

ComAnDOS: advanced visualizations for cross-species comparisons

Nikolaos Papadopoulos

July 14, 2023

Abstract

Recent experimental progress has made it possible to obtain single-cell transcriptomes for entire animals, reigniting the debate about the definition of cell types and enabling their study in an evolutionary context. The SAMap algorithm has been instrumental in this endeavor, allowing cross-species comparisons without discarding many-to-one orthologous genes. However, SAMap’s lack of downstream analysis and visualization tools effectively requiring advanced programming skills to unlock the full potential of its analysis. Here, we present ComAnDOS (COMparative ANalysis Downstream Of Samap), a Python package that bundles many useful downstream analysis and visualization tools for SAMap. ComAnDOS is freely available at <https://github.com/galicae/comandos>.

1 Introduction

Even before single-cell people had noticed that cells came in groups that looked similar; and that cells in different parts of the body did different things. They guessed that cells that looked similar probably worked in a similar way. It was wild to see that you could find the same cell types (as in: very, very similarly looking) in different parts of the same body, never mind very, very similar cells in different species; even ones as remotely related as humans and jellyfish.

As molecular techniques and knowledge about DNA came along we even figured out that many of the very similar cells used many similar genes; in fact, in many cases even the regulatory apparatus that governed what these cells were supposed to become was shared. This was a big deal, because it meant that possibly the ancestor of these animals, who must have had cell types that did similar things, was already using this regulatory mechanism to build these cell types - maybe things like neurons, vision, muscles, and so on were invented once. The limitation here was that we could only look at a one or maybe a handful of genes at once, so lots of effort had to go into figuring out just one cell type or just one regulatory relationship.

Single-cell RNA sequencing (scRNA-seq) changed all that, because we can now look at the cell type complement of entire organisms. People mostly looked

at human and mouse stuff, because that’s what was easy and medically (financially) interesting, but as time went on and things became cheaper, you could suddenly single-cell sequence entire weird animals. And people did.

And then smart people like Günter and Detlev noticed that there seems to be a hierarchy in cell types - not only in their morphology, location in the body, etc. but also in their gene expression profiles. Duplication and divergence is a very plausible and widespread mechanism for evolution to make new things, so why could it not have been the source of cell type evolution too? They proposed that cooperating groups of transcription factors are the key to cell type evolution, and that these should be conserved between species.

Seeing as we can’t query evolution but rather its phenotypic footprint in extant animals, going from A to B in that theory is not straightforward, but the hope is that by looking at many animals at once we might be able to see the red thread that connects it all. By comparing phenotypic similarity between cell types and seeing who is similar to whom, we hope to figure out which cell types in extant animals used to be the same cell type in their last common ancestor. By seeing which genes are still used by both cell types today, we hope to figure out what function the ancestral cell type was performing.

Originally, people just kept one-to-one orthologs and reduced the scRNA-seq data of both species to that. This has the problem that it assumes too many things - A) that the orthologs are conserved not only in sequence, but also in function, location, timing, and magnitude of expression, and B) that all the subfunctionalisation that happened throughout evolution is basically meaningless and we’ll get the important details despite discarding something like 60-80% of the information. This worked surprisingly well [3], but it was clear that it was not the best way to do it.

This changed with SAMap [2], which takes a BLAST map of genes as input, and takes turns optimizing a cell-cell graph and a gene-gene graph until convergence. SAMap was a game-changer for cross-species comparisons, and is now the de-facto standard for this kind of work. However, SAMap is not without its limitations. The two most important ones are that it is rather hard to install and use (also owing to suboptimal documentation), and that it doesn’t offer intuitive downstream analysis tools. The Sankey plot is fit for infographics but nothing more, and circle plots were never useful for anything in the recorded history of mankind.

2 Advanced visualizations for cross-species comparisons

2.1 Replacing Sankey plots with annotated heatmaps

One of the main outputs of SAMap is a table of mapping scores between query and target clusters. This table naturally lends itself to a heatmap visualization. We provide wrappers around the heatmap functions of the `seaborn` Python library [1] to automate the plotting of annotated heatmaps that can visualize

cell type mapping scores along with a coarse level clustering, such as tissue or cell type family. By visualizing all pairwise cluster mapping scores the hierarchical nature of cell type relationships becomes evident, with cell type families across species clustering together in the heatmap. This is in contrast to the Sankey plot, which only shows the higher mapping scores, implying mostly one-to-one relationships between cell types.

2.2 Visualizing cross-species gene expression

One of the more challenging aspects of cross-species comparisons is finding the conserved gene expression patterns that will lend credence to hypothesized evolutionary relationships between cell types. SAMap proposes pairs of genes that show correlated expression between cell types, but offers no way to visualize this. ComAndOS offers a re-implementation of the popular dotplot visualization that allows plotting the expression of multiple genes for two species at once. The dotplots of the query and target species are arranged to the left and right, accordingly. Furthermore, linking lines can be drawn between genes, visualizing relationships between them, such as homology (e.g. solid lines connecting orthologs and dashed lines connecting paralogs).

2.3 High-quality documentation and modularity for easy extension

The nature of scRNA-seq data analysis in general and cross-species comparisons in particular is exceedingly exploratory. Comparative plots must reflect this exploratory nature, and be easy to repurpose and extend. To this end, the functions in ComAndOS are deliberately kept generic, exposed to the users, and documented extensively. For example, the paired dotplot function can be used to visualize any combination of gene groups, connected in any way desired by the user.

References

- [1] Ekaba Bisong and Ekaba Bisong. Matplotlib and seaborn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 151–165, 2019.
- [2] Alexander J Tarashansky, Jacob M Musser, Margarita Khariton, Pengyang Li, Detlev Arendt, Stephen R Quake, and Bo Wang. Mapping single-cell atlases throughout metazoa unravels cell type evolution. *Elife*, 10:e66747, 2021.
- [3] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.