The Journal of Open Source Software

scTree: An R package to generate antibody-compatible classifiers from single-cell sequencing data

J. Sebastian Paez[1, 2], **Michael K. Wendt**[1, 2], **and Nadia Atallah Lanman**[1, 3]

1 Purdue University, Center for Cancer Research **2** Purdue University, Department of Medicinal Chemistry and Molecular Pharmacology **3** Purdue University, Department of Comparative Pathobiology

DOI: 10.21105/joss.02061

Software

- Review
- Repository
- Archive

Editor: William Rowe

Reviewers:

- @mschubert
- @jenzopr

Submitted: 14 January 2020
Published: 25 April 2020

Summary

Single-cell RNA sequencing (scRNA-seq) is now a commonly used technique to measure the transcriptome of populations of cells. Clustering heterogeneous cells based on these transcriptomes enables identification of cell populations (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018; Trapnell et al., 2014). There are multiple methods available to identify "marker" genes that differ between these populations (Butler et al., 2018; Love, Huber, & Anders, 2014; Robinson, McCarthy, & Smyth, 2009). However, there are usually too many genes in these lists to directly suggest an experimental follow-up strategy for selecting them from a bulk population (e.g. via FACS (Tung et al., 2007)). Here we present scTree, a tool that aims to provide biologists using the R programming language and scRNA-seq analysis programs a minimal set of genes that can be used in downstream experiments. The package is free, open source and available though GitHub at github.com/jspaezp/sctree

Implementation and results

The underlying model behind scTree is a combination of random forest for variable selection and a classification tree; having this model as a classifier relies on the fact that classification trees are analogous to many approaches in biology such as the gating strategy employed in flow cytometry or Fluorescence assisted cell sorting (FACS) experiments. In flow cytometry and FACS experiments, populations are identified and sorted based on expression levels of distinct markers that entail the identity or state of the chosen population. Usually such experiments use only relative levels of marker expression, using terms such as "High" and "Low" (Coquery, Loo, Buszko, Lannigan, & Erickson, 2012; Robertson & Scadden, 2005).

In a similar manner, scTree produces accurate, biologically relevant, and easily interpretable results, which can be used for subsequent subpopulation sorting and biological validation by fitting shallow decision trees analogous to FACS sorting strategies and is able to output these classifiers in a format easily interpretable in a wet-lab setting.

The method to calculate variable importances based on random forests has been previously described, and has been implemented in R by the `ranger` package (Altmann, Toloși, Sander, & Lengauer, 2010; Janitza, Celik, & Boulesteix, 2018; Wright & Ziegler, 2017). The suggestion of gating strategies is achieved by fitting a classification tree using the implementation provided by the `partykit` R package (Hothorn & Zeileis, 2015).

In order to benchmark the quality of markers, we utilized a recall-based strategy. Briefly, each dataset was split randomly into two sections, a training set with 80% of the cells and a testing set consisting of the 20% remaining. A classifier was trained by selecting the top 5 markers suggested for each cluster by either scTree (Altman method) or by two of the most commonly used marker gene detection methods for scRNA-seq data: t-tests or wilcoxon-tests (as implemented by Seurat v3.0.1).

These classifiers were then used to predict the identity of the testing set and the quality was assesed by comparing the recall, accuracy and precision of the prediction. We were concerned that the forest-based markers would artificially favor scTree, therefore we utilized several classifiers for the markers derived from either scTree, t-tests or wilcoxon-tests. As shown in **Figures 1 and 2**, bias was not observed, and regardless of the final classification model, the features selected by using scTree provide a comparable accuracy, precision and recall to those acquired using traditional differential expression methods. It is important to note that many of the wrongly assigned labels happen between cell populations that are hard to define in-vivo and are not resolved clusters in the UMAP dimensional reduction, such as macrophage subtypes and between NK and Tc cells.

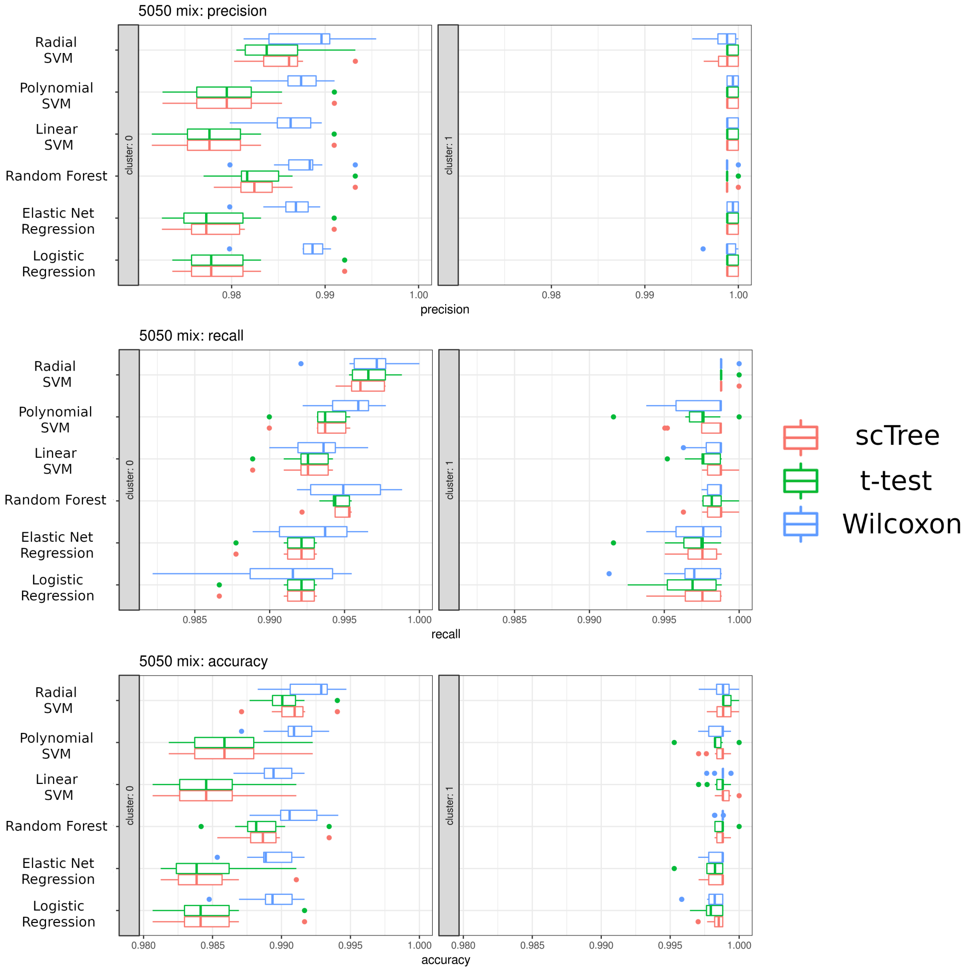
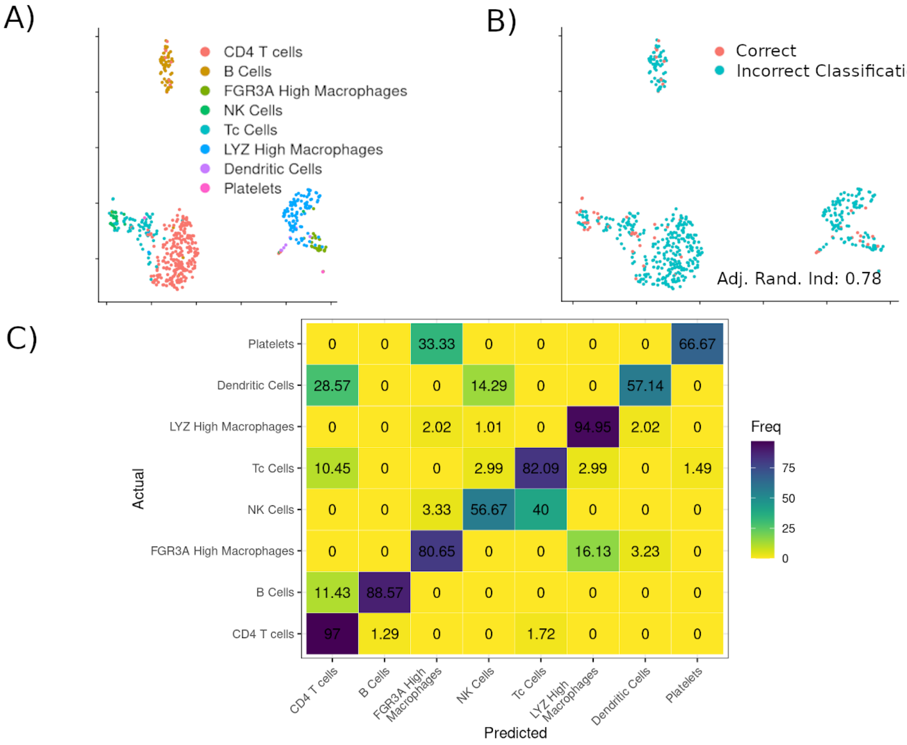


Figure 1: Depiction of the classification performance achieved in the Jurkat:293 50:50 dataset. A number of machine learning algorithms were tested to ensure that scTree performed as well as traditional marker identification approaches, regardless of the classifier used.



Actual \ Predicted	CD4 T cells	B Cells	FGR3A High Macrophages	NK Cells	Tc Cells	LYZ High Macrophages	Dendritic Cells	Platelets
Platelets	0	0	33.33	0	0	0	0	66.67
Dendritic Cells	28.57	0	0	14.29	0	0	57.14	0
LYZ High Macrophages	0	0	2.02	1.01	0	94.95	2.02	0
Tc Cells	10.45	0	0	2.99	82.09	2.99	0	1.49
NK Cells	0	0	3.33	56.67	40	0	0	0
FGR3A High Macrophages	0	0	80.65	0	0	16.13	3.23	0
B Cells	11.43	88.57	0	0	0	0	0	0
CD4 T cells	97	1.29	0	0	1.72	0	0	0

Figure 2: Depiction of the predicted identities in the PBMC 3k dataset dataset. A. Real identities are the identities as identified through unsupervised clustering performed using Seurat and annotated based on prior knowledge (Butler et al., 2018). **B.** The scTree package was then used to classify cells based on the top 5 markers for each cluster chosen by scTree and accurately recapitulates the original classification determined by Seurat. **C.** Confusion matrix showing the assigned classification to each cluster.

Example Output from the package

Predictor generation

As mentioned previously, a main focus in the development of scTree was the biological interepretability of the models. Therefore the models can be expressed as a Garnett file, as shown in **Code Section 1**, as specified originally in the Garnett manuscript by the Trapell lab (Pliner, Shendure, & Trapnell, 2019). Visualizations are designed to resemble flow cytometry results, as show in **Figure 3** and connections with several antibody vendors are provided to query the availability of probes for the genes found to be usefull for classification.

```
as.garnett(tree_model, rules_keep = "^NK")
# > NK Cells     (n = 88)
# expressed above: CST7 1.636, TYROBP 2.451
```

Code Section 1. *Suggested classification scheme for NK cell cluster of the PBMC dataset. The data depicts how the cluster corresponding to NK cells can be predominantly identified as GNLY High/GZMB High.*

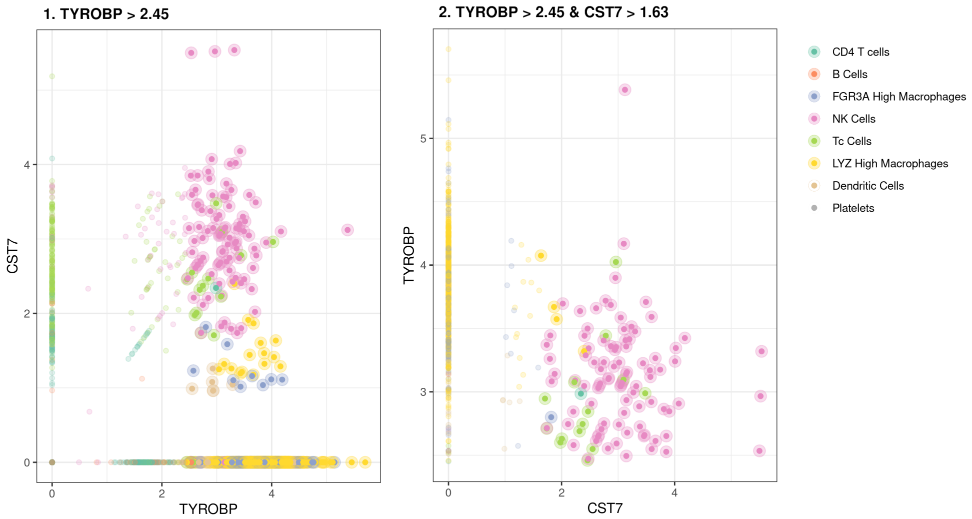


Figure 3: Scatterplot showing the progressive gating that would be used to classify node 11 in the 3K PBMC dataset. Filtering in each pane is done on the gene presented on the X-axis of the plot and cells kept during that filtering step are highlighted

Despite scTree being originally developed for single cell sequencing, we recognize it could also be used for other high content single-cell workflows, such as CyTOF or data driven multiple-channel flow cytometry.

Antibody querying interface

The provided interface with antibody databases, further enhances the utility of scTree by fulfilling the need to interface *in silico* models and data with *in vitro* followup. Therefore, a package interface with common antibody vendors and search engines are provided. This interface is exemplified in **Code section 2**.

```r
require(sctree)
head(query_biocompare_antibodies("CD11b"))
#>                                    title            vendor
#> 1         Anti-CD11b antibody [EPR1344]             Abcam
#> 2             Anti-CD11b/ITGAM Antibody         BosterBio
#> 3    Anti-CD11b/ITGAM Picoband Antibody         BosterBio
#> 4 Anti-CD11b Rabbit Monoclonal Antibody         BosterBio
#> 5  Monoclonal Antibody to CD11b (human)   MyBioSource.com
#> 6                Anti-CD11b (Mouse) mAb MBL International
#>                                                            specification
#> 1 Applications: WB, IHC-p; Reactivity: Hu, Ms, Rt, Pg, RhMk; Conjugat
#> 2             Applications: Western Blot (WB); Reactivity: Hu, Ms, Rt
#> 3   Applications: WB, FCM, ICC, IHC-fr, IHC-p; Reactivity: Hu, Ms, Rt
#> 4                       Applications: WB, IF, IHC; Reactivity: Hu, Ms
#> 5              Applications: Flow Cytometry (FCM); Reactivity: Human
#> 6              Applications: Flow Cytometry (FCM); Reactivity: Mouse
```

Code Section 2. *Example of the automated antibody query interface*

Additional usage cases and up-to-date code snippets of the common functions can be found in the package documentation website (jspaezp.github.io/sctree/) and the readme file hosted in the github repository (github.com/jspaezp/sctree).

Methods

Testing dataset processing

The filtered raw counts for each dataset were downloaded from the 10x website single cell expression datasets (10X-Genomics, 2019) and were processed by the standard Seurat workflow, as described in the package tutorial ("Satija-Lab", 2019). This process was carried out for the following datasets:

1. 3k PBMC, Peripheral Blood Mononuclear Cells (PBMC)
2. 50%:50% Jurkat:293T Cell Mixture, originally published by Wan, H. et al. in 2017

Description of the benchmarking process

Briefly, each dataset was split into a testing and a training set. For each cluster, each of the different marker identification methodologies was used and the top five markers were selected. These five markers were used to train a prediction model on the training set and the predicitons were carried out on the testing set. These predictions were compared with the assigned cluster identity and performance metrics were calculated.

Formulas defining the prediction quality

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$accuracy = \frac{True\ Positives + True\ Negatives}{Total}$$

Acknowledgments

This study was supported by the Computational Genomics Shared Resource at the Purdue University Center for Cancer Research (NIH grant P30 433 CA023168), IU Simon Cancer Center (NIH grant P30 CA082709), and the Walther Cancer Foundation.

References

10X-Genomics. (2019). 10X Genomics datasets. https://web.archive.org/web/20190827115031/https://support.10xgenomics.com/single-cell-gene-expression/datasets.

Altmann, A., Toloși, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. doi:10.1093/bioinformatics/btq134

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411–420. doi:10.1038/nbt.4096

Coquery, C. M., Loo, W., Buszko, M., Lannigan, J., & Erickson, L. D. (2012). Optimized protocol for the isolation of spleen-resident murine neutrophils. *Cytometry Part A*, *81A*(9), 806–814. doi:10.1002/cyto.a.22096

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909. Retrieved from http://jmlr.org/papers/v16/hothorn15a.html

Janitza, S., Celik, E., & Boulesteix, A.-l. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*(4), 885–915. doi:10.1007/s11634-016-0276-4

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. doi:10.1186/s13059-014-0550-8

Pliner, H. A., Shendure, J., & Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nature Methods*, *16*(10), 983–986. doi:10.1038/s41592-019-0535-3

Robertson, P., & Scadden, D. T. (2005). Differentiation and Characterization of T Cells. *Current Protocols in Immunology*. doi:10.1002/0471142735.im22f08s69

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. doi:10.1093/bioinformatics/btp616

"Satija-Lab". (2019). Seurat - guided clustering tutorial. https://web.archive.org/web/20200309135737/https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381–386. doi:10.1038/nbt.2859

Tung, J. W., Heydari, K., Tirouvanziam, R., Sahaf, B., Parks, D. R., Herzenberg, L. A., & Herzenberg, L. A. (2007). Modern flow cytometry: A practical approach. *Clin Lab Med*, *27*(3), 453–68, v. Journal Article. doi:10.1016/j.cll.2007.05.001

Wright, M. N., & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1). doi:10.18637/jss.v077.i01