Discovery of cell type classification marker genes from single cell RNA sequencing data using NS-Forest

Angela Liu[1], Beverly Peng[1], Ajith V. Pankajam[2], Richard H. Scheuermann[2], Yun Zhang[1, #]

1. Department of Informatics, J. Craig Venter Institute, La Jolla, CA, United States of America.
2. National Library of Medicine, National Institutes of Health, Bethesda, MD, United States of America
Correspondence: zhangy@jcvi.org

Abstract

The use of single-cell transcriptomic technologies that quantitively describe cell transcriptional phenotypes using single cell/nucleus RNA sequencing (scRNA-seq) 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. 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 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. By modularizing the final decision tree step, NS-Forest v4.0 can be used to compare the performance of user-defined 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 to1, with a metric of 1 given to markers that are only expressed within their target cell types and not in cells of any other cell types. We have applied NS-Forest v4.0 on scRNA-seq datasets from three human organs, including the brain, kidney, and lung. We observe that NS-Forest v4.0 outperforms previous versions on its ability to identify markers with higher On-Target Fraction values for closely related cell types and outperforms other marker gene selection approaches on the classification performance with significantly higher F-beta scores.

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 their 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 BioMolecular 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 BRAIN Initiative 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 and reviewed 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 granular cell types. Alternatively, DE genes selected by a modified Wilcoxon rank sum test (performed using the "presto" R package) is used in the popular Azimuth web-based reference[12] 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 a "cell type classification marker gene combinations" for scRNA-seq data, which must meet the following criteria: 1) each gene 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), 3) gene combinations optimized for cell type classification using metrics that quantify classification confidence, and 5) can be reproduced by a generalizable method.

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 applications.

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 gene features (marker candidate) for each cell type ranked by Gini index calculated in the random forest model. (Figure 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 then 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 all individual cells of the target cell type and not in cells of any other cell type. Version 2.0 uses this Binary Expression Score as an additional ranking step to preferentially select genes with this 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 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 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 step to the feature selection process 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, allowing for the preferential selection of informative binary markers during the iterative random forest process, resulting in a more distinct 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 (Figure 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 (Figure 2). In version 4.0, four options for the threshold used in the BinaryFirst step were implemented: '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. In this case, the algorithm is the same as version 3.9. The other threshold values correspond to mild, moderate, and high criteria that are calculated based on the distribution of Binary Expression Scores from all genes, and so these values vary depending on the input dataset, controlling for gene expression distribution variability arising from many factors, including the organ/tissue type, sample pre-processing, sequencing platform, etc. As the thresholding values increase, the number of selected genes decrease. The mild threshold is set as the median Binary Expression Score; the moderate threshold is set as the mean Binary Expression Score plus one standard deviation of all scores; the high threshold is set as the mean Binary Expression Score plus two times the standard deviation. In version 4.0, the default threshold is set to 'BinaryFirst_high,'

which is the most stringent threshold that filters the most genes, based on the performance evaluation described below.

The BinaryFirst strategy is designed to reduce the feature space that the random forest classifier must search over in the following workflow step of NS-Forest, with the goal of improving the algorithm's ability to identify marker genes for clusters that have previously been defined as difficult to distinguish. There exists multiple groups of closely-related cell types in the human MTG dataset[5]. The input datasets used for marker gene selection typically have tens of thousands of genes, but the majority of these will not be useful as the marker genes for each cluster in the dataset. 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. In previous versions of NS-Forest, this large number of genes in the input datasets corresponded to the large total number of input features that were considered for random selection at each decision tree node, and this was the case for all decision trees that made up the random forest model. Our hypothesis was that by reducing the size of the set of gene candidates, BinaryFirst would be able to simplify the task of the random forest classifier, while simultaneously reducing the overall runtime of NS-Forest. 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]. The BinaryFirst approach was designed to enrich for the most useful classification features prior to random selection. In this sense, BinaryFirst can be viewed as an informative method of dimensionality reduction approach that retains and evaluates genes that are most uniquely expressed in their corresponding target cluster.

We note that often the median Binary Expression Score in a scRNA-seq dataset is 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. For the analyses presented in this paper, we will treat the results obtained from running version 4.0 with the 'BinaryFirst_mild' threshold as 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.

Improved marker gene selection on the human brain dataset

The performance of NS-Forest v4.0 was initially assessed on the same human middle temporal gyrus (MTG) brain dataset that was used to evaluate previous versions of NS-Forest. Figure 3 shows heatmaps of marker gene expression across different versions of NS-Forest. The heatmaps in Figure 3A visualize the median expression of each marker gene selected by the three different versions of NS-Forest in the 75 clusters in the human MTG dataset. The horizontal axis lists the 75 clusters (in the order corresponding to the dendrogram in Figure 1c from the MTG study[5]) and the marker genes corresponding to the clusters in the same order are listed on the vertical axis. In the ideal scenario where there exist marker genes that are exclusively expressed for each cluster, expression would only be observed along the diagonal axis. In these heatmaps, we observe a noticeable decrease in the amount of off-target expression across the three heatmaps from versions 1.3 to 2.0/3.9 to 4.0, as the amount of off-target expression of each marker gene is generally decreased. The median gene expression of each marker in their associated target clusters also increases in version 4.0, as we observe more red and yellow squares on the diagonal axis. Comparing versions 2.0/3.9 and 4.0, we note an increase in the median PPV across all 75 clusters as well as a

significant increase in the median On-Target Fraction (Figure 3B). The improvement in PPV (precision) means that version 4.0 of NS-Forest does a better job at identifying 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 median F-beta score between the two versions due to slightly lower recall. However, since the false negative component of recall is a combination of misclassification and the dropout technical artifact, we view the improvement in PPV to be a better measure of method improvement. The boxplots in Figure 3C display the difference in the distribution of the On-Target Fraction values for each of the 75 cell types in this dataset, and they showcase the significant difference between the distributions in v2.0/3.9 versus v4.0 (mild versus high BinaryFirst threshold on the x-axis). It is also interesting to note that the p-values for all three paired t-tests comparing the On-Target Fraction values between each possible pair of BinaryFirst thresholds on the human MTG dataset are statistically significant (mild vs. moderate p-value = 0.01, moderate vs. high p-value = 1.79e-04, mild vs. high p-value = 1.68e-06). This suggests that the filtering done in the BinaryFirst step is very effective in helping NS-Forest better identify markers that are more exclusively expressed at higher levels in their target cell types for this brain dataset.

Tradeoffs in the F-beta score when compared to markers ranked strictly by feature importance (Gini index) from the off-the-shelf random forest model in NS-Forest version 1.3 without enriching for the binary expression pattern were previously noted[13]. When comparing v1.3 and v2.0, Aevermann et al. demonstrated that the genes selected by v2.0 with a high Binary Expression Score were usually not those that are ranked highest by Gini index in random forest models, leading to a slightly lower median F-beta score. However, those genes selected in v2.0 had significantly higher Binary Expression Scores, making them more useful for downstream assays such as RT-PCR and spatial transcriptomics. We observed a similar trend in this current comparison of versions 2.0/3.9 and version 4.0: the median F-beta score for v4.0 markers is slightly lower than for v2.0/3.9 markers, but the Binary Expression Scores are higher for v4.0 than for v2.0/3.9, with an average of 0.971compared to 0.936. This suggests that version 4.0 of NS-Forest selects marker genes with greater Binary Expression Scores by providing a shorter list of filtered gene candidates from the filtering done by the BinaryFirst step to the random forest classifier, resulting in a decreased likelihood that the machine learning algorithm becomes trapped in the local optima only quantified by the Gini index. Interestingly, the average Gini index values for the v4.0 markers also seems to be greater than that for the v2.0/3.9 markers (0.024 for v4.0 versus 0.008 for v2.0/3.9). This suggests that in this comparison of v2.0/3.9 and v4.0, there is greater correlation between genes with a high Binary Expression Score and those that are ranked highest by Gini index.

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 (v4.0 results in Supplementary Table 1). 85 of the total markers (~51% of the markers identified by v2.0/3.9) were identified for the same cluster for both sets.

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

The distribution of On-Target Fractions for each of the three subclades is consistent with these visual patterns (Supplementary Figure 2). The p-value for comparing the mild and high BinaryFirst thresholds for the VIP subclade is statistically significant (p-value = 0.01), however, the p-values for the L4 subclade and the PVALB subclade are not significant (p-value = 0.37 and 0.07, respectively), likely due to the small sample size for the t-test (the L4 subclade has 5 distinct cell types and the PVALB subclade has 6 whereas the VIP subclade has 21).

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 NS-Forest v4.0 performance improvement. 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 (Figure 4A), which matches the observed high On-Target Fraction values for these datasets (Figure 4B). These values are higher than 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 (Figure 4B), the median F-beta scores are very similar in both versions for each organ. In the kidney dataset, we observe a slight improvement in PPV and a slight decrease in On-Target Fraction 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 positives classifications is much fewer in version 4.0 (dropped from an average of 312.5 in version 2.0/3.9 to 198 in version 4.0), indicating that the version 4.0 marker genes are better markers for classification in general. In the lung L5 subclass dataset, we observe increases in both PPV and On-Target Fraction in 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 distinguish between the 75 cell types (v4.0 results in Supplementary Tables 2). 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 for the lung L5 dataset (v4.0 results in Supplementary Tables 3), 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 Figure 3 and Supplementary Tables 4-5), suggesting that it is easier to select marker genes at less granular levels.

Improvement in runtime in version 4.0

In addition to the improvements in the classification performance, NS-Forest v4.0 also significantly reduces the overall runtime of the algorithm compared to versions 2.0/3.9 (Table 1). We observe that the overall runtime is much lower in v4.0 of NS-Forest (by default, 'BinaryFirst_high' is used) as compared to that of v2.0/3.9 in all three human organ datasets, with the ratio of runtime (v4.0 to v2.0/3.9) ranging from 0.10 to 0.26 across the datasets. All runtimes were obtained from running NS-Forest through jobs that were submitted to the *Expanse* supercomputer system in the San Diego SuperComputer Center. 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 (Supplementary Figure 4). We note that for all three datasets, there is a large decrease in runtime going from v2.0/3.9 to 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 high threshold was used, 1-7% of the total original genes passed the BinaryFirst threshold in the three datasets. This demonstrates 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 Figure 4), 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 while providing improvement of 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 resources 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]). To compare the performance of these different approaches, the Human Lung Cell Atlas (HLCA) core dataset of ~0.5 million cells divided into 61 cell types from healthy lung tissue from 107 individuals was used[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). For this comparison, we consider four marker gene lists for healthy lung cell types: NS-Forest markers, HLCA markers, Azimuth markers, and ASCT+B markers. The HLCA markers are available in Supplementary Table 6 from the publication[18]. The Azimuth markers derived from the same dataset are pre-calculated and available at the Azimuth portal (https://azimuth.hubmapconsortium.org/). The ASCT+B markers (lung v1.4) are available at the Human Reference Atlas portal (https://humanatlas.io/asctb-tables). For ASCT+B cell types, Cell Ontology[19] IDs were used to map to HLCA cell types, resulting in 18 exact matches, following the Simple Standard for Sharing Ontological Mappings (SSSOM) guideline[20].

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 dotplot of the NS-Forest marker genes, where cell types are in the rows and marker genes are in the columns. The 61 cell types are ordered according to the dendrogram in Figure 5B generated by the scanpy dendrogram function[15] using all genes and default settings. In the dendrogram, similar cell types are grouped according to the hierarchical clustering of the transcriptome profiles of these cell types; three major branches 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 of the dotplot, with few off-diagonal expressions (Figure 5A). Similar dotplots were produced for the other

three marker gene lists (Supplementary Figure 5). The HLCA authors provided 165 marker genes (1-5 markers per type) to characterize the 61 cell types, which also show a diagonal pattern in the dotplot but with more off-diagonal expressions (Supplementary Figure 5A). The Azimuth marker list contains 560 marker genes (10 markers per type) for 56 of these cell types (AT0, Hematopoietic stem cells, Hillock-like, Smooth muscle FAM83D+, 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 Figure 5B). 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 Figure 5C) because the manual approach based on existing knowledge/literature does not capture the granularity obtained in single cell-resolution data.

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 four marker gene sets was compared using the HLCA data, calculating F-beta, precision, and recall (Figure 5C) for the 18 cell types matched across all four references (Figure 5D-F) and for all identified cell types (Figure 5G-I). For the fair comparison of the 18 common cell types, the NS-Forest marker genes have the highest median F-beta score (0.78), followed by the HLCA markers (0.57), Azimuth markers (0.17), and ASCT+B markers (0.17). Similar results were found using all cell types (NS-Forest: 0.71, HLCA: 0.48, Azimuth: 0.17, ASCT+B: 0.17). The F-beta score reflects the global patterns observed in the dotplots (Figure 5A). The distribution of F-beta scores are significantly different between NS-Forest and all other methods (NS-Forest vs. HLCA: $p = 0.001$, NS-Forest vs. Azimuth: $p = 1.8e-7$, and NS-Forest vs. ASCT+B: $p = 2.2e-8$). Surprisingly, median PPV (precision) scores are high for all four methods (0.82-0.94), with the ASCT+B being the highest, showing no significant difference between NS-Forest and the other methods (NS-Forest vs. HLCA: $p = 0.97$, NS-Forest vs. Azimuth: $p = 0.21$, NS-Forest vs. ASCT+B: $p = 0.85$). In contrast, the median recall value show a similar trend to the F-beta scores, with NS-Forest being the highest (0.48) and ASCT+B being the lowest (0.05) in the 18 common cell types. The distribution of recall scores are significantly different between NS-Forest and all other methods (NS-Forest vs. HLCA: $p = 0.01$, NS-Forest vs. Azimuth: $p = 1.5e-7$, and NS-Forest vs. ASCT+B: $p = 1.2e-7$). Scatter plots for pairwise direct 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 Figure 6), while having comparable performance in PPV (precision). Low F-beta and recall values for the other three methods is driven by excessive false negative predictions. Comparing all methods, NS-Forest produces the most comprehensive and concise list of marker genes with dominantly higher F-beta scores for cell type classification.

Discussion

This paper describes major algorithmic refinements made in NS-Forest v4.0 and its performance on the human brain MTG dataset used to develop the previous versions, as well as 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 reduced substantially. 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, we found that this new version demonstrated clear improvement in the human MTG dataset, as the brain is the most complex organ that we evaluated in this study. Additional datasets representing the human kidney and lung were used to validate NS-Forest's performance in 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 marker lists from popular resources. We formally established the notion of a cell type classification marker gene, which is different from the notion of differentially expression gene. In the head-to-head comparison using the HLCA dataset with half a million cells, the NS-Forest marker genes showed superior performance than the HLCA, Azimuth, and ASCT+B marker lists.

NS-Forest marker genes can be used for multiple downstream experimental investigations 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 that can later be adopted into the official Cell Ontology[19], as NS-Forest provides the minimum combinations of marker genes that can serve as a set of definitional characteristics of the cell types[19]. Such efforts have already begun, and NS-Forest has contributed to the BRAIN Initiative Cell Census Network (BICCN) data ecosystem to derive the necessary and sufficient marker gene knowledge[21]. As such, the Provisional Cell Ontology (PCL) is generated in this manner for human, mouse, and marmoset primary motor cortex[22]. 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 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 (Figure 2). The Binary Expression Scores are first calculated for each gene-cluster pair in the dataset, and this score is calculated using the formula defined in the paper detailing version 2.0[13]. When running NS-Forest, 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. If 'none' is used, the algorithm is the same as version 3.9. The other threshold values correspond to mild, moderate, and high criteria that 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.

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, implying that using the mild threshold is the same as version 3.9. 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, respectively.

On-Target Fraction metric

In version 4.0, we provide a new metric, the On-Target Fraction, 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, thus prompting the creation of the On-Target Fraction metric. To make a clear distinction, we refer to F-beta score and PPV as *classification* metrics, and On-Target Fraction as an *expression* metric.

The On-Target Fraction is defined as, for each marker gene g in target cluster T,

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

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 $i = 1, \ldots, n$. 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 in practice. At the cluster level, we report the On-Target Fraction for each cluster using the median $\text{Fraction}_{\text{gT}}$ 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 data.

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, the NS-Forest v4.0 package is now modularized, and it consists 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 the 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 amount 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 threshold (None: 0, BinaryFirst_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. 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 F-beta score, PPV, TP, FP, TN, FN, and On-Target Fraction.

The evaluating module can be performed independently without calling the NSForesting module. It 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 heatmap figures for visualization of the NS-Forest 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[23] boxplots and scatter plots, which are useful for comparing metrics and identifying clusters of interest.

Data Availability

The human brain Middle Temporal Gyrus (MTG) dataset is downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq. The human kidney dataset is 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 unpublished (manuscript under submission). The Human Lung Cell Atlas (HLCA) dataset is downloaded from https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293 (core).

Code Availability

NS-Forest v4.0 is openly available at https://github.com/JCVenterInstitute/NSForest, with documentation and example code in the tutorials.

Acknowledgments

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

Author 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. A.L., 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.

Competing interests

None.

<table>
<tr><th>Workflow step</th><th>NS-Forest v1.3</th><th>NS-Forest v2.0</th><th>NS-Forest v3.9</th><th>NS-Forest v4.0</th></tr>
<tr><td>Input data format</td><td colspan="2">Cell-by-gene expression matrix (log-transformed/normalized) and cell cluster labels</td><td colspan="2">Annotated data (anndata) object in h5ad format</td></tr>
<tr><td>Gene pre-selection</td><td>None</td><td colspan="2">Select positive genes (i.e., median gene expression > 0 in at least one cluster)</td><td>a. Select positive genes
b. Select by Binary Scoring Criterion (“BinaryFirst”)</td></tr>
<tr><td>Random Forest feature selection</td><td>Rank gene features by Gini Impurity</td><td colspan="3">a. Rank gene features by Gini Impurity
b. Rank gene features by Binary Score</td></tr>
<tr><td>Feature evaluation</td><td>Calculate median gene expression in the target cluster for each top-ranked gene feature</td><td colspan="3">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</td></tr>
<tr><td>Combinatorial marker gene determination</td><td>Calculate the F1 scores for stepwise addition of the top-ranked gene features and find the maximum F1 score</td><td colspan="3">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</td></tr>
<tr><td>Performance metrics</td><td>F1 score</td><td>F-beta score</td><td>a. F-beta score
b. Positive Predictive Value (PPV)</td><td>a. F-beta score
b. PPV
c. On-Target Fraction</td></tr>
</table>

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

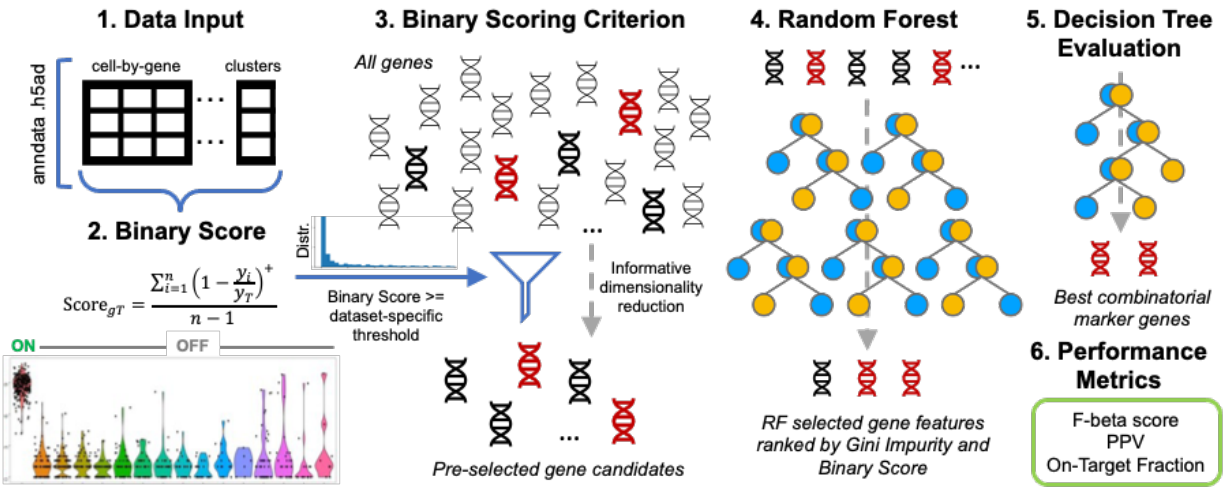
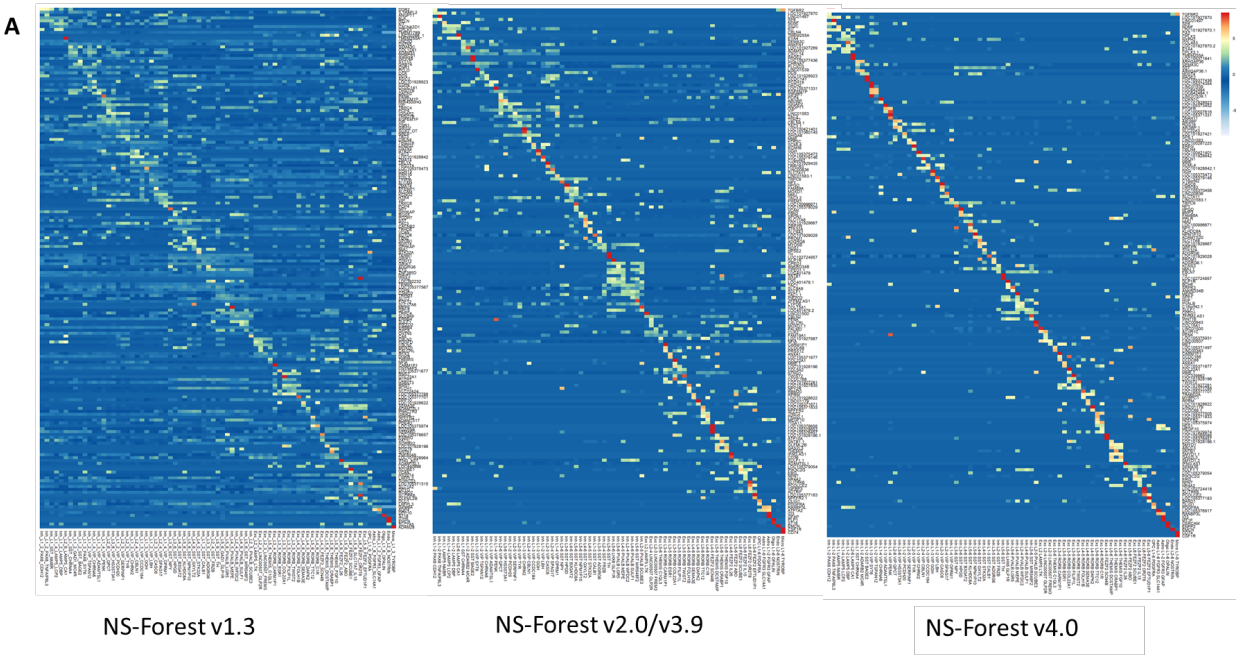


Figure 2. NS-Forest version 4.0 workflow. The algorithm starts with an anndata object in .h5ad format, containing the cell-by-gene expression matrix and cluster labels for each cell as data input (step 1). Based on the input data, the algorithm pre-calculates the median gene expression for each gene in each cluster (i.e., a cluster-by-gene median matrix) and pre-selects genes that have positive median expression in at least one cluster (not shown in this figure). The Binary Expression Score is then calculated for each cluster-gene pair (step 2), i.e., a cluster-by-gene Binary Score matrix is produced (note that a gene may have different values of Binary Score in different clusters), and the dataset-specific threshold is calculated based on the Binary Score distribution and user's selection of mild, moderate, or high criterion. This threshold value is used to filter the genes that are considered as candidate genes for each cluster in the next step, as genes with a Binary Expression Score greater than or equal to the threshold are selected (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 important for RF classification 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), and On-Target Fraction are reported for each cluster, serving as metrics for evaluating the performance of the final maker gene combinations (step 6).



B

	NS-Forest v1.3	NS-Forest v2.0/v3.9	NS-Forest v4.0
Median F-beta	0.63*	0.702	0.633
Median PPV	NA	0.857	**0.880**
Median On-Target Fraction	NA	0.280	0.399

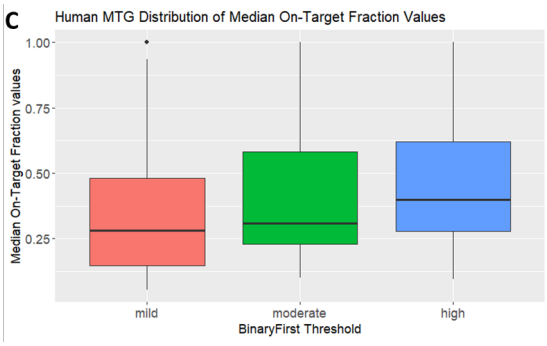
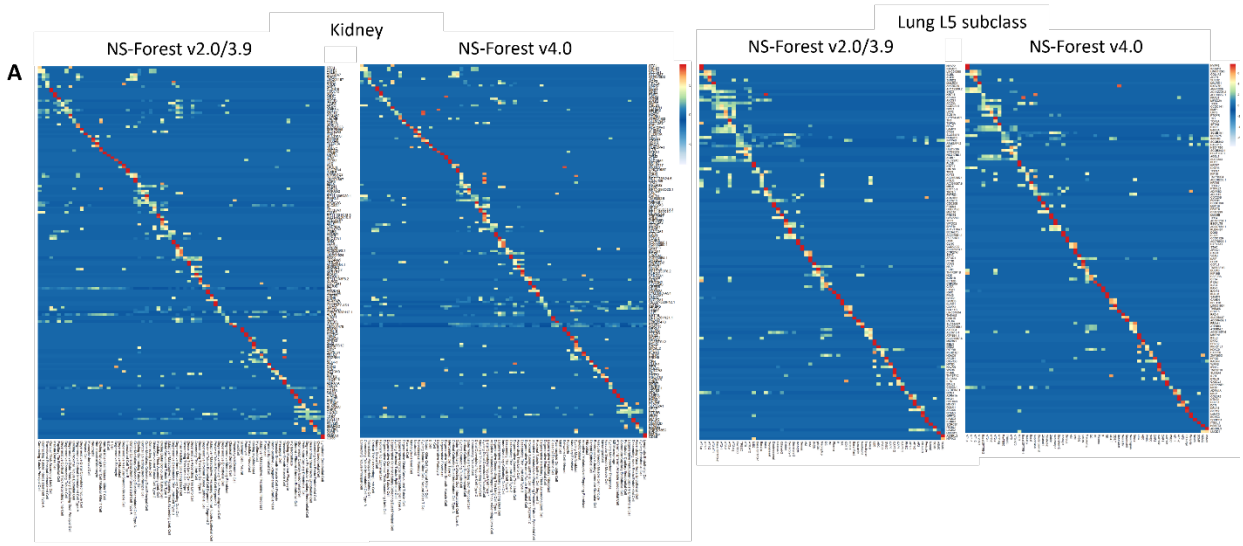


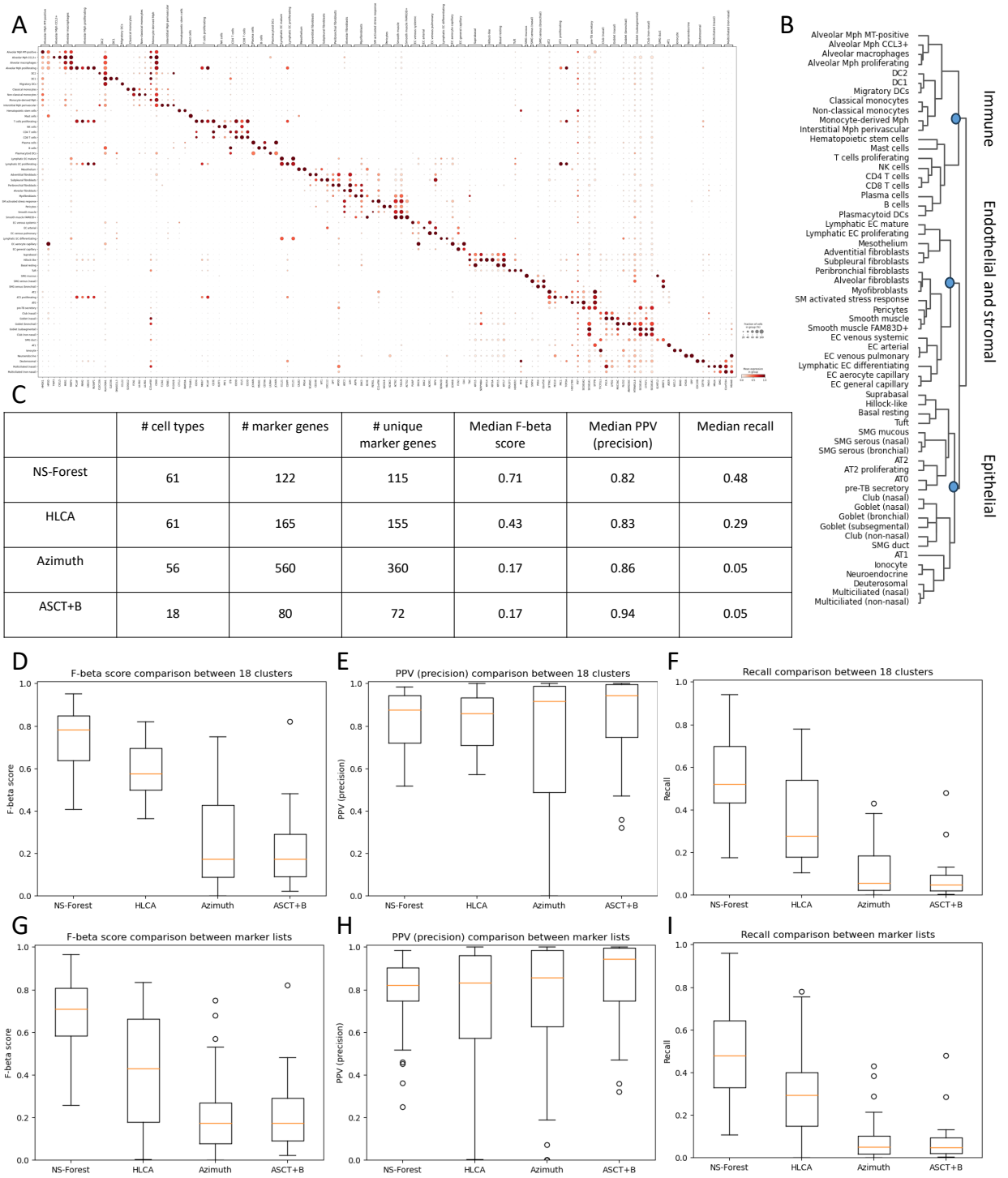
Figure 3. Comparing performance of previous versions of NS-Forest to version 4.0 on human MTG dataset. (A) Heatmaps of NS-Forest v1.3, v2.0/v3.9, and v4.0 marker genes for the 75 cell type clusters from human middle temporal gyrus. The colors correspond to the normalized median expression level (log2-transformed 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 (Figure 1c in Hodge et al. 2019). **(B)** Comparison of the performance metrics of the corresponding versions of NS-Forest results shown in (A). *Note that the unweighted F1 score is used in v1.3. **(C)** Boxplots displaying the distribution in the median On-Target Fraction values across the 75 clusters in the human MTG dataset across the results from running NS-Forest v4.0 with the three possible BinaryFirst configurations.



B

	Kidney		Lung L5 subclass	
NS-Forest version	*NS-Forest v2.0/3.9*	*NS-Forest v4.0*	*NS-Forest v2.0/3.9*	*NS-Forest v4.0*
Median F-beta	0.555	0.559	0.630	0.630
Median PPV	0.708	0.709	0.767	0.794
Median On-Target Fraction	0.640	0.607	0.592	0.669

Figure 4. Evaluating performance of NS-Forest 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).

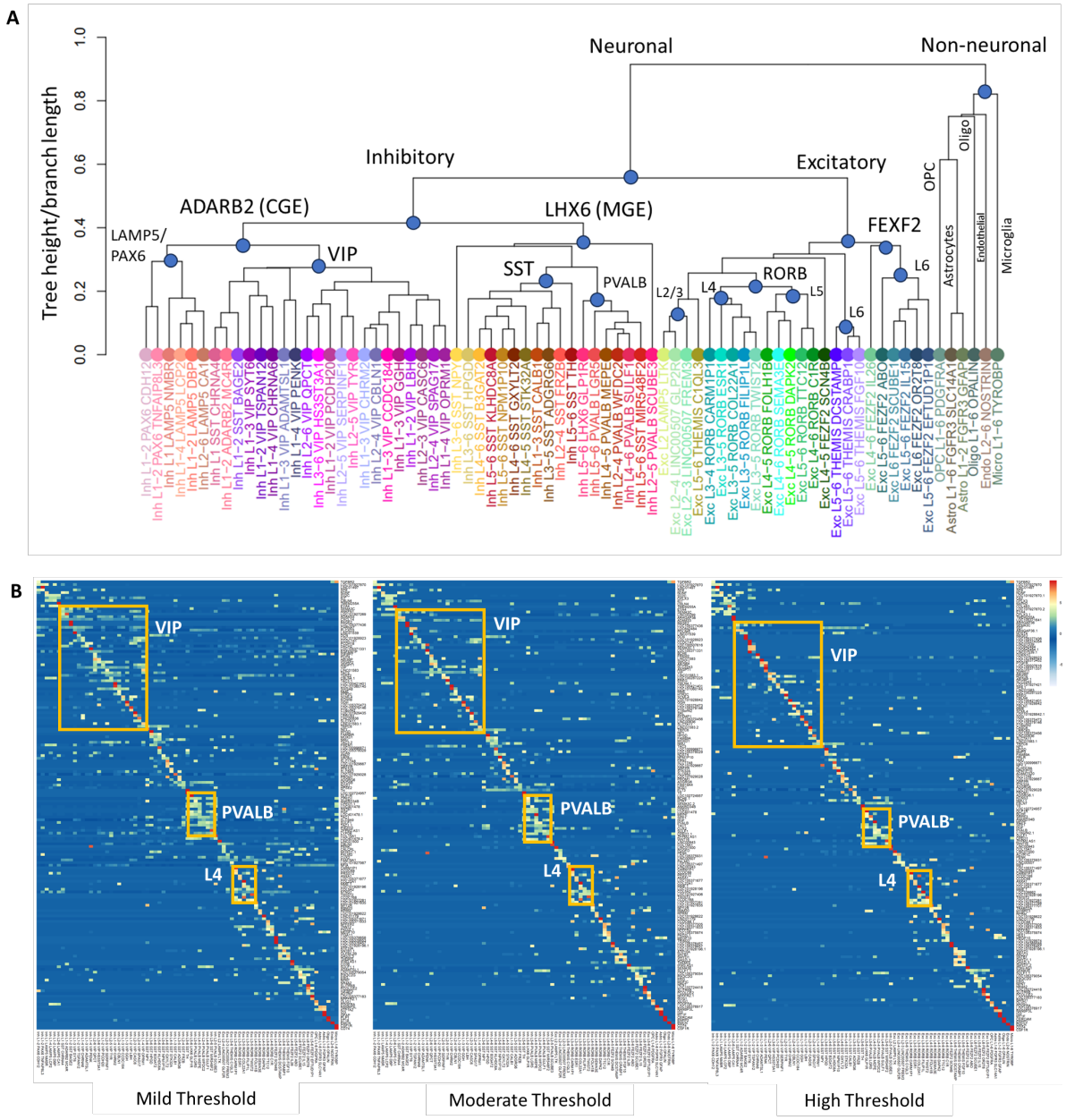


	# cell types	# marker genes	# unique marker genes	Median F-beta score	Median PPV (precision)	Median recall
NS-Forest	61	122	115	0.71	0.82	0.48
HLCA	61	165	155	0.43	0.83	0.29
Azimuth	56	560	360	0.17	0.86	0.05
ASCT+B	18	80	72	0.17	0.94	0.05

Figure 5. NS-Forest marker genes compared to other published HLCA marker gene lists. (A) Dotplot of the 122 NS-Forest marker genes on HLCA core cell types. **(B)** Dendrogram of 61 ann_finest_level from HLCA core dataset. Representing the rows in (A). Generated before preprocessing. **(C)** Comparing number of cell types, number of markers, number of unique markers, median f-beta score, median PPV, and median recall for the NS-Forest, HLCA, Azimuth, and ASCT+B marker lists. **(D-F)** Boxplots of F-beta score, PPV (precision), and recall for the 18 cell types commonly characterized by all methods. **(G-I)** Boxplots of F-beta score, PPV (precision), and recall for all available cell types across methods.

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)	05:13:02	1	**1** (13945/13945)
	BinaryFirst: moderate (mean + 1 std. dev.)	01:44:54	0.335	**0.164** (2292.47/13945)
	BinaryFirst: high (mean + 2 std. dev.)	01:00:38	0.194	**0.073** (1022.99/13945)
Kidney	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	68:17:37	1	**1** (33920/33920)
	BinaryFirst: moderate (mean + 1 std. dev.)	22:14:28	0.326	**0.013** (447.79/33920)
	BinaryFirst: high (mean + 2 std. dev.)	06:53:04	0.101	**0.013** (446.39/33920)
Lung: L5	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	02:37:58	1	**1** (29800/29800)
	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.

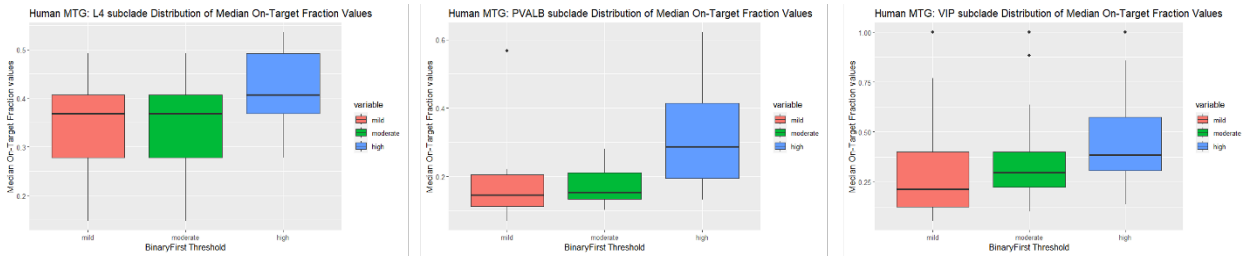


C

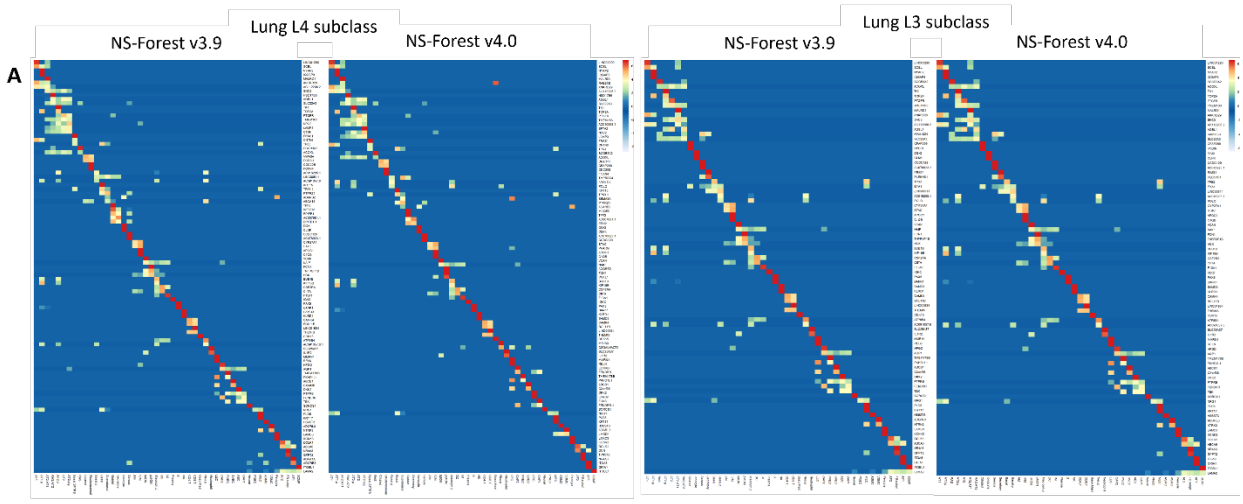
Subclade	Approach	Median On-Target Fraction
L4	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	0.368
	BinaryFirst: moderate (mean + 1 std. dev.)	0.368
	BinaryFirst: high (mean + 2 std. dev.)	0.406
PVALB	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	0.146
	BinaryFirst: moderate (mean + 1 std. dev.)	0.153
	BinaryFirst: high (mean + 2 std. dev.)	0.286
VIP	NS-Forest v2.0/3.9/BinaryFirst: mild (median)	0.213
	BinaryFirst: moderate (mean + 1 std. dev.)	0.296
	BinaryFirst: high (mean + 2 std. dev.)	0.385

Supplementary Figure 1. Comparing performance of using different BinaryFirst thresholds in NS-Forest version 4.0 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 v2.0/v3.9 (equivalent to v4.0 with the 'BinaryFirst_mild' threshold), v4.0 with the 'BinaryFirst_moderate' threshold, and v4.0 with the 'BinaryFirst_high' threshold. 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 specific versions of NS-Forest.



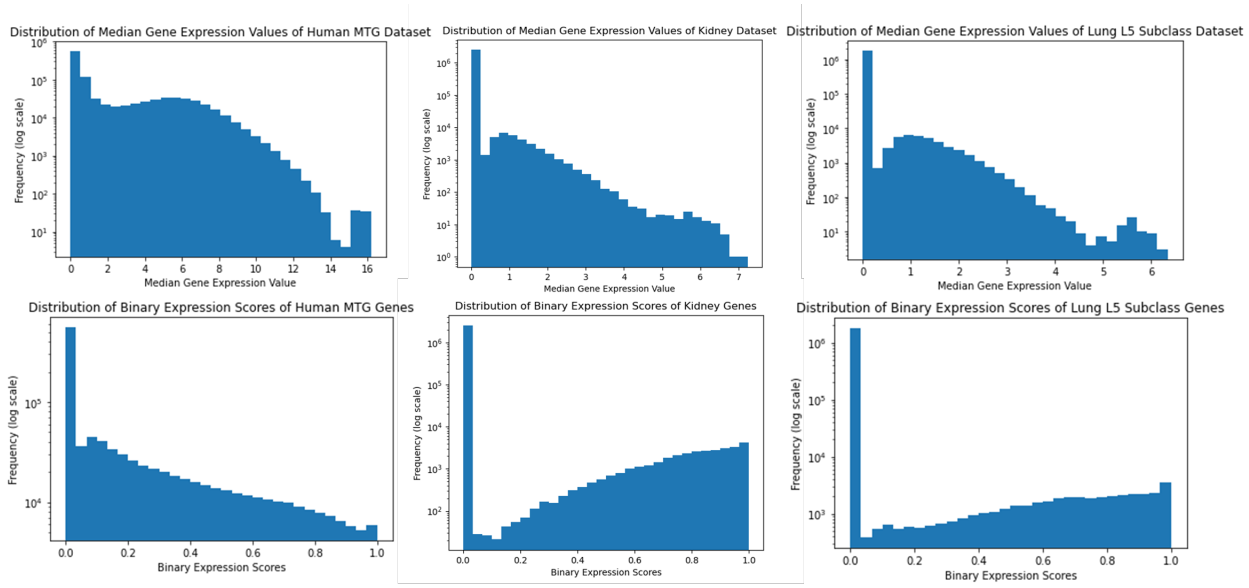
Supplementary Figure 2. Comparison of median On-Target Fraction distribution in major subclades of Human MTG dataset across BinaryFirst thresholds. These three plots display boxplots comparing the distribution of the median On-Target Fraction values within each of the three specific subclades in the human MTG dataset (L4, PVALB, and VIP subclades) across the mild, moderate, and high BinaryFirst thresholds. The boxplots demonstrate how the distribution of On-Target Fraction values in the results where the 'BinaryFirst_high' threshold is used (blue boxplots) is clearly higher than that for the 'BinaryFirst_mild' and 'BinaryFirst_moderate' thresholds (red and green boxplots).



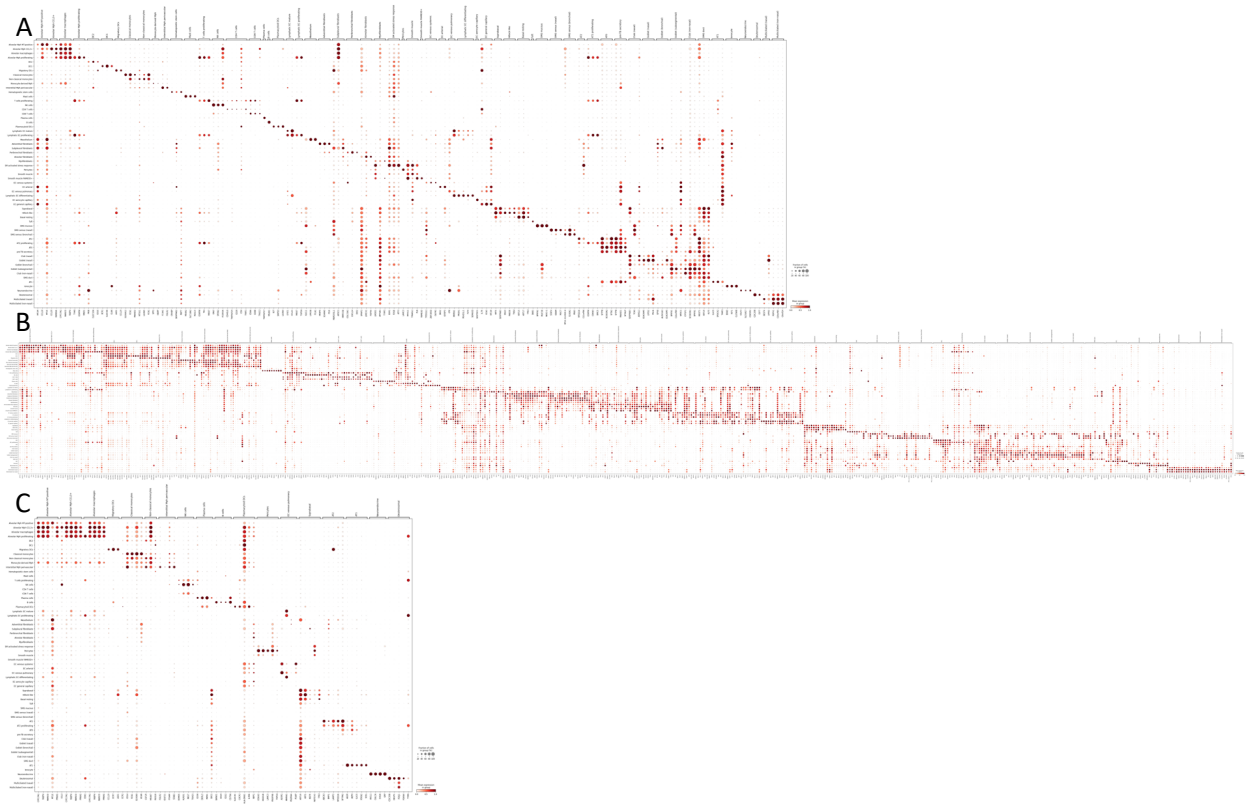
B

	Lung L4 subclass		Lung L3 subclass	
NS-Forest version	*NS-Forest v3.9*	*NS-Forest v4.0*	*NS-Forest v3.9*	*NS-Forest v4.0*
Median f-beta	0.676	0.649	0.694	0.692
Median PPV	0.859	0.848	0.877	0.873
Median On-Target Fraction	0.637	0.746	0.749	0.768

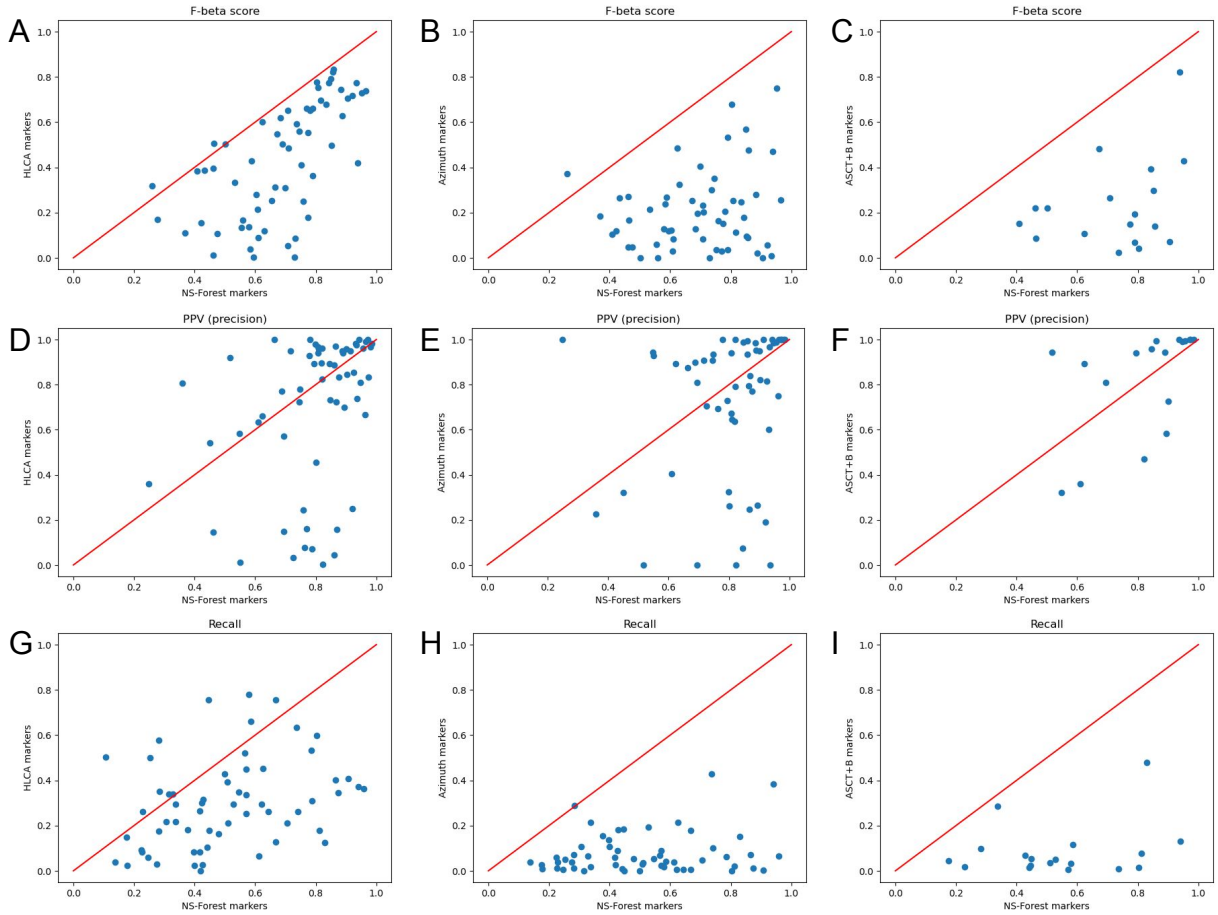
Supplementary Figure 3. Evaluating performance of NS-Forest on lung L4 and L3 subclass datasets. (A) Heatmaps of NS-Forest v2.0/v3.9 and v4.0 markers from the L4 and L3 subclasses of the human lung. The colors correspond to the normalized median expression level (log2-transformed counts per million) 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 4. 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 on the logarithmic 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 on the logarithmic scale. All distributions are highly right-skewed.



Supplementary Figure 5. Dotplots of HLCA, Azimuth, and ASCT+B marker genes on HLCA core cell types. (A) HLCA with 165 markers across 61 cell types. **(B)** Azimuth with 560 markers across 56 cell types. **(C)** ASCT+B with 80 markers across 18 cell types.



Supplementary Figure 6. NS-Forest consistently outperforms other published methods in classification metrics. (A-C) Scatter plots comparing F-beta scores for each cell type of NS-Forest markers vs. HLCA, Azimuth, and ASCT+B. **(D-F)** Scatter plots comparing PPV (precision) for each cell type of NS-Forest markers vs. HLCA, Azimuth, and ASCT+B. **(G-I)** Scatter plots comparing recall for each cell type of NS-Forest markers vs. HLCA, Azimuth, and ASCT+B.

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.

References

1. 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. Epub 2018/01/13. doi: 10.1186/s12859-017-1977-1. PubMed PMID: 29322913; PMCID: PMC5763450.
2. 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. Epub 2017/12/06. doi: 10.7554/eLife.27041. PubMed PMID: 29206104; PMCID: PMC5762154.
3. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. 2019;574(7777):187-92. Epub 2019/10/11. doi: 10.1038/s41586-019-1629-x. PubMed PMID: 31597973; PMCID: PMC6800388.
4. Mott MC, Gordon JA, Koroshetz WJ. The NIH BRAIN Initiative: Advancing neurotechnologies, integrating disciplines. PLoS Biol. 2018;16(11):e3000066. Epub 2018/11/27. doi: 10.1371/journal.pbio.3000066. PubMed PMID: 30475794; PMCID: PMC6283590.
5. 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. Epub 2019/08/23. doi: 10.1038/s41586-019-1506-7. PubMed PMID: 31435019; PMCID: PMC6919571.
6. 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. Epub 2023/10/12. doi: 10.1126/science.add7046. PubMed PMID: 37824663.
7. 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.
8. Duong TE. Single-nucleus RNA-seq data for the Lung Airways and Parenchymal Map. Under submission. 2024.
9. 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. Epub 2023/07/20. doi: 10.1038/s41556-023-01194-w. PubMed PMID: 37468756; PMCID: PMC10681365.

10. 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. Epub 2021/11/10. doi: 10.1038/s41556-021-00788-6. PubMed PMID: 34750582; PMCID: PMC10079270.
11. Boerner K, Blood PD, Silverstein JC, Ruffalo M, Teichmann SA, Pryhuber G, Misra RS, Purkerson JM, Fan J, Hickey JW. Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas Construction and Usage. bioRxiv. 2024:2024.03. 27.587041.
12. 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. Epub 2021/06/02. doi: 10.1016/j.cell.2021.04.048. PubMed PMID: 34062119; PMCID: PMC8238499.
13. 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. Epub 2021/06/06. doi: 10.1101/gr.275569.121. PubMed PMID: 34088715; PMCID: PMC8494219.
14. 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. Human molecular genetics. 2018;27(R1):R40-R7; PMCID: PMC5946857.
15. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. Epub 2018/02/08. doi: 10.1186/s13059-017-1382-0. PubMed PMID: 29409532; PMCID: PMC5802054.
16. 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.
17. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.
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. Epub 2023/06/09. doi: 10.1038/s41591-023-02327-2. PubMed PMID: 37291214; PMCID: PMC10287567 currently acting Chief Executive Officer of Iolux. F.J.T. consults for Immunai, Singularity Bio, CytoReason and Omniscope and has ownership interest in Dermagnostix and Cellarity. In the past 3 years, M.D.L. was a contractor for the CZI and received remuneration for talks at Pfizer and Janssen Pharmaceuticals. J.A.K. reports grants/contracts from Boehringer Ingelheim and Bristol Myers Squibb, consulting fees from Janssen and Boehringer Ingelheim and

study support from Genentech and is a member of the scientific advisory board of APIE Therapeutics. In the past 3 years, S.A.T. has received remuneration for consulting and Scientific Advisory Board membership from Genentech, Roche, Biogen, GlaxoSmithKline, Foresite Labs and Qiagen. S.A.T. is a co-founder and board member of and holds equity in Transition Bio. D.S. is a founder of Pliant Therapeutics and a member of the Genentech Scientific Advisory Board and has a sponsored research agreement with AbbVie. N.K. served as a consultant to Boehringer Ingelheim, Third Rock, Pliant, Samumed, NuMedii, Theravance, LifeMax, Three Lakes Partners, Optikira, AstraZeneca, RohBar, Veracyte, Augmanity, CSL Behring, Galapagos and Thyron over the past 3 years and reports equity in Pliant and Thyron, grants from Veracyte, Boehringer Ingelheim and Bristol Myers Squibb and nonfinancial support from miRagen and AstraZeneca. N.K. owns intellectual property on novel biomarkers and therapeutics in IPF licensed to biotechnology. O.R.-R. is a co-inventor on patent applications (PCT/US2016/059233, PCT/US2018/064553, PCT/US2018/060860, PCT/US2017/016146, PCT/US2019/055894, PCT/US2018/064563, PCT/US2020/032933) filed by the Broad Institute for inventions related to single-cell genomics. O.R.-R. has been an employee of Genentech since 19 October 2020 and has equity in Roche. O.E. serves in an advisory capacity to Pieris Pharmaceuticals, Blade Therapeutics, Delta 4 and YAP Therapeutics. Y.B. holds a Canada Research Chair in the Genomics of Heart and Lung Diseases. The remaining authors declare no competing interests.

19. 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. Epub 2016/07/06. doi: 10.1186/s13326-016-0088-7. PubMed PMID: 27377652; PMCID: PMC4932724.

20. 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. Epub 2022/05/27. doi: 10.1093/database/baac035. PubMed PMID: 35616100; PMCID: PMC9216545.

21. 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. Epub 2023/06/30. doi: 10.1371/journal.pbio.3002133. PubMed PMID: 37390046; PMCID: PMC10313015.

22. 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 data-driven cell-type ontologies. Sci Data. 2023;10(1):50. Epub 2023/01/25. doi: 10.1038/s41597-022-01886-2. PubMed PMID: 36693887; PMCID: PMC9873614.
23. Sievert C. Interactive web-based data visualization with R, plotly, and shiny: Chapman and Hall/CRC; 2020.