Bioinformatics, 37(6), 2021, 882–884
doi: 10.1093/bioinformatics/btaa755
Advance Access Publication Date: 26 August 2020
Applications Note

Systems biology

STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data

Massimo Andreatta [1,2,3] **and Santiago J. Carmona** [1,2,3,*]

[1]Ludwig Institute for Cancer Research Lausanne, University of Lausanne, CH-1066 Epalinges, Switzerland, [2]Department of Oncology, CHUV, UNIL CHUV, CH-1066 Epalinges and [3]University of Lausanne, Lausanne, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received and revised on June 30, 2020; editorial decision on August 17, 2020; accepted on August 19, 2020

Abstract

Summary: STACAS is a computational method for the identification of integration anchors in the Seurat environment, optimized for the integration of single-cell (sc) RNA-seq datasets that share only a subset of cell types. We demonstrate that by *(i)* correcting batch effects while preserving relevant biological variability across datasets, *(ii)* filtering aberrant integration anchors with a quantitative distance measure and *(iii)* constructing optimal guide trees for integration, STACAS can accurately align scRNA-seq datasets composed of only partially overlapping cell populations.

Availability and implementation: Source code and R package available at https://github.com/carmonalab/STACAS; Docker image available at https://hub.docker.com/repository/docker/mandrea1/stacas_demo.

Contact: santiago.carmona@unil.ch

1 Introduction

Massively parallel single-cell transcriptomics (scRNA-seq) has emerged as a transformative technology that enables measuring molecular profiles at single-cell resolution. However, despite the highly multiplexed technologies, single-cell data are produced separately for different tissues and organs and are affected by multiple batch effects, such as different sample processing and scRNA-seq protocols. As such, integration of single-cell data might be the ultimate challenge in the field towards the generation of single-cell atlases (Eisenstein, 2020; Lähnemann *et al.*, 2020; Regev *et al.*, 2017).

Seurat (Stuart *et al.*, 2019) is currently one of the most popular and best performing algorithms for single-cell data integration, and can be effortlessly integrated into complex analysis pipelines (Tran *et al.*, 2020). At the core of the Seurat integration algorithm is the identification of mutual nearest neighbors (MNN) across single-cell datasets, named 'anchors', in a reduced space obtained from canonical correlation analysis (CCA). These anchors and their scores are used to compute correction vectors for each query cell, transforming (i.e. batch-correcting) its expression profile (Haghverdi *et al.*, 2018). Transformed cell profiles can then be jointly analyzed as part of an integrated space. To handle more than two datasets, a guide tree based on pairwise batch similarities is used to dictate the batch integration order. While Seurat has proven very powerful for the removal of technical artifacts between replicated experiments or even different sequencing technologies (Tran *et al.*, 2020), it tends to overcorrect batch effects and performs poorly when integrating heterogeneous datasets (Luecken *et al.*, 2020), where only a fraction of cell types are shared between individual samples. This is crucial for the creation of reference cell type-specific single-cell atlases where the datasets to integrate were obtained from different tissues or experimental conditions (e.g. T cells from blood versus tumor-infiltrating T cells), and as a consequence are composed of different, partially overlapping cell states or sub-types.

2 Results

STACAS is a package for determining integration anchors between heterogeneous datasets, and it is designed to be easily incorporated into Seurat dataset integration pipelines. STACAS uses a reciprocal principal component analysis (PCA) procedure to calculate anchors, where each dataset in a pair is projected onto the reduced PCA space of the other dataset; mutual nearest neighbors are then calculated in these reduced spaces. Crucially, and in contrast to the CCA reduction used by Seurat, the expression values of genes used in generating the PCA spaces are not rescaled to have zero mean and unit variance. When integrating heterogeneous datasets, for instance composed only of $CD4^+$ or $CD8^+$ T cells, such rescaling can cancel out important biological differences between the datasets (Fig. 1A).

A second innovation introduced in STACAS is the filtering of anchors based on anchor pairwise distance, which is calculated on the reduced PCA spaces used to determine the anchors. We observed that the distribution of anchor distances between datasets with shared cell subtypes (i.e. a sample containing both $CD4^+$ and $CD8^+$ T cells, compared to a sample of $CD8^+$ T cells only) is centered on lower pairwise anchor distances compared with dataset pairs with limited or no overlap (e.g. a $CD4^+$ sample and a $CD8^+$ sample)

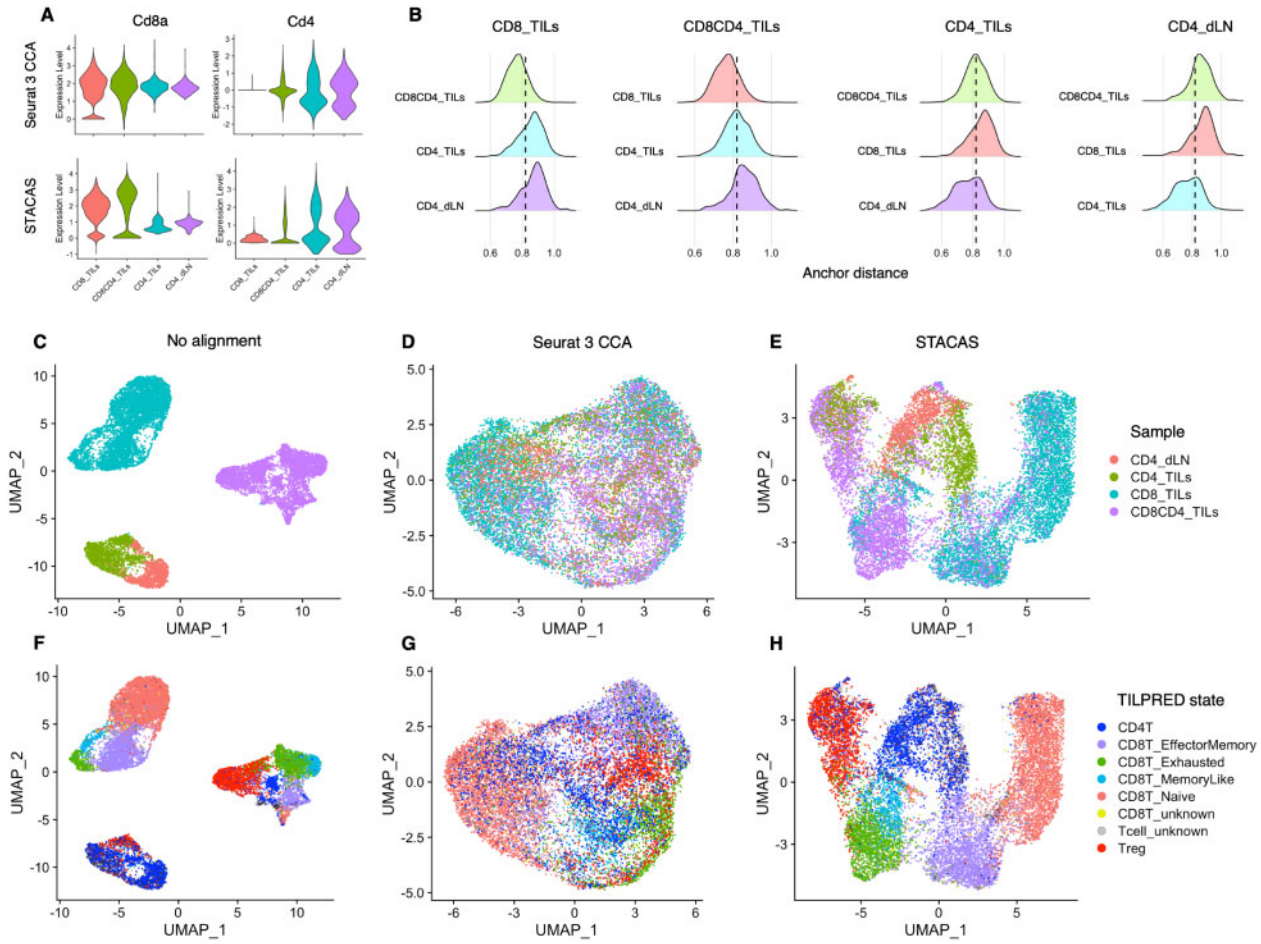

Fig. 1. Anchor finding and dataset integration using STACAS. (**A**) Expression level (log [normalized UMI counts + 1]) of *Cd8a* and *Cd4* after integration with Seurat CCA (top) or STACAS (bottom); important biological differences between the samples are lost by data rescaling and sub-optimal anchoring by Seurat 3 CCA. (**B**) Anchor distance distribution between pairs of samples prior to anchor filtering by STACAS; poor anchors with distance higher than threshold (represented with a vertical dashed line) are filtered out by STACAS. (**C–E**) Low-dimensionality UMAP visualization of scRNA-seq data, colored by sample, without batch correction (C), using Seurat CCA anchors (D) and using STACAS anchors (E) for dataset alignment. (**F–H**) UMAP visualization of scRNA-seq data, colored by TILPRED state prediction, without batch correction (F), using Seurat CCA anchors (G) and using STACAS anchors (H) for dataset alignment

(Fig. 1B); anchor distance can therefore be used as a quantitative measure to filter spurious anchors and improve dataset integration. In STACAS, the anchor filtering threshold defaults to the 80th percentile of the distance distribution between the two most similar datasets included in the integration task.

Finally, the anchors determined by STACAS can be used directly for dataset integration using the IntegrateData function in Seurat 3. STACAS suggests a guide tree to determine the order in which datasets are to be integrated. In contrast to the Seurat default guide tree, which favors datasets with the highest total number of cells in any given pair, STACAS prioritizes samples with the highest total number of anchors; the rationale being that datasets with many anchors are likely to contain more cell types and represent the 'centroid' of the integrated map.

In the example in Figure 1, we integrated four scRNA-seq datasets of mouse T cells from public repositories, composed of (i) $CD8^+$ tumor-infiltrating lymphocytes (TILs) (Carmona *et al.*, 2020); (ii) $CD4^+$ and $CD8^+$ TILs (Xiong *et al.*, 2019); (iii) $CD4^+$ T cells from tumors (Magen *et al.*, 2019) and (iv) $CD4^+$ T cells from tumor-draining lymph nodes (dLN) (Magen *et al.*, 2019). There is an evident batch effect between the samples, with the cells of each sample clustering together regardless of their type (Fig. 1C and F). Consistently with a recent benchmark (Luecken *et al.*, 2020), dataset alignment using Seurat 3 appears to overcorrect these batch effects, overlaying samples with little in common such as $CD4^+$ dLN and $CD8^+$ TILs (Fig. 1D and G). In contrast, STACAS only aligns cells with similar states across samples, limiting the superposition of $CD4^+$ with $CD8^+$ cells (Fig. 1E). Supervised cell state classification using TILPRED (Carmona *et al.*, 2020) confirms that in most cases STACAS was able to cluster cell types across different, heterogeneous datasets (Fig. 1H). We obtained similar, consistent results on larger-scale integration tasks toward the construction of reference T cell maps in cancer and chronic infection (Andreatta *et al.*, 2020). An interactive TIL reference atlas constructed using STACAS can be explored at: http://tilatlas.unil.ch

Funding

This research was supported by the Swiss National Science Foundation (SNF) Ambizione [180010 to S.J.C.].

Conflict of Interest: none declared.

Data availability

The data analyzed in this article are publicly available from NCBI Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/ under the identifiers GSE124691 and GSE116390, and from EMBL-EBI ArrayExpress at https://www.ebi.ac.uk/arrayexpress/ under entry E-MTAB-7919.

References

Andreatta,M. *et al.* (2020) Projecting single-cell transcriptomics data onto a reference T cell atlas to interpret immune responses. bioRxiv 2020.06.23.166546; doi: 10.1101/2020.06.23.166546

Carmona,S.J. *et al.* (2020) Deciphering the transcriptomic landscape of tumor-infiltrating CD8 lymphocytes in B16 melanoma tumors with single-cell RNA-Seq. *OncoImmunology*, **9**, 1737369.

Eisenstein,M. (2020) Single-cell RNA-seq analysis software providers scramble to offer solutions. *Nat. Biotechnol.*, **38**, 254–257.

Haghverdi,L. *et al.* (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421–427.

Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.

Luecken,M. *et al.* (2020) Benchmarking atlas-level data integration in single-cell genomics. bioRxiv 2020.05.22.111161; doi: 10.1101/2020.05.22.111161

Magen,A. *et al.* (2019) Single-cell profiling defines transcriptomic signatures specific to tumor-reactive versus virus-responsive CD4+ T cells. *Cell Rep.*, **29**, 3019–3032.e6.

Regev,A. *et al.*(2017) The human cell atlas.eLife Dec 5;6:e27041. doi: 10.7554/eLife.27041.

Stuart,T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.e21.

Tran,H.T.N. *et al.* (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.

Xiong,H. *et al.* (2019) Coexpression of inhibitory receptors enriches for activated and functional CD8+ T cells in murine syngeneic tumor models. *Cancer Immunol. Res.*, **7**, 963–976.