

Gene expression

Ensemble learning for classifying single-cell data and projection across reference atlases

Lin Wang[†], Francisca Catalan[†], Karin Shamardani, Husam Babikir and Aaron Diaz  *

Department of Neurosurgery, University of California, San Francisco, CA 94158, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on July 5, 2019; revised on December 12, 2019; editorial decision on February 20, 2020; accepted on February 24, 2020

Abstract

Summary: Single-cell data are being generated at an accelerating pace. How best to project data across single-cell atlases is an open problem. We developed a boosted learner that overcomes the greatest challenge with status quo classifiers: low sensitivity, especially when dealing with rare cell types. By comparing novel and published data from distinct scRNA-seq modalities that were acquired from the same tissues, we show that this approach preserves cell-type labels when mapping across diverse platforms.

Availability and implementation: <https://github.com/diazlab/ELSA>

Contact: aaron.diaz@ucsf.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Results

There are multiple efforts ongoing to assemble atlases of cells derived from complex tissues via single-cell RNA-sequencing (scRNA-seq). As these projects scale from thousands to millions of cells, extending cell-type annotations to newly generated data becomes challenging. How best to project single-cell data from one reference atlas to another, or onto bulk-sequencing databases remains an open problem.

We developed an ensemble classifier of scRNA-seq, single-nuclei RNA-sequencing (snRNA-seq) and bulk-extraction RNA-sequencing (RNA-seq) data: ensemble learning for classifying single-cell data and projection across reference Atlases (ELSA; <https://github.com/diazlab/ELSA>). We trained ELSA on public atlases and tested it on published single-cell data, novel scRNA-seq and snRNA-seq of human glioma tissues (four patients, >11K cells, [Supplementary Tables S1 and S2](#)).

ELSA first identifies optimal gene sets for cross-platform analyses using a random-forest approach ([Fig. 1A](#)). Training and classification are then performed using a boosted learner, incorporating random under-sampling to accommodate class imbalances ([Supplementary Methods](#)). The resulting classifier can then be used to project single-cell data between reference atlases or to update an existing atlas with newly acquired data ([Fig. 1B](#)).

Multinomial regression has also been used for classification of single-cell data ([Pliner *et al.*, 2019](#)). However, we find that regression models, support-vector machines, and other status-quo approaches have low sensitivity for this purpose, as assessed via 10-fold cross-validation ([Fig. 1C](#)). This is especially the case

when regression is applied to datasets with imbalanced class labels and/or rare cell types ([Fig. 1D](#)). ELSA's boosting approach builds an ensemble learner by averaging a sequence of weak learners, each emphasizing training instances that were misclassified by previous models. Moreover, random under-sampling of the training data is used to achieve a balanced class distribution. Thereby, rare cell types are classified more accurately. In our tests, we found that under-sampling was critical to achieve optimal performance. Boosting alone, and even boosting with over-sampling did not achieve the same performance as ELSA ([Fig. 1C and D](#)), with the most pronounced differences in performance observed in rare cell types (e.g. CD34+ stem cells; [Supplementary Fig. S1A and B](#)).

[Kiselev *et al.* \(2018\)](#) showed that single-cell data can be projected across experiments, using similarity metrics to compare cells to each other or to a cell-cluster centroid. However, their approach does not explicitly accommodate differences in platform used for data acquisition. To tackle this problem, ELSA uses a random-forest method to optimize gene selection for downstream comparisons. We found that this approach for feature selection improved both the sensitivity and specificity of most methods ([Fig. 1E and F](#)). To illustrate our approach, we trained ELSA using thusly optimized gene sets and projected single-cell data from Smart-seq2 to Drop-seq platforms ([Supplementary Fig. S1C–E](#)) and from snRNA-seq to scRNA-seq ([Supplementary Fig. S1F](#)). Our results show with greater sensitivity than competing approaches ([Supplementary Fig. S1G](#)).

To further test ELSA using heterogeneous tissues, we performed scRNA-seq of four human glioma specimens and combined this

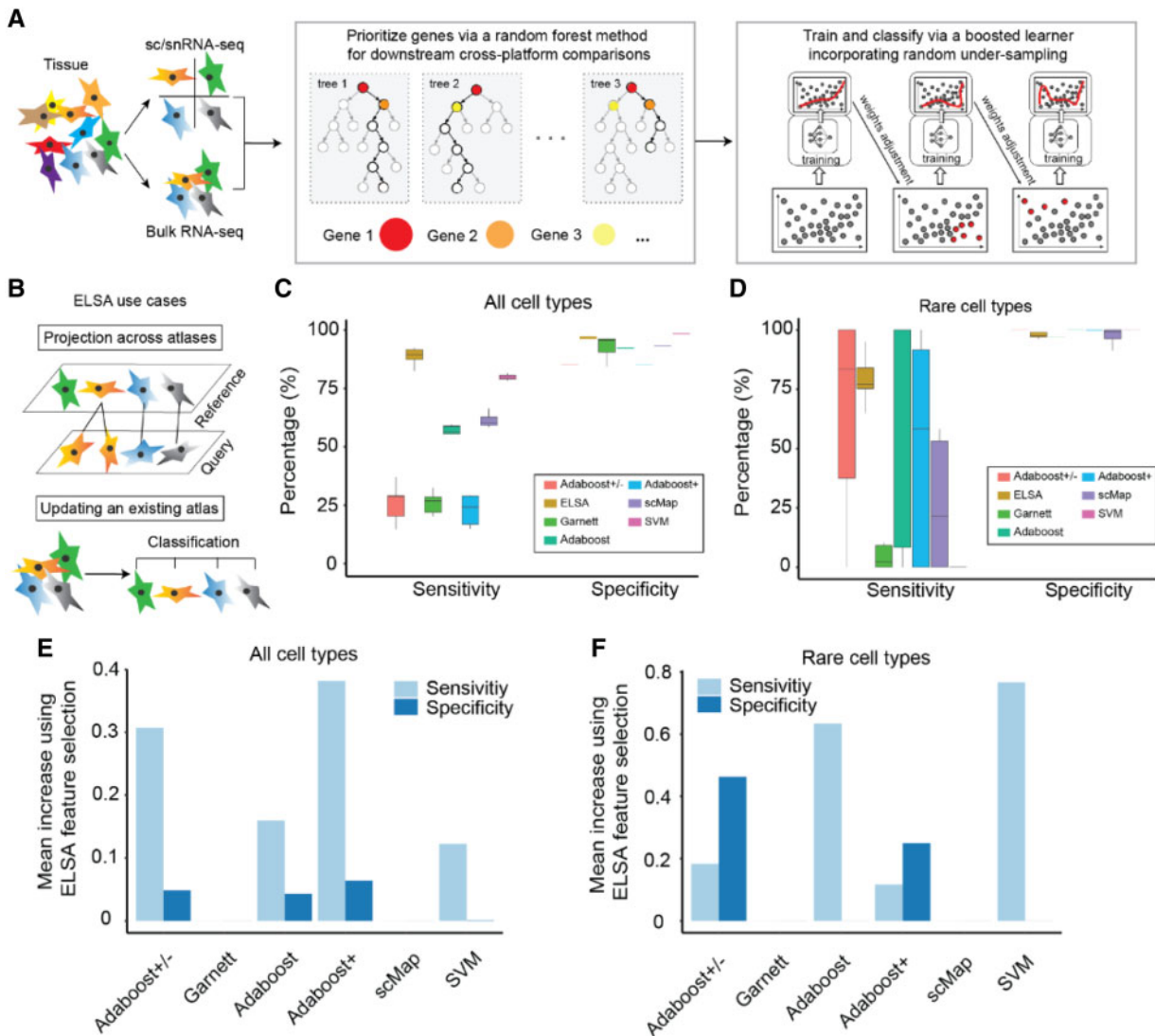


Fig. 1. (A) Summary diagram of ELSA's methodology. (B) ELSA use cases. (C) A comparison of sensitivity and specificity between ELSA and competing algorithms on the PBMC dataset, performed via 10-fold cross-validation. Adaboost +/- refers to Adaboost with combined under and synthetic over sampling. Adaboost+ refers to Adaboost with synthetic over-sampling. (D) Sensitivity and specificity as in (C) for rare cell types, defined as those representing <10% of the training data. (E) Increases in sensitivity and specificity after pre-processing via ELSA's feature selection method. (F) Sensitivity and specificity increase as in (E) for rare cell types

with data from four published cases from the same pipeline (Wang et al., 2019). We then mapped those cells to cell types found in the non-malignant brain. The resulting classification maps immune and stromal cells accurately and associates neoplastic glioblastoma cells with their non-malignant glial counterparts (Supplementary Fig. S2).

Recent advances underscore the importance of feature selection for comparing single-cell data across platforms (Kiselev et al., 2018, 2019) and feature extraction for cell-type identification (Risso et al., 2018). Random-forest classifiers implicitly rank of features by importance, which can be assessed from the misclassification error incurred when a feature's values are permuted. Thus, random forests provide a principled approach to feature selection.

Under-sampling corrects for the class imbalances that occur in most single-cell datasets from complex tissue. This enables the accurate classification of rare cell types. We conclude from our cross-validation analysis that ensemble methods such as ours are less prone to overfitting than status quo approaches.

Ensemble methods like these are not without their limitations. Ensemble methods are often more difficult to interpret than regression models. Variable importance analysis can provide insight. But, if an ensemble method outperforms a linear regression model then it is likely exploiting some non-linear interaction effect that variable importance analysis won't account for.

Garnett (Pliner et al., 2019) defines a hierarchical markup language for annotating cell type based on marker gene expression. This functionality is beyond the scope of ELSA. However, Garnett has a modular structure for marker gene analysis and training data selection, followed by ELSA for classification. ELSA also isn't suitable for cell-type discovery, many codes exist for that purpose (e.g. Butler et al., 2018).

Funding

This work has been supported by awards from the American Cancer Society (IRG-17-180-19), the Brain Tumor Funders' Collaborative, the University of California, San Francisco Glioma Precision Medicine Program, the National

Institutes of Health (P30CA082103) and the University of California Cancer Research Coordinating Committee to A.D.

Conflict of Interest: none declared.

References

- Butler, A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
- Kiselev, V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
- Kiselev, V.Y. *et al.* (2018) Scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.
- Pliner, H.A. *et al.* (2019) Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, **16**, 983–986.
- Risso, D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284–301.
- Wang, L. *et al.* (2019) The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.*, **9**, 1708–1719.