Caltech Division of Biology and Biological Engineering Department of Computing and Mathematical Sciences

Lambda Moses and Lior Pachter
1200 E. California Blvd MC 140-18,
Pasadena, CA 91125, USA
(424) 666-5899
dlu2@caltech.edu

2023-07-31

Dear Dr. Doerr,

We are submitting our article "Voyager: exploratory single-cell genomics data analysis with geospatial statistics" to be considered for publication in Nature Methods. Voyager is an exploratory data analysis (EDA) framework for spatial single-cell 'omics analysis in the vein of packages such as Seurat, scanpy/squidpy, Giotto, and STUtility. However, it introduces several novel ideas, methods, and implementations that we believe will be transformative for users conducting exploratory *spatial* data analysis (ESDA). Specifically, Voyager is unique in providing a bridge between the vast field of geospatial statistics, and spatial single-cell 'omics.

The paper describes several key advances:

1. We describe next-generation infrastructure in R for spatial 'omics via a new S4 class we have developed called SpatialFeatureExperiment (SFE) that frames the spatial 'omics data first and foremost as a geometric map. SFE bundles geometries such as cell segmentations and tissue regions with gene expression and SFE objects can be operated on with the geometries. This package is already approved in Bioconductor, which upholds high standards of robustness and reproducibility. SFE builds on well-established data infrastructure. Hence many existing spatial and non-spatial analysis methods outside Voyager can be applied to SFE objects, as Voyager aims to complement existing methods with the geospatial tradition.
2. We showcase numerous geospatial ESDA methods that have not been previously explored for spatial single-cell 'omics: for example, we show, via analysis of biological data, that *negative* spatial autocorrelation can be identified with single-cell spatial transcriptomics, and has interesting biological implications.
3. We describe a parallel implementation of the R version of Voyager in Python (VoyagerPy), that is concordant with Voyager via automated compatibility testing of core methods. We believe this will represent a major, and transformative advance for the field. We show that results from basic computations such as principal component analysis and log fold change differ substantially between the widely used packages Seurat (in R) and scanpy (in Python). This means that currently, the results users obtain in analyzing data differ depending on the programming language chosen. We not only fix this problem with Voyager, but describe a framework that can be generally adopted to ensure coherence in bioinformatics methodology independent of the programming language chosen.

Primarily, however, Voyager is important because it places spatial information front and center for EDA of spatial, and even non-spatial 'omics data. According to John Tukey who spearheaded EDA, "it is important to understand what you *can do* before you learn to measure how *well* you seem to have *done* it". Methods for specific tasks in spatial -omics data analysis have been previously published in Nature Methods, such as SpatialDE and C-SIDE to identify spatially variable genes. These methods focus on "how *well* you seem to have *done* it", while leaving "what you *can* do" unaddressed. The squidpy EDA framework

Division of Biology and Biological Engineering
Department of Computing and Mathematical
Sciences

Lambda Moses and Lior Pachter
1200 E. California Blvd MC 140-18,
Pasadena, CA 91125, USA
(424) 666-5899
dlu2@caltech.edu

for spatial 'omics, also published on Nature Methods, implements several spatial analysis methods. With a wide range of ESDA methods united by a consistent user interface and a geospatial mindset centering on the spatial information, Voyager allows users to explore more on what they can do as exemplified in the case studies, hence taking squidpy's mission to the next level. Like squidpy, Voyager is extensible, as the ESDA tradition is ever evolving.

Earlier versions of the work have been presented at the BioC 2022 conference as a package demo and as seminars at the Harvard Medical School and Cedars-Sinai Medical Center. An earlier version of this manuscript is a chapter in the first author's PhD thesis. A draft of this manuscript has been posted as a preprint on bioRxiv. No part of this manuscript has been published or considered for publication in any other journal. This work has not been discussed with any Nature Methods editor. We have no conflicts of interest to disclose.

Thank you for your consideration; we look forward to hearing from you.

Sincerely,
Lambda Moses and Lior Pachter