The Journal of Open Source Software

ELeFHAnt: Ensemble Learning for Harmonization and Annotation of Single Cell Data

Konrad Thorner[1] **and Praneet Chaturvedi**[1]

1 Cincinnati Children's Hospital Medical Center

DOI: 10.21105/joss.03516

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 19 July 2021
Published: 21 July 2021

Summary

Single cell sequencing has become a powerful method for understanding biological systems at an increasingly granular level. For single cell RNA-seq data specifically, one of the primary questions is using the transcriptome to determine cell identity. It is common to visualize such data and find clusters of cells with similarity in gene expression, but assigning each cluster a cell type is a much more open-ended task. Taking advantage of publicly-available, annotated datasets in combination with supervised learning is a powerful method for addressing this question. Ensemble Learning for Harmonization and Annotation of Single Cells (ELeFHAnt) provides an easy to use R package for users to annotate clusters of single cells, harmonize labels across single cell datasets to generate a unified atlas, and infer relationships among cell types between two datasets. It provides users with the flexibility of choosing between random forest and SVM (Support Vector Machine) based classifiers or letting ELeFHAnt apply both in combination to make predictions.

Statement of Need

As an alternative to manual annotation, there are many automatic cell annotation tools currently available that employ either gene marker, correlation, or machine learning-based methods, each with varying levels of performance (Pasquini et al., 2021). The label transfer functionality of Seurat is among the most well-known, but occurs on an individual cell level rather than a community level, often leading to over-annotation (Stuart et al., 2019). There are also deep learning-based tools emerging such as scANVI, which utilizes generative models but requires significantly more computation time (Xu et al., 2021).

ELeFHAnt is a supervised machine learning-based tool that enables researchers to identify cell types in their scRNA-seq data while providing additional unique features. ELeFHAnt gives users the ability to use and compare not just one but multiple classification algorithms simultaneously through its ensemble method, weighting their predictions to produce the best consensus among them. In this ensemble, SVM and random forest are our two classifiers selected for their superior accuracy and computation time in a benchmarking study (Abdelaal et al., 2019). Additionally, selecting the optimal reference is a challenge addressed by harmonization, that allows users to integrate multiple datasets together into an atlas. A standardized set of labels is generated across all of them, which can subsequently be used to annotate new datasets. Relationships between two datasets can also be deduced to better understand how each was annotated. This is provided in an easy to interpret heatmap format that compares all the cell types between them. Finally, a subsampling procedure is used to enable faster predictions while being shown not to influence reproducibility.

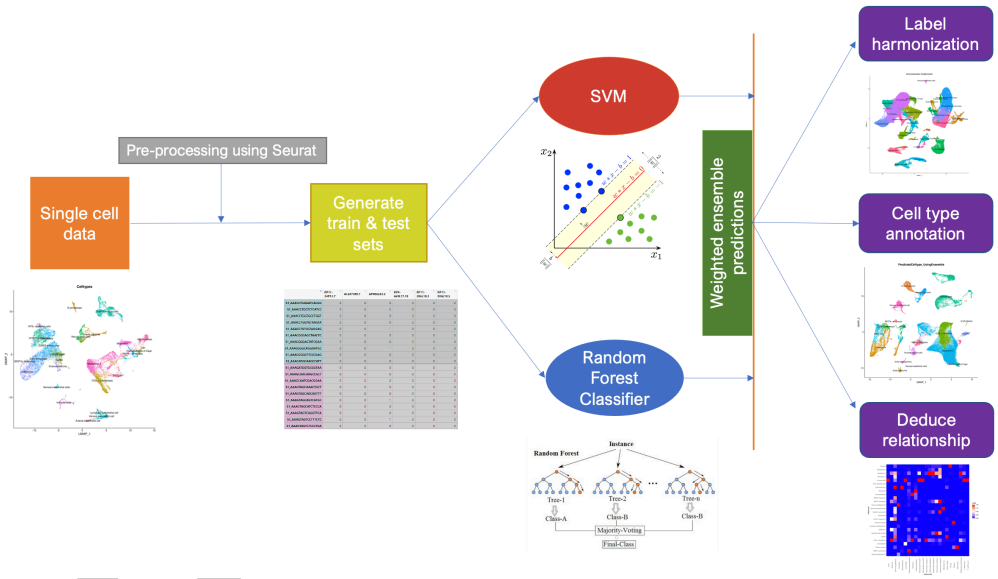
ELeFHAnt has been tested on multiple public datasets involving multiple time points in fetal development, where it was still able to identify cell types. It has also been used internally within

Cincinnati Children's Hospital Medical Center across different projects, including annotation of snRNA (single nucleus RNA) sequencing with a harmonized gut cell atlas.

Overview and Examples

ELeFHAnt makes use of Seurat for the initial input data and pre-processing. It will then generate the training and test sets from the reference and query respectively, with optional subsampling. SVM and Random Forest are the classifiers that can be used seperately or in an ensemble. Classification accuracy of both are used to assign weights to the predictions from each classifier. These weighted confusion matrices are normalized based on the largest number of cells shared among celltypes and assigned clusters. They are then added together for the final ensemble predictions.

Figure 1



The attributes of our three example datasets of early gut development are shown below, as well as those of the integrated dataset. "Fetal" refers to a subset of terminal ileum (TI) data from an atlas for human fetal intestinal development called "STAR-FINDer" (Fawkner-Corbett et al., 2021). "Gut" refers to a subset of duojejunum cell data from the Gut Cell Atlas, which also examines intestinal development from 6-10 weeks post-conception (Elmentaite et al., 2020). Lastly, "Spence" refers a subset of fetal intestinal data from a multi-endodermal organ atlas (Yu et al., 2021).

Table 1

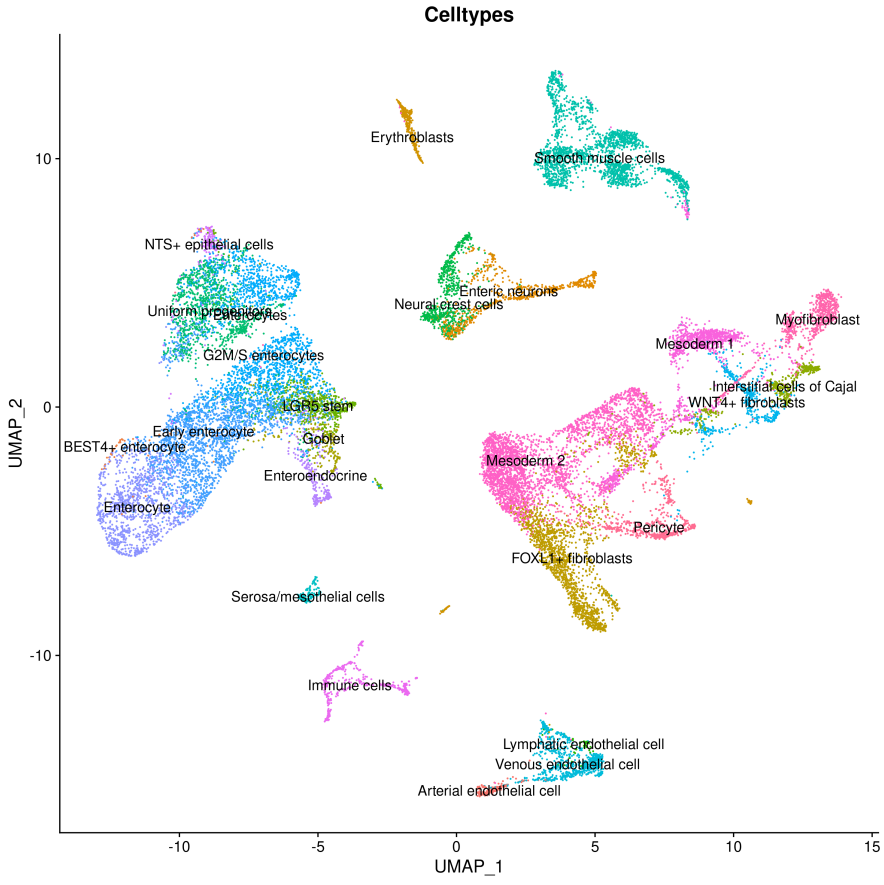
	Fetal	Gut	Spence	Integrated
# of Cells	24787	21592	77672	124245
# of Genes	33538	33694	26879	2000
# of Clusters	25	31	26	30
# Cell Types	95	26	24	145
Median nFeature	2029	2858	1913	2121
Median nCount	5141	9988	3754	4833

Celltype Annotation

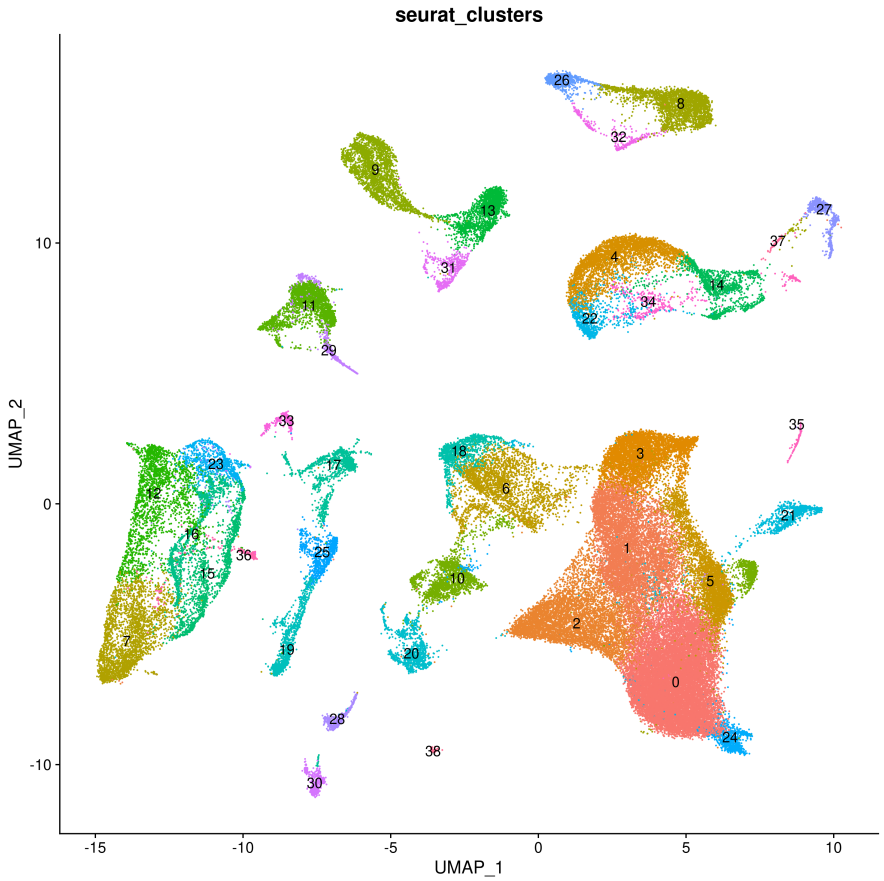
To demonstrate Celltype Annotation using ELeFHAnt we used Gut Cell Atlas (Elmentaite et al., 2020) as reference and Fetal intestinal data (Yu et al., 2021) as query. Reference and query were downsampled to 200 cells per Celltypes and seurat_clusters respectively to enable fast computation.

Figure 2

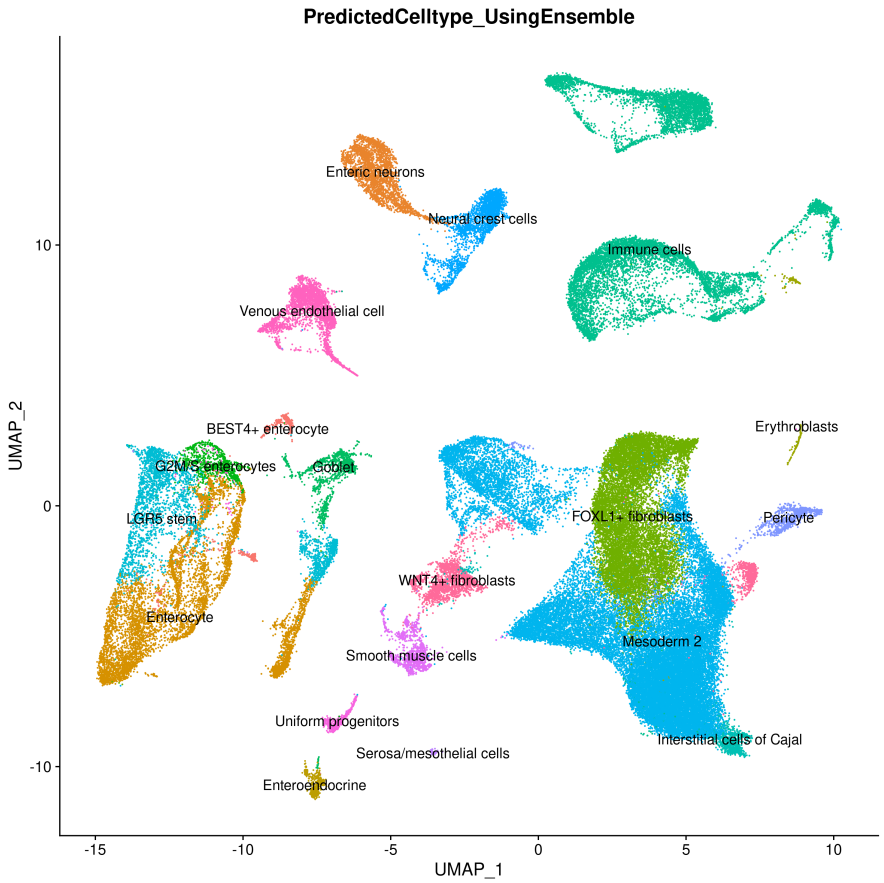
A



B


seurat_clusters
UMAP_2
UMAP_1
10
0
-10
-15
-10
-5
0
5
10
26
8
32
9
13
27
37
4
31
14
11
34
22
29
33
35
18
3
23
17
6
12
21
16
36
25
1
15
10
5
2
7
19
20
0
28
24
38
30

C



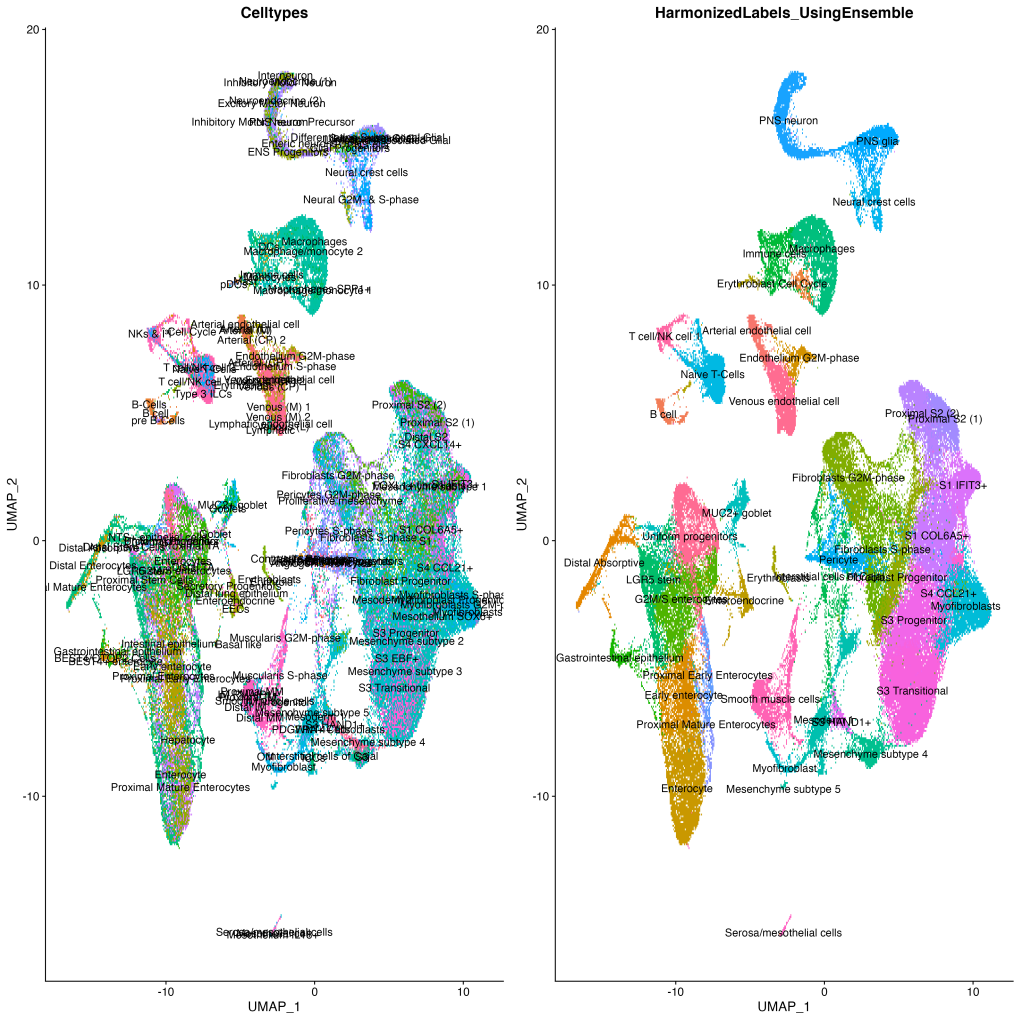
4

(A) represents celltypes in the reference dataset displayed on a UMAP. (B) represents Seurat clusters displayed on the query, and (C) represents the predicted celltypes as determined by ELeFHAnt's ensemble approach.

Label Harmonization

To demonstrate LabelHarmonization we used three datasets: 1) Gut Cell Atlas (Elmentaite et al., 2020)) 2) Fetal intestinal data (Yu et al., 2021) from Dr. Spence's Lab 3) Fetal intestine data from STAR-FINDer (Fawkner-Corbett et al., 2021). Data shown below is based on subsetting 200 cells per celltype in each dataset to harmonize the atlas of ~125k cells.

Figure 3

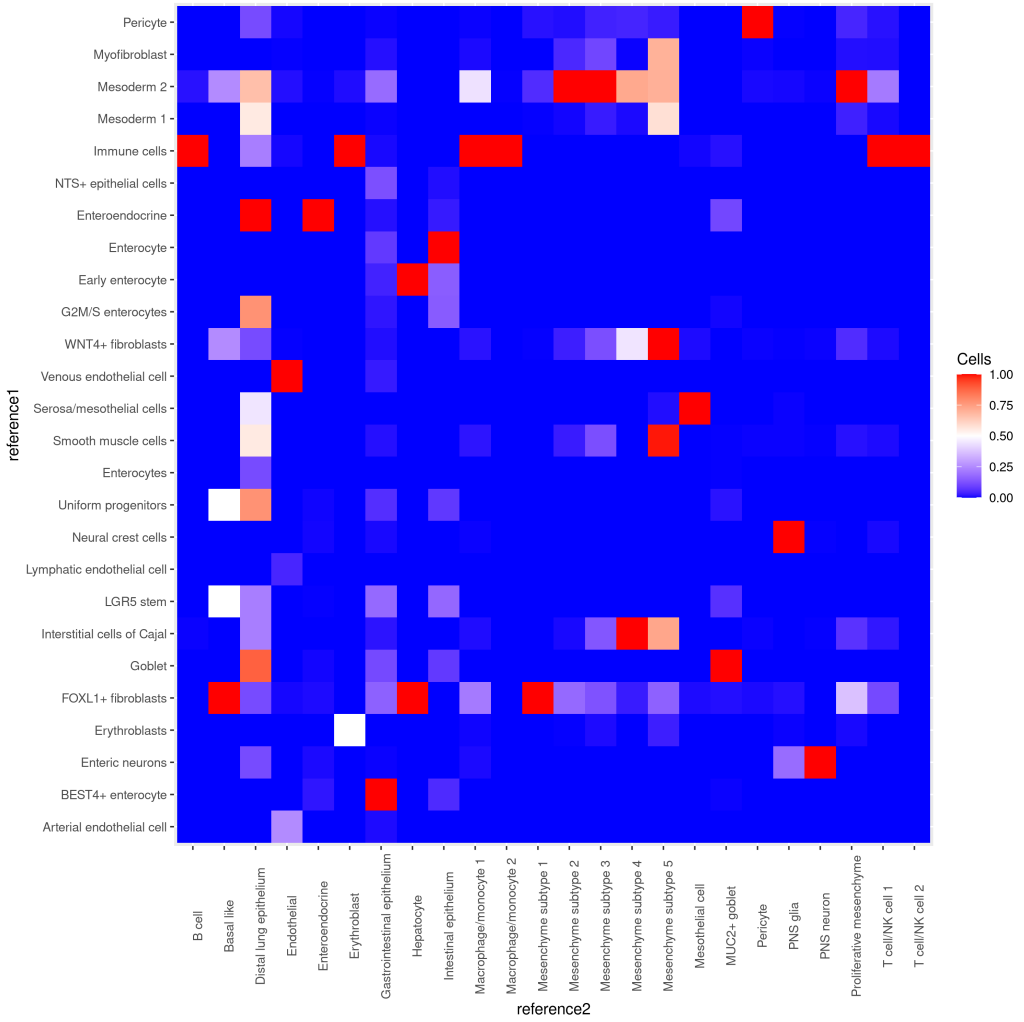


The UMAP on the left represents the 144 total celltypes from the three datasets after integration. The UMAP on the right is the result of ELeFHAnt's harmonization with the ensemble method, showing the resulting labels for each cluster.

Deduce Relationship

To demonstrate Deduce Relationship we used two datasets that were also uses in the harmonization example: 1) Gut Cell Atlas (Elmentaite et al., 2020) 2) Fetal intestinal data (Fawkner-Corbett et al., 2021) from Dr. Spence's Lab. Data shown below is based on subsetting to 500 cells per celltype in each dataset.

Figure 4



The heatmap depicts the relationship between celltypes for the two references, with each red square showing which cell type in reference 2 best matches a particular celltype in reference 1.

Acknowledgements

We would like to thank Drs. Emily Miraldi, Nathan Salomonis, Anil Jegga, and Aaron Zorn at Cincinnati Children's Hospital Medical Center for their valuable feedback.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, H., D.and Mei, & Reinders, A., M.and Mahfouz. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*. https://doi.org/10.1186/s13059-019-1795-z

Elmentaite, R., Ross, A., Roberts, K., James, K., Ortmann, D., Gomes, T., Nayak, K., Tuck, L., Pritchard, S., Bayraktar, O., Heuschkel, R., Vallier, L., Teichmann, S., & Zilbauer, M. (2020). Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links

to Childhood Crohn's Disease. *Developmental Cell*, *55*. https://doi.org/10.1016/j.devcel.2020.11.010

Fawkner-Corbett, D., Antanaviciute, A., Parikh, K., Jagielowicz, M., Sousa Gerós, A., Gupta, T., Ashley, N., Khamis, D., Fowler, D., Morrissey, E., Cunningham, C., Johnson, P., Koohy, H., & Simmons, A. (2021). Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, *184*. https://doi.org/10.1016/j.cell.2020.12.016

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, *19*. https://doi.org/10.1016/j.csbj.2021.01.015

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*. https://doi.org/10.1016/j.cell.2019.05.031

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., & Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, *17*. https://doi.org/10.15252/msb.20209620

Yu, Q., Kilik, U., Holloway, E., Tsai, Y., Harmel, C., Wu, A., Wu, J., Czerwinski, M., Childs, C., He, Z., Capeling, M., Huang, S., Glass, I., Higgins, P., Treutlein, B., Spence, J., & Camp, J. (2021). Charting human development using a multi-endodermal organ atlas and organoid models. *Cell*, *184*. https://doi.org/10.1016/j.cell.2021.04.028