. 6A). The CellRef marker list contains 115 marker genes (2-7 markers per type) for the 33 exact matched cell types. The CellRef dotplot shows a relatively clean diagonal expresion pattern, although some genes show high levels of expression for multiple cell types (Supplementary Fig. 6B). The ASCT + B marker list contains 80 marker genes (3-5 markers per type) for the 18 exact matched cell types. The ASCT + B dotplot is sparse (Supplementary Fig. 6C) because the manual curation approach based on existing knowledge from the scientific literature does not capture the granularity obtained in single cell-resolution data. The Azimuth marker list contains 535 marker genes (8-10 markers per type) for 56 of these cell types (AT0, Hematopoietic stem cells, Hillock-like, Smooth muscle FAM83D+,



core cell types. **B** Dendrogram of 61 ann_finest_level cell types from HLCA core dataset corresponding to the rows in (A), generated before preprocessing and by the scanpy package (scanpy.tl.dendrogram). **C** Comparing the number of cell types, number of markers, number of unique markers, median F-beta score, median PPV (precision), median recall, and median On-Target Fraction for the NS-Forest, HLCA, CellRef, ASCT + B, and Azimuth marker lists. **D-G** Boxplots of F-beta score, PPV (precision), recall, and On-Target Fraction for the 13 cell types commonly characterized by all methods. **H–K** Boxplots of F-beta score, PPV (precision), recall, and On-Target Fraction for all available cell types across methods

and pre-TB secretory markers are not available). The Azimuth dotplot lacks a clean diagonal pattern with many of the genes being expressed at high levels across many similar cell types (Supplementary Fig. 6D).

The NS-Forest v4.0 Python package was also modularized to enable user-defined marker gene evaluation, allowing for direct comparison of the cell type classification metrics between different input marker lists. Using this approach, the performance of the five marker gene sets was compared using the HLCA data by calculating F-beta score, PPV (precision), recall, and On-Target Fraction for all cell types and directly comparing the medians (Fig. 5C), along with the distributions of these performance metrics for the 13 common

cell types matched across all five reference datasets (Fig. 5D-G) and for all matched cell types (Fig. 5H-K). For the fair comparison of the 13 common cell types, the NS-Forest marker genes have the highest median F-beta score (0.80), followed by the HLCA markers (0.65), Cell-Ref markers (0.39), ASCT+B markers (0.22), and Azimuth markers (0.18). Similar results were found using all cell types (NS-Forest: 0.71, HLCA: 0.43, CellRef: 0.38, ASCT+B: 0.17, and Azimuth: 0.17 using paired t-test, same below). The F-beta scores reflect the global patterns observed in the dotplots of these marker gene lists (Fig. 5A and Supplementary Fig. 6). The distributions of F-beta scores are significantly different between NS-Forest and all other methods (NS-Forest vs. HLCA: p = 0.0039, NS-Forest vs. CellRef: p = 2.0e-5, NS-Forest vs. ASCT+B: p=2.1e-6, NS-Forest vs. Azimuth: p = 1.6e-5). Surprisingly, the median PPV (precision) values are high for all five methods (0.89-0.99), with the ASCT + B being the highest, showing no significant difference between NS-Forest and the other methods (NS-Forest vs. HLCA: p = 0.70, NS-Forest vs. CellRef: p = 0.93, NS-Forest vs. ASCT + B: p = 0.81, and NS-Forest vs. Azimuth: p = 0.098). In contrast, the median recall values show a similar trend to the F-beta scores, with NS-Forest being the highest (0.58) and ASCT + B being the lowest (0.054) in the 13 common cell types. The distributions of recall values are significantly different between NS-Forest and all other methods (NS-Forest vs. HLCA: p=0.036, NS-Forest vs. CellRef: p=3.5e-7, NS-Forest vs. ASCT+B: p=2.0e-6, and NS-Forest vs. Azimuth: p = 3.9e-6). The median On-Target Fraction ranges from the highest (0.55) for NS-Forest and the lowest (0.19) for Azimuth in the 13 common cell types. The difference of the overall distributions of On-Target Fraction between NS-Forest and the other methods are not significant for HLCA and CellRef, and are significant for ASCT + B and Azimuth (NS-Forest vs. HLCA: p=0.61, NS-Forest vs. CellRef: p = 0.73, NS-Forest vs. ASCT + B: p = 0.0071, and NS-Forest vs. Azimuth: p = 1.3e-4). Though the difference in median On-Target Fraction values between NS-Forest and ASCT + B is small in Fig. 5G, the significant *p*-value from the paired t-test can be expained in Supplementary Fig. 7O, where most of the pairs have higher values in NS-Forest. Scatter plots for pairwise comparison between the NS-Forest markers and other markers for each cell type also show superior F-beta and recall results using NS-Forest marker combinations (Supplementary Fig. 7), while having comparable performance in PPV (precision) and On-Target Fraction. It is interesting to note that there are several clusters that have perfect On-Target Fractions for the CellRef markers, but lower recall. These genes are exclusively expressed in the

target cluster, but in a much smaller proportion of cells (an example is highlighted in Supplementary Figs. 6 and 7), which would result in more false negative cells and thus lower recall. It is generally true that the low F-beta and recall values for the other four methods are driven by excessive false negative predictions. Comparing all methods, NS-Forest produces the most comprehensive and concise list of marker genes with consistently higher F-beta scores for cell type classification.

Comparison with other marker gene selection methods

To directly compare NS-Forest v4.0 for cell-type-specific marker gene selection, four other methods: COMET [22], RankCorr [23], scGeneFit [24], and MarkerMap [25] were considered. Two of the methods were previously directly compared with NS-Forest v2.0 [13]. Here we present the results of all six methods (two versions of NS-Forest and four other methods) applied on the Villani et al. dataset [26], which deep sequenced ~ 1,000 cells from healthy blood samples across six dendritic cell (DC) and four monocyte (Mono) populations. To better delineate cell types, we utilized the Louvain clustering done in Aevermann et al. [13], which merged the DC2 and DC3 clusters as well as the Mono1 and Mono3 clusters, resulting in eight distinct clusters (Fig. 6A). Each method was iteratively run, generating a set of marker genes for each cluster (Supplementary Table 7). Dotplots of the markers identified by each method are shown in Fig. 6B-G, where NS-Forest v4.0 showed the cleanest cluster-specific expression with genes that show strong binary expression patterns. NS-Forest v4.0 identified the fewest marker genes (14 unique markers for the 8 clusters), followed by COMET, MarkerMap, and scGeneFit (15 unique markers), NS-Forest v2.0 (16 unique markers), and RankCorr (28 unique markers). As expected, NS-Forest v4.0 had the highest binary scores for its selected genes (Fig. 6H). The classification metrics of all six methods are shown side-by-side in Fig. 6I-L. Comparing the median F-beta scores, NS-Forest v2.0 had the highest score (0.87), followed by scGeneFit (0.82), RankCorr (0.80), NS-Forest v4.0 (0.79), COMET (0.68), and MarkerMap (0.24). The slightly lower F-beta for NS-Forest v4.0 is an expected trade-off for the higher binary scores. Most methods had relatively high PPVs of 0.8 or greater. The recall values tended to be lower with broader ranges, of which scGeneFit had the highest median recall close to 0.8. It is interesting to note that scGeneFit identified some negative marker genes (e.g., HLA-DRB4 and TYROBP in Fig. 6E), which is helpful for ruling out false negatives in a classification model but would be difficult to use for certain downstream experiments in practice. The On-Target Fractions displayed greater variabilities, where the two



Fig. 6 Direct comparison of six marker gene selection methods on an immune cell dataset. A UMAP of the clusters produced by Louvain clustering for the monocyte and dendritic cell types described in Villani et al. [26]. B-G Dotplots of the marker genes selected by NS-Forest v2.0 (B), NS-Forest v4.0 (C), RankCorr (D), scGeneFit (E), COMET (F), and MarkerMap (G). H Boxplots of Binary Expression Scores of the marker genes selected by each method. I-L Boxplots of F-beta score, PPV (precision), recall, and On-Target Fraction for performance comparison across all six methods on the Villani et al. dataset

versions of NS-Forest had the top two median On-Target Fraction values, consistent with the amount of off-diagonal expression observed in the dotplots.

Performance evaluation in simulation studies

To further understand the properties of the metrics used in NS-Forest, simulation studies were conducted using a zero-inflated three-component mixture model with varying zero inflation levels (see Methods). Figure 7A shows the simulated gene expression patterns. Genes 1-5 are true marker genes with simulated gene expression patterns (i-v); genes 6–10 are non-marker genes with expression patterns (vi-x); genes 11-100 are null genes with expression pattern (xi). The null genes can be considered as "white noise" genes, the inclusion of which simulates the long list of non-informative genes as input features to the random forest model in this simplistic simulation design. Figure 7B-C show the performance of the genecentric metrics Binary Score and On-Target Fraction with respect to the varying zero inflation levels. Evaluating in cluster 1, i.e. the target cluster, both metrics show large error bars for larger zero inflation levels. The order of the performance curves of genes 1-5 are expected, where gene 1 is the most ideal marker gene for cluster 1, with a Binary Score = 1 and an On-Target Fraction = 1 across all zero inflation levels. In Fig. 7B, genes 1-3 have high Binary Scores, while genes 4-5 have decreased Binary Scores and high variabilities when the zero inflation levels become high, suggesting that the Binary Score is robust to patterns (i-iii) where there is little-to-no expression in the non-target clusters, and effective for patterns (iv-v) where there are quantitatively lower expression in the non-target clusters. In Fig. 7C, the On-Target Fraction effectively captures the drop of genes 2–3 from gene 1 due to the changes in patterns (i-iii), suggesting that the Binary Score and On-Target Fraction capture complementary properties of a marker genes.

NS-Forest v2.0/v3.9 and v4.0 were both applied to the simulated data. Since true marker genes were only simulated for clusters 1–5, the cluster-centric performance metrics are shown in Fig. 8 for these clusters. While the F-beta score, PPV, and On-Target Fraction of the two versions of NS-Forest are very similar, recall is significantly higher (t-test p-value=2.2e-16) and less variable in v4.0. The performance metrics that show a prominent decreasing trend with respect to the zero inflation levels are F-beta score for cluster 1 and recall for all clusters. The impact on recall values is expected as a direct consequence of increased false negatives with zero inflation. The decreasing trend of F-beta score in cluster 1 suggests that the misclassification of cells in cluster 1, where there is a perfect marker (i.e., gene 1), comes from the false negatives when the perfect marker expression was impacted by dropouts. The simulation studies also showed an interesting observation of the chances of selecting non-marker genes for the two NS-Forest versions. Out of the 20 iterations (Supplementary Fig. 8), the non-marker genes 6-10 were randomly selected by



Fig. 7 Performance evaluation in simulation studies. A Examples of simulated gene expression patterns with 20% zero inflation. Simulation design is summarized in the text box. B-C Evaluation of the gene-centric metrics – Binary Score and On-Target Fraction – with respect to zero inflation in the simulations

NS-Forest v2.0/v3.9 at all zero inflation levels, but were not selected by NS-Forest v4.0 in most of the simulations. Among all simulation results, the chance of selecting any non-marker gene is 13% (223/1672=total count of selecting genes 6–10 as marker genes / total count of selected marker genes) in v2.0/v3.9 and 0.3% (5/1481) in v4.0.

Discussion

This paper describes major algorithmic refinements made in NS-Forest v4.0 and its improved performance on the human brain MTG dataset used to develop the previous versions, as well as its performance on datasets from the human kidney and lung. The main motivation for developing version 4.0 was to improve the marker gene selection performance on closely related cell types by enriching for markers that exhibit the pattern of being highly and uniquely expressed in their target cell types without losing a significant amount of classification power. The BinaryFirst step was introduced in the NS-Forest workflow to enrich for candidate genes that exhibit the desired gene expression pattern. It is essentially an informative dimensionality reduction approach that effectively reduces the size of the set of candidate genes prior to the random forest classification step, which is usually the most time-consuming part of the algorithm. As a result, the overall runtime of the algorithm is substantially reduced. To explicitly demonstrate the improvements made by this algorithmic refinement, we introduced the On-Target Fraction metric that quantifies how well the NS-Forest marker genes are exclusively expressed in each distinct cell type.

Overall, this new version of NS-Forest demonstrated clear improvement in the human MTG brain dataset, which is the most complex organ evaluated in this study. Additional datasets representing the human kidney and



Fig. 8 NS-Forest performance in simulation studies. Evaluation of the cluster-centric performance metrics – F-beta score, PPV, recall and On-Target Fraction – for NS-Forest v2.0/3.9 and v4.0 with respect to zero inflation in the simulations. Using two-sample t-test (v2.0/3.9 vs. v4.0), *p*-values = 0.7919 for F-beta score, 0.5738 for PPV, 2.2e-16 for recall, 0.2798 for On-Target Fraction

lung were used to validate NS-Forest's performance on data from other organs. NS-Forest v4.0 is now a comprehensive Python package that not only implements the algorithm for obtaining the marker genes, but also supports the marker gene evaluation functions in a machine learning framework for cell type classification. In this paper, we also presented a comparative analysis of the NS-Forest marker genes and other popular marker gene lists. We formally established the notion of cell type classification marker genes, which are different from the notion of differentially expression genes. In the headto-head comparison using the HLCA dataset with half a million cells, the NS-Forest marker genes showed superior performance over the HLCA, CellRef, ASCT + B and Azimuth marker sets.

One of the innovations of the NS-Forest approach is the enrichment of binary genes ranked by the novel Binary Expression Score. The algorithm also outputs the top 10 binary genes as part of the supplementary results, which may serve as an extended list of genes of interest for future study or experiment design. As one future direction to explore, we also looked at the potential co-expression pattern of these binary genes in the HLCA dataset (the binary genes can be found in Supplementary Table 6). In the co-expression heatmap (Supplementary Fig. 9), we highlighted regions where the cell-type-specific binary genes showed very strong coexpression (e.g., the black box) and where the cell-typespecific binary genes of several related clusters showed overlapping co-expression (e.g., the yellow box). These patterns can be observed along the diagonal for many of the HLCA cell types at the finest level annotation. Though NS-Forest is not explicitly designed for detecting cell-type-specific co-expression gene networks, the binary gene list does identify upregulated co-expression gene networks for many cell types, which may be futher explored to provide a complementary perspective to those co-expression inference methods based on statistical approaches [27, 28].

NS-Forest marker genes can be used for multiple downstream experimental investigations, such as spatial transcriptomics gene panel design in the SpaceTx consortium [29], and to produce marker gene sets designed to capture specific cell type properties. One of the main applications of NS-Forest identified marker genes is contributing to the definition of ontological classes of scRNA-seq data-driven cell types for incorporation into the official Cell Ontology [21], as NS-Forest provides the minimum combinations of marker genes that can serve as a set of definitional characteristics of the cell types [21]. Such efforts have already begun, as NS-Forest has contributed to the BRAIN Initiative Cell Census Network (BICCN) data ecosystem to derive the necessary and sufficient marker gene knowledge [30]. As such, the Provisional Cell Ontology (PCL) is generated in this manner for the human, mouse, and marmoset primary motor cortex [31]. Among the general single cell community, there is a current lack of a formal, standardized

representation of cell type clusters derived from the tremendous amount of scRNA-seq data and their transcriptional characterization that is widely accepted by the scientific community. One of the challenges associated with formalizing such a representation is the aspect of scaling up the semantic knowledge representations to keep up with the rate at which single cell transcriptomic data and analyses are being produced today. To this end, NS-Forest appears well-suited to help alleviate some of the challenges associated with such a task, especially with the enhancements introduced in version 4.0.

Methods

BinaryFirst step in NS-Forest v4.0

The BinaryFirst step is introduced in version 4.0 of NS-Forest and is implemented in the gene pre-selection step of the workflow. Essentially, this process reduces the number of genes that are later considered in the random forest step as candidate marker genes for each cluster by only selecting genes with a Binary Expression Score greater than or equal to a dataset-specific threshold based on the distribution of this score (Fig. 2). This step is important because it substantially reduces the feature space that is input into the random forest, significantly decreasing the runtime of NS-Forest. The Binary Expression Scores are first calculated for each gene-cluster pair in the dataset using the formula defined in the paper detailing version 2.0 [13], and the formula is restated in the Binary Expression Score section below. Users can specify which type of threshold is used in the BinaryFirst step: 'none', 'BinaryFirst_mild', 'BinaryFirst_moderate', or 'BinaryFirst_high'. If the threshold is 'none,' no filtering is performed, and all genes in the input anndata object are considered for the iterative search in the random forest step. The other threshold values are calculated based on the distribution of Binary Expression Scores from all genes, and so these values vary depending on the input dataset used. The mild threshold is set as the median Binary Expression Score, the moderate threshold is set as the mean Binary Expression Score plus the standard deviation of all scores, and the high threshold is set as the mean Binary Expression Score plus two times the standard deviation. In version 4.0, the default is set to 'BinaryFirst_high,' which is the most stringent threshold that filters the most genes.

The median Binary Expression Score in a scRNA-seq dataset is often 0, as is the case for the human MTG, kidney, and lung datasets used in this paper. This is expected because most genes are not useful for distinguishing granular cell types at single cell resolution. In such cases, the 'BinaryFirst_mild' model in version 4.0 is equivalent to the model used in version 2.0/3.9, since no initial filtering is done when NS-Forest is run with this threshold. Thus, the results obtained from running version 4.0 with the 'BinaryFirst_mild' threshold is equivalent to results obtained from running NS-Forest version 2.0 or 3.9 (no algorithmic difference between versions 2.0 and 3.9). By default, NS-Forest v4.0 uses the 'BinaryFirst_high' threshold, unless otherwise stated.

For the human middle temporal gyrus (MTG) dataset, the median Binary Expression Score is 0, indicating that more than half of the original input genes have zero median expression in these clusters and are therefore non-informative. In the human MTG dataset, the mean Binary Expression Score is 0.167 and the standard deviation is 0.249, implying the thresholds for this specific dataset are as follows: mild=0, moderate=0.176+0.249=0.415, high=0.176+2*0.249=0.664. For the human kidney and lung datasets, the median Binary Expression Score is also 0. The moderate and high thresholds are 0.102 and 0.194 for the kidney dataset, respectively, and 0.118 and 0.222 for the lung dataset,

All runtimes discussed in the Results section detailing the improvement obtained with the BinaryFirst step were obtained from running NS-Forest through jobs that were submitted to the *Expanse* supercomputer system in the San Diego SuperComputer Center.

Binary expression score

Here is a recap of the Binary Expression Score from our earlier study [13]. The Binary Expression Score is defined as below. For each gene g evaluated in target cluster T,

Score_{gT} =
$$\frac{\sum_{i=1}^{n} \left(1 - \frac{m_{gi}}{m_{gT}}\right)^{+}}{n-1}$$
,

where m_{gT} = median expression of gene g in the target cluster T, and m_{gi} = median expression of gene g in cluster i for i = 1, ..., n. The mathematical symbol $(\cdot)^+$ denotes the positive part of a real valued function, meaning max $(\cdot, 0)$. In the most ideal case where $m_{gT} = x$ for some positive value x > 0 and $m_{gi} = 0$ for all $i \neq T$, the score Score_{gT} = 1; in the least ideal case where $m_{gT} < m_{gi}$ for all $i \neq T$, the score Score_{gT} = 0. The Binary Expression Score has a range of [0,1].

On-Target fraction metric

In version 4.0, a new On-Target Fraction metric is provided, to quantify the expression specificity of the marker genes with respect to their target cell types. In previous versions, the algorithm reported the F-beta score and Positive Predictive Value (PPV) together with the true/ false positives/negatives (i.e., TP, FP, TN, FN), which are metrics that quantify the discriminative power of each set of markers for their cell type classification performance. However, these metrics do not fully capture how well NS-Forest achieves the ideal scenario of identifying markers for each cluster that are exclusively expressed in that cluster. To make a clear distinction, we refer to F-beta score, PPV, and recall as *classification* metrics, and On-

Target Fraction as an *expression* metric. The On-Target Fraction is defined for each marker gene *g* in target cluster *T* as, cells in each cluster (Fig. 7). Genes 1–5 are true marker genes with expression patterns (i-v), where $\mu_T = 10$ for the true targeting cluster and $\mu_i = 0,0,0,5,7, \forall i \neq T$ for genes 1–5, respectively. Genes 6–10 are non-marker genes receiving the same level of signals across all clusters as shown in expression patterns (vi-x), where $\mu_i = 7,6,5,4,3, \forall i = 1,...,20$ for genes 6–10, respectively. In this simulation design, the Gamma component

OnTarget Fraction -	median expression in target cluster	m_{gT}
Onrarget Tractiong	sum of median expression across all clusters	$-\frac{1}{\sum_{i=1}^{n}m_{gi}}$,

where m_{gT} = median expression of marker gene g in target cluster T, and m_{gi} = median expression of marker gene g in cluster i for i = 1, ..., n. This metric has a range of [0,1], with a value of 1 being the ideal case where this marker gene is "exclusively" expressed in more than half of the cells in its target cluster and in fewer than half of the cells in all other clusters. Due to the zero-inflation nature of scRNA-seq data, this metric can effectively capture those genes that have abundant expression exclusively within the target cluster. At the cluster level, we report the On-Target Fraction for each cluster using the median Fraction_{gT} of its marker genes. We used the median as a summary statistic when reporting the On-Target Fraction to account for the non-normal nature of scRNA-seq gene expression distribution.

Simulation design

The same simulation model is used to generate the simulated data as reported earlier [13]. A zero-inflated three-component mixture model is used in the simulation design to reflect the dropout technical artifact (zero-inflation), background and positive expression signals as observed in real data distributions. Let X denote the gene expression value. X follows a mixture distribution such that.

$$P(X = x) = \pi_1 \cdot \delta_0(x) + \pi_2 \cdot f_{Gamma}(x) + \pi_3 \cdot f_{Normal}(x)$$

The zero-inflation component is $\delta_0(x) = \text{probabil-}$ ity density function of the degenerate distribution at 0; the Gamma distribution component $f_{\text{Gamma}}(x)$ = probability density function of the distribution Gamma($\alpha = 1, \beta = 1$) with mean=1, representing the background expression; and the Normal component $f_{\text{Normal}}(x) = \text{probability density function of the distribu$ $tion Normal(<math>\mu = \mu_i, \sigma^2 = 1$) for cluster *i*, representing the positive expression signals. The parameters π_1, π_2 , and π_3 corresponds to the weights of each component such that $\pi_1, \pi_2, \pi_3 > 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$.

Based on the above model, 10 genes with expression patterns (i-x) were simulated across 20 clusters with 300

is set at constant $\pi_2 = 0.1$. While the zero-inflation component π_1 varies from 0.05 to 0.45, the Normal component is $\pi_3 = 0.9 - \pi_1$. Other than the 10 genes receiving expression patterns, 90 non-expression genes were simulated as shown in pattern (xi), where $\pi_3 = 0$. In total, 100 genes and 6000 cells were simulated. This simulation was repeated 20 times.

NS-Forest python package

With the continuous refinements of the NS-Forest algorithm and its marker evaluation metrics, NS-Forest has become a comprehensive software package. To provide a user-friendly software package, NS-Forest v4.0 is now modularized, consisting of 4 main functional modules: preprocessing, NSForesting, evaluating, and plotting. NS-Forest v4.0 takes in an anndata [15] object in.h5ad format that contains a cell-by-gene expression matrix and the cell type cluster membership column stored in the observation-level (.obs) metadata matrix of the data object.

An optional but suggested step before preprocessing is to generate a hierarchical clustering dendrogram of the cell type clusters on the full dataset. This occurs before any gene filtering because the dendrogram should be consistent between various preprocessing methods. In the preprocessing module, the first step is to calculate the median expression matrix for each gene in each cluster. The default positive_genes_only parameter is true, which filters for genes with a positive median expression in at least one cluster. (Note that this preprocessing step based on medians is specific to NS-Forest and should not be used if evaluating the performance of marker genes produced by other approaches.) Based on the median expression matrix, the Binary Expression Score of each gene (positive genes only by default) is calculated in each cluster. The Binary Expression Score has a range of [0,1], with values closer to 1 indicating a higher level of binary expression (i.e., the gene is expressed in the target cluster and not others). The pre-calculated median expression matrix and the Binary Expression Score matrix are saved in the unstructured metadata slot (.uns) of the data object.

In the NSForesting module, the BinaryFirst step is implemented with the gene_selection parameter that determines the BinaryFirst criterion (None: 0, Binary-First_mild: median, BinaryFirst_moderate: mean+std, and BinaryFirst_high: mean+2std), which filters out genes below the chosen threshold. The main step in the NS-Forest algorithm is building a random forest classifier for each cluster that is trained on the genes that passed the gene_selection criteria. In this step, each gene is ranked by the Gini Impurity index and the *n* top genes with the highest Gini index are then reranked by their pre-calculated Binary Expression Scores. The n_gene_eval value indicates how many genes from the reranked candidate gene list are input into the decision tree evaluation for determining the best combinatorial marker genes as final output. A single split decision tree is built for each evaluated gene. Gene combinations of all set lengths are evaluated, and the combination with the highest F-beta score is considered the final set of NS-Forest marker genes for that cluster. The performance metrics returned from this module for each cluster are the F-beta score, PPV (precision), recall, TP, FP, TN, FN, and **On-Target Fraction.**

The evaluating module can be called independently without calling the preprocessing and NSForesting module. This module is useful for calculating the metrics for a user-input marker gene list with paired cluster names to compare the cell type classification performance and marker expression across different marker gene lists. We provide an option of using mean instead of median for the On-Target Fraction calculation, to account for cases where a user-input gene is only expressed in a small proportion of cells in the target cluster and has absolutely no expression in other clusters.

The plotting module creates the scanpy [15] dot plot, stacked violin plot, and matrix plot figures for visualization of the NS-Forest or user-input marker genes with clusters organized according to the dendrogram order (from the preprocessing step) or in order corresponding to user-input. Other plotting functions include creating interactive plotly [32] boxplots and scatter plots, which are useful for comparing metrics and identifying clusters of interest.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44330-024-00015-2.

Supplementary Figure 1. Comparing performance of using different BinaryFirst thresholds in NS-Forest v4.0 on specific subclades within human MTG dataset. (A) Hierarchical dendrogram derived in the original human MTG study with labelled and color-coded subclades (https://github.com/ AllenInstitute/MOp_taxonomies_ontology/tree/main). (B) Heatmaps of markers from the human middle temporal gyrus (MTG) dataset generated from NS-Forest v4.0 with 'BinaryFirst_midd,' BinaryFirst_moderate', and 'BinaryFirst_high' thresholds. The regions on the heatmaps highlighted by the orange boxes correspond to the identified markers for the cell types in specific subclades (VIP, PVALB, and L4 subclades) that are known to be more similar to each other and thus, more difficult to distinguish. The colors correspond to the normalized median expression level (log2-transformed counts per million) for the marker gene (rows) in a given cell type cluster (columns), with high expression in red/yellow, and low expression in blue/white. The clusters are ordered according to the hierarchical dendrogram provided in the original study shown in (A). (C) Median On-Target Fraction values within each of the three specific subclades across these three BinaryFirst thresholds.

Supplementary Figure 2. Comparison of On-Target Fraction distribution in major subclades of Human MTG dataset across BinaryFirst thresholds. Boxplots comparing the distribution of the On-Target Fraction values within each of the three specific subclades in the human MTG dataset (L4, PVALB, and VIP subclades) across the mild, moderate, and high BinaryFirst thresholds are shown.

Supplementary Figure 3. Additional investigation of mean + 3 standard deviations (SD) BinaryFirst threshold evaluated in the human MTG dataset. (A-B) Heatmaps of NS-Forest marker genes using the BinaryFirst threshold of mean + 3 SD in the human MTG dataset, without and with the VIP, PVALB, and L4 subclades highlighted. (C-D) Performance metrics using the mean + 3 SD threshold in the human MTG dataset, directly comparable with Fig. 3D and Supplementary Fig. 1C. (E) Scatter plots and the best linear relationship fitted for the number of input genes to the random forest (RF) step after BinaryFirst filtering using different thresholds with respect to the On-Target Fraction values per cluster.

Supplementary Figure 4. Evaluating performance of NS-Forest on lung L4 and L3 subclass datasets. (A) Heatmaps of NS-Forest v2.0/v3.9 and v4.0 markers from the L4 and L3 subclasses of the human lung. The colors correspond to the normalized median expression level (log2-transformed counts per million) for the marker gene (rows) in a given cell type cluster (columns), with high expression in red/yellow, and low expression in blue/white. The clusters are ordered according to the hierarchical ordering in the dendrogram generated by the scanpy package (*scanpy.tl.dendrogram*) using default settings. (B) Comparison of the performance metrics corresponding to the NS-Forest results shown in (A).

Supplementary Figure 5. Distribution of median gene expression per cluster and Binary Expression Score in human MTG, kidney, and lung datasets. First row: histograms of distribution of median gene expression values of genes expressed in all clusters in human MTG, kidney, and lung datasets. X-axis displays the range of median gene expression values in each dataset, and the y-axis displays the frequency of each median gene expression value (log scale). Second row: histograms of distribution of Binary Expression Score values of genes in these three datasets. X-axis ranges from 0 to 1 (representing the possible values the binary expression score can be), and the y-axis displays the frequency of each binary score value (log scale). All distributions are highly right-skewed.

Supplementary Figure 6. Dotplots of HLCA, CellRef, ASCT + B, and Azimuth marker genes on HLCA core cell types. (A) 162 HLCA markers across 61 HLCA cell types. (B) 115 CellRef markers across 33 HLCA cell types. The highlighted example is where the On-Target Fraction is perfect but recall is low. (C) 80 ASCT + B markers across 18 HLCA cell types. (D) 535 Azimuth markers across 56 HLCA cell types.

Supplementary Figure 7. NS-Forest consistently outperforms other published lung marker genes in classification performance. (A-D) Scatter plots comparing F-beta scores for each cell type using NS-Forest markers vs. HLCA, CellRef, ASCT + B, and Azimuth markers. (E–H) Scatter plots comparing PPV (precision) for each cell type using NS-Forest markers vs. HLCA, CellRef, ASCT + B, and Azimuth markers. (I-L) Scatter

Page 18 of 20

plots comparing recall for each cell type using NS-Forest markers vs. HLCA, CellRef, ASCT + B, and Azimuth markers. (M-P) Scatter plots comparing On-Target Fraction for each cell type using NS-Forest markers vs. HLCA, CellRef, ASCT + B, and Azimuth markers. The highlighted example is where the On-Target Fraction is perfect but recall is low.

Supplementary Figure 8. Simulation results of NS-Forest v2.0/3.9 and v4.0. Heatmap of the number of times that a gene is selected as NS-Forest marker gene in the simulations. By the simulation design, genes 1–5 are true marker genes, genes 6–10 are non-marker genes, and genes 11–100 are null genes (not shown as they won't be selected by design).

Supplementary Figure 9. Heatmap of co-expression for the binary genes outputted from the NS-Forest algorithm with the HLCA dataset.

Supplementary Table 1. NS-Forest v4.0 results of the human brain middle temporal gyrus dataset.

Supplementary Table 2. NS-Forest v4.0 results of the human kidney dataset.

Supplementary Table 3. NS-Forest v4.0 results of the human lung L5 subclass dataset.

Supplementary Table 4. NS-Forest v4.0 results of the human lung L4 subclass dataset.

Supplementary Table 5. NS-Forest v4.0 results of the human lung L3 subclass dataset.

Supplementary Table 6. NS-Forest v4.0 results of the Human Lung Cell Atlas dataset.

Supplementary Table 7. Marker genes selected for the immune dataset by six methods.

Authors' contributions

A.L., R.H.S. and Y.Z. conceived the project. A.L. and Y.Z. designed and implemented the algorithm. B.P. built the software package. A.L., B.P. and A.V.P. conducted the data analyses. T.E.D. and G.P. provided the unpublished data and supervised the analysis on it. A.L., B.P., R.H.S., and Y.Z. interpreted the results. A.L., B.P., R.H.S., and Y.Z. wrote the manuscript. All authors agreed on the contents of the manuscript.

Funding

This work was supported by the U.S. National Institutes of Health (1RF1MH123220, 1R03OD036499, OT2OD033756, U54HL165443, and U54HL145608) and the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health. The funding bodies had no role in the design or conclusions of this study.

Data availability

The human brain Middle Temporal Gyrus (MTG) dataset was downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq. The human kidney dataset was downloaded from https://cellxgene.cziscience. com/collections/bcb61471-2a44-4d00-a0af-ff085512674c (Integrated Single-nucleus and Single-cell RNA-seq of the Adult Human Kidney). The human lung dataset is currently unpublished (manuscript under submission). The Human Lung Cell Atlas (HLCA) dataset was downloaded from https://cellxgene.czisc ience.com/collections/6f6d381a-7701-4781-935c-db10d30de293 (core).

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹ Department of Informatics, J. Craig Venter Institute, La Jolla, CA, USA. ²Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ³Department of Pediatrics, Division of Respiratory Medicine, University of California, La Jolla, San Diego, CA, USA. ⁴Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA.

Received: 13 July 2024 Accepted: 17 October 2024 Published online: 04 November 2024

References

- Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, Miller JA, Lee A, Chang I, McCorrison J, Pulendran B, Qian Y, Schork NJ, Lasken RS, Lein ES, Scheuermann RH. Cell type discovery and representation in the era of high-content single cell phenotyping. BMC Bioinformatics. 2017;18(Suppl 17):559. https://doi.org/10.1186/s12859-017-1977-1. Epub 2018/01/13; PubMed PMID: 29322913; PMCID: PMC5763450.
- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Gottgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N, Human Cell Atlas Meeting P. The Human Cell Atlas. Elife. 2017;6:e27041. https://doi.org/10.7554/eLife.27041.
- HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. 2019;574(7777):187–92. https://doi.org/10.1038/s41586-019-1629-x. Epub 2019/10/11; PubMed PMID: 31597973; PMCID: PMC6800388.
- Mott MC, Gordon JA, Koroshetz WJ. The NIH BRAIN Initiative: Advancing neurotechnologies, integrating disciplines. PLoS Biol. 2018;16(11):e3000066. https://doi.org/10.1371/journal.pbio.3000066. Epub 2018/11/27; PubMed PMID: 30475794; PMCID: PMC6283590.
- Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, Yao Z, Eggermont J, Hollt T, Levi BP, Shehata SI, Aevermann B, Beller A, Bertagnolli D, Brouner K, Casper T, Cobbs C, Dalley R, Dee N, Ding SL, Ellenbogen RG, Fong O, Garren E, Goldy J, Gwinn RP, Hirschstein D, Keene CD, Keshk M, Ko AL, Lathia K, Mahfouz A, Maltzer Z, McGraw M, Nguyen TN, Nyhus J, Ojemann JG, Oldre A, Parry S, Reynolds S, Rimorin C, Shapovalova NV, Somasundaram S, Szafer A, Thomsen ER, Tieu M, Quon G, Scheuermann RH, Yuste R, Sunkin SM, Lelieveldt B, Feng D, Ng L, Bernard A, Hawrylycz M, Phillips JW, Tasic B, Zeng H, Jones AR, Koch C, Lein ES. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573(7772):61–8. https://doi.org/10.1038/s41586-019-1506-7. Epub 2019/08/23; PubMed PMID: 31435019; PMCID: PMC6919571.
- Siletti K, Hodge R, Mossi Albiach A, Lee KW, Ding SL, Hu L, Lönnerberg P, Bakken T, Casper T, Clark M, Dee N, Gloe J, Hirschstein D, Shapovalova NV, Keene CD, Nyhus J, Tung H, Yanny AM, Arenas E, Lein ES, Linnarsson S. Transcriptomic diversity of cell types across the adult human brain. Science. 2023;382(6667):eadd7046. https://doi.org/10.1126/science.add70 46. Epub 2023/10/12; PubMed PMID: 37824663.
- Lake BB, Menon R, Winfree S, Hu Q, Ferreira RM, Kalhor K, Barwinska D, Otto EA, Ferkowicz M, Diep D. An atlas of healthy and injured cell states and niches in the human kidney. Nature. 2023;619(7970):585–94.
- Multimodal human lung atlas with transcriptomic and regulatory profiling from the same nucleus. Manuscript in preparation. Data available at https://portal.hubmapconsortium.org/.
- Jain S, Pei L, Spraggins JM, Angelo M, Carson JP, Gehlenborg N, Ginty F, Gonçalves JP, Hagood JS, Hickey JW, Kelleher NL, Laurent LC, Lin S, Lin Y, Liu H, Naba A, Nakayasu ES, Qian WJ, Radtke A, Robson P, Stockwell BR, Van de Plas R, Vlachos IS, Zhou M, Börner K, Snyder MP. Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). Nat Cell

Biol. 2023;25(8):1089–100. https://doi.org/10.1038/s41556-023-01194-w. Epub 2023/07/20; PubMed PMID: 37468756; PMCID: PMC10681365.

- Börner K, Teichmann SA, Quardokus EM, Gee JC, Browne K, Osumi-Sutherland D, Herr BW 2nd, Bueckle A, Paul H, Haniffa M, Jardine L, Bernard A, Ding SL, Miller JA, Lin S, Halushka MK, Boppana A, Longacre TA, Hickey J, Lin Y, Valerius MT, He Y, Pryhuber G, Sun X, Jorgensen M, Radtke AJ, Wasserfall C, Ginty F, Ho J, Sunshine J, Beuschel RT, Brusko M, Lee S, Malhotra R, Jain S, Weber G. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. Nat Cell Biol. 2021;23(11):1117–28. https://doi.org/10.1038/ s41556-021-00788-6. Epub 2021/11/10; PubMed PMID: 34750582; PMCID: PMC10079270.
- Börner K, Blood PD, Silverstein JC, Ruffalo M, Satija R, Teichmann SA, Pryhuber G, Misra RS, Purkerson J, Fan J, Hickey JW, Molla G, Xu C, Zhang Y, Weber G, Jain Y, Qaurooni D, Kong Y; HRA Team; Bueckle A, Herr BW 2nd. Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas Construction and Usage. bioRxiv [Preprint]. 2024:2024.03.27.587041. https:// doi.org/10.1101/2024.03.27.587041.
- Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29. https://doi.org/10.1016/j. cell.2021.04.048. Epub 2021/06/02; PubMed PMID: 34062119; PMCID: PMC8238499.
- Aevermann B, Zhang Y, Novotny M, Keshk M, Bakken T, Miller J, Hodge R, Lelieveldt B, Lein E, Scheuermann RH. A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. Genome Res. 2021;31(10):1767–80. https://doi.org/10.1101/gr.275569.121. Epub 2021/06/06; PubMed PMID: 34088715; PMCID: PMC8494219.
- Aevermann BD, Novotny M, Bakken T, Miller JA, Diehl AD, Osumi-Sutherland D, Lasken RS, Lein ES, Scheuermann RH. Cell type discovery using single-cell transcriptomics: implications for ontological representation. Hum Mol Genet. 2018;27(R1):R40–7.
- Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/ s13059-017-1382-0. Epub 2018/02/08; PubMed PMID: 29409532; PMCID: PMC5802054.
- Le H, Peng B, Uy J, Carrillo D, Zhang Y, Aevermann BD, Scheuermann RH. Machine learning for cell type classification from single nucleus RNA sequencing data. PLoS ONE. 2022;17(9): e0275070.
- Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd edition. New York: Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7.
- 18. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, Markov NS, Zaragosi LE, Ji Y, Ansari M, Arguel MJ, Apperloo L, Banchero M, Bécavin C, Berg M, Chichelnitskiy E, Chung MI, Collin A, Gay ACA, Gote-Schniering J, Hooshiar Kashani B, Inecik K, Jain M, Kapellos TS, Kole TM, Leroy S, Mayr CH, Oliver AJ, von Papen M, Peter L, Taylor CJ, Walzthoeni T, Xu C, Bui LT, De Donno C, Dony L, Faiz A, Guo M, Gutierrez AJ, Heumos L, Huang N, Ibarra IL, Jackson ND, Kadur Lakshminarasimha Murthy P, Lotfollahi M, Tabib T, Talavera-López C, Travaglini KJ, Wilbrey-Clark A, Worlock KB, Yoshida M, van den Berge M, Bossé Y, Desai TJ, Eickelberg O, Kaminski N, Krasnow MA, Lafyatis R, Nikolic MZ, Powell JE, Rajagopal J, Rojas M, Rozenblatt-Rosen O, Seibold MA, Sheppard D, Shepherd DP, Sin DD, Timens W, Tsankov AM, Whitsett J, Xu Y, Banovich NE, Barbry P, Duong TE, Falk CS, Meyer KB, Kropski JA, Pe'er D, Schiller HB, Tata PR, Schultze JL, Teichmann SA, Misharin AV, Nawijn MC, Luecken MD, Theis FJ. An integrated cell atlas of the lung in health and disease. Nat Med. 2023;29(6):1563-77. https://doi.org/10.1038/ s41591-023-02327-2. Epub 2023/06/09; PubMed PMID: 37291214; PMCID: PMC10287567
- Guo M, Morley MP, Jiang C, Wu Y, Li G, Du Y, Zhao S, Wagner A, Cakar AC, Kouril M, Jin K, Gaddis N, Kitzmiller JA, Stewart K, Basil MC, Lin SM, Ying Y, Babu A, Wikenheiser-Brokamp KA, Mun KS, Naren AP, Clair G, Adkins JN, Pryhuber GS, Misra RS, Aronow BJ, Tickle TL, Salomonis N, Sun X, Morrisey EE, Whitsett JA, Xu Y. Guided construction of single cell reference for human and mouse lung. Nat Commun. 2023;14(1):4566. https://doi.org/10.1038/

s41467-023-40173-5. Epub 2023/07/30; PubMed PMID: 37516747; PMCID: PMC10387117.

- Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, Chute CG, Duncan WD, Evelo CT, Gabriel D, Graybeal J, Gray A, Gyori BM, Haendel M, Harmse H, Harris NL, Harrow I, Hegde HB, Hoyt AL, Hoyt CT, Jiao D, Jiménez-Ruiz E, Jupp S, Kim H, Koehler S, Liener T, Long Q, Malone J, McLaughlin JA, McMurry JA, Moxon S, Munoz-Torres MC, Osumi-Sutherland D, Overton JA, Peters B, Putman T, Queralt-Rosinach N, Shefchek K, Solbrig H, Thessen A, Tudorache T, Vasilevsky N, Wagner AH, Mungall CJ. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database (Oxford). 2022;2022:baac035. https://doi.org/10.1093/database/baac035.
- Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, Van Slyke CE, Vasilevsky NA, Haendel MA, Blake JA, Mungall CJ. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7(1):44. https://doi.org/10.1186/s13326-016-0088-7. Epub 2016/07/06; PubMed PMID: 27377652; PMCID: PMC4932724.
- Delaney C, Schnell A, Cammarata LV, Yao-Smith A, Regev A, Kuchroo VK, Singer M. Combinatorial prediction of marker panels from single-cell transcriptomic data. Mol Syst Biol. 2019;15(10):e9005. https://doi.org/10. 15252/msb.20199005. Epub 2019/10/28; PubMed PMID: 31657111; PMCID: PMC6811728.
- Vargo AHS, Gilbert AC. A rank-based marker selection method for high throughput scRNA-seq data. BMC Bioinformatics. 2020;21(1):477. https:// doi.org/10.1186/s12859-020-03641-z. Epub 2020/10/25; PubMed PMID: 33097004; PMCID: PMC7585212.
- Dumitrascu B, Villar S, Mixon DG, Engelhardt BE. Optimal marker gene selection for cell type discrimination in single cell analyses. Nat Commun. 2021;12(1):1186. https://doi.org/10.1038/s41467-021-21453-4.
- Gregory W, Sarwar N, Kevrekidis G, Villar S, Dumitrascu B. MarkerMap: nonlinear marker selection for single-cell studies. NPJ Syst Biol Appl. 2024;10(1):17. https://doi.org/10.1038/s41540-024-00339-3. Epub 2024/02/14; PubMed PMID: 38351188; PMCID: PMC10864304.
- Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, Griesbeck M, Butler A, Zheng S, Lazo S, Jardine L, Dixon D, Stephenson E, Nilsson E, Grundberg I, McDonald D, Filby A, Li W, De Jager PL, Rozenblatt-Rosen O, Lane AA, Haniffa M, Regev A, Hacohen N. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356(6335):eaah4573. https://doi.org/10.1126/science.aah4573.
- Qiu Y, Wang J, Lei J, Roeder K. Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. Bioinformatics. 2021;37(19):3228–34. https://doi.org/10.1093/bioinformatics/btab257. Epub 2021/04/28; PubMed PMID: 33904573; PMCID: PMC8504631.
- Su C, Xu Z, Shan X, Cai B, Zhao H, Zhang J. Cell-type-specific co-expression inference from single cell RNA-sequencing data. Nat Commun. 2023;14(1):4846. https://doi.org/10.1038/s41467-023-40503-7. Epub 2023/08/11; PubMed PMID: 37563115; PMCID: PMC10415381.
- Zhang Y, Miller JA, Park J, Lelieveldt BP, Long B, Abdelaal T, Aevermann BD, Biancalani T, Comiter C, Dzyubachyk O, Eggermont J, Langseth CM, Petukhov V, Scalia G, Vaishnav ED, Zhao Y, Lein ES, Scheuermann RH. Referencebased cell type matching of in situ image-based spatial transcriptomics data on primary visual cortex of mouse brain. Sci Rep. 2023;13(1):9567. https://doi.org/10.1038/s41598-023-36638-8. Epub 2023/06/14; PubMed PMID: 37311768; PMCID: PMC10264402.
- 30. Hawrylycz M, Martone ME, Ascoli GA, Bjaalie JG, Dong HW, Ghosh SS, Gillis J, Hertzano R, Haynor DR, Hof PR, Kim Y, Lein E, Liu Y, Miller JA, Mitra PP, Mukamel E, Ng L, Osumi-Sutherland D, Peng H, Ray PL, Sanchez R, Regev A, Ropelewski A, Scheuermann RH, Tan SZK, Thompson CL, Tickle T, Tilgner H, Varghese M, Wester B, White O, Zeng H, Aevermann B, Allemang D, Ament S, Athey TL, Baker C, Baker KS, Baker PM, Bandrowski A, Banerjee S, Bishwakarma P, Carr A, Chen M, Choudhury R, Cool J, Creasy H, D'Orazi F, Degatano K, Dichter B, Ding SL, Dolbeare T, Ecker JR, Fang R, Fillion-Robin JC, Fliss TP, Gee J, Gillespie T, Gouwens N, Zhang GQ, Halchenko YO, Harris NL, Herb BR, Hintiryan H, Hood G, Horvath S, Huo B, Jarecka D, Jiang S, Khajouei F, Kiernan EA, Kir H, Kruse L, Lee C, Lelieveldt B, Li Y, Liu H, Liu L, Markuhar A, Mathews J, Mathews KL, Mezias C, Miller MI, Mollenkopf T, Mufti S, Mungall CJ, Orvis J, Puchades MA, Qu L, Receveur JP, Ren B, Sjoquist N, Staats B, Tward D, van Velthoven CTJ, Wang Q, Xie F, Xu H, Yao Z, Yun Z, Zhang YR, Zheng WJ, Zingg B. A guide to the BRAIN Initiative Cell Census Network data ecosystem. PLoS Biol. 2023;21(6):e3002133. https://doi.org/10.1371/journal.pbio.3002133. Epub 2023/06/30; PubMed PMID: 37390046; PMCID: PMC10313015.

- Tan SZK, Kir H, Aevermann BD, Gillespie T, Harris N, Hawrylycz MJ, Jorstad NL, Lein ES, Matentzoglu N, Miller JA, Mollenkopf TS, Mungall CJ, Ray PL, Sanchez REA, Staats B, Vermillion J, Yadav A, Zhang Y, Scheuermann RH, Osumi-Sutherland D. Brain Data Standards - A method for building datadriven cell-type ontologies. Sci Data. 2023;10(1):50. https://doi.org/10.1038/ s41597-022-01886-2. Epub 2023/01/25; PubMed PMID: 36693887; PMCID: PMC9873614.
- Sievert C. Interactive web-based data visualization with R, plotly, and shiny. 1st edition. New York: Chapman and Hall/CRC; 2020. https://doi.org/10. 1201/9780429447273.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.