fcoex: using coexpression to explore cell type diversity in scRNA-seq data

This manuscript ([permalink](#)) was automatically generated from [lubianat/fcoex_report@ce7a723](#) on May 14, 2021.

Authors

- **Tiago Lubiana**

  ⓘ [0000-0003-2473-2313](#) · ⓞ [lubianat](#)

  Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil

- **Helder I Nakaya**

  ⓘ [0000-0001-5297-9108](#)

  Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil

Abstract

Introduction

Single-cell RNA sequencing (scRNA-seq) data analysis is at the core of the current quest to describe all human cell types. [1] The annotation of cell events in scRNA-seq is commonly done by clustering single cells and manually identifying cell types on the dataset. [2]:

Clusters (and their markers) are a prime tool for discovering new groupings with biological relevance. Fine-grained Louvain clustering can highlight out rare, uniform population like the newly identified airway ionocytes. [3] [3] Complementarily, hierarchical clustering provides a multilevel perspective on cell identity, providing knowledge on upper cell classes, prone for ontology building[4])

While such methodologies are already powerful, we identified a gap: current works seldom explore the multi-hierarchy clustering. Biologists are used to tree-like; single-hierarchy classifications, such as the so-called tree-of-life. That tree-like structure rises as a natural side-effect of the macroevolutionary process of vertebrates, where species give rise to one (or more) others. Cell type classifications, however, are functional in essence[5] and, thus, do not need to be tree-like. In fact, formal ontologies of cell types (like the Cell Ontology) catalog many cell types with multiple direct parents. [6] [7]

We build fcoex, an R package that builds coexpression networks as an scaffold for multi-hierarchy classification of cell types, and describe its application to some datasets.

Results

The_fcoex_ method

The_fcoex_ tool was built from first principles to provide better understandability. Our first goal was to develop a smaller set of genes that globally captured the cellular diversity of a dataset. For that, we decided to explore feature selection by symmetrical uncertainty, the correlation metric of FCBF, a popular feature selection algorithm for machine learning with little previous use in biomedical sciences . Symmetrical uncertainty relies on entropy (in the information-theory sense), which relies on categories for calculation. Thus, we implemented a set of heuristics to binarize gene expressions which can be accessed via the_fcoex_::discretize() function.
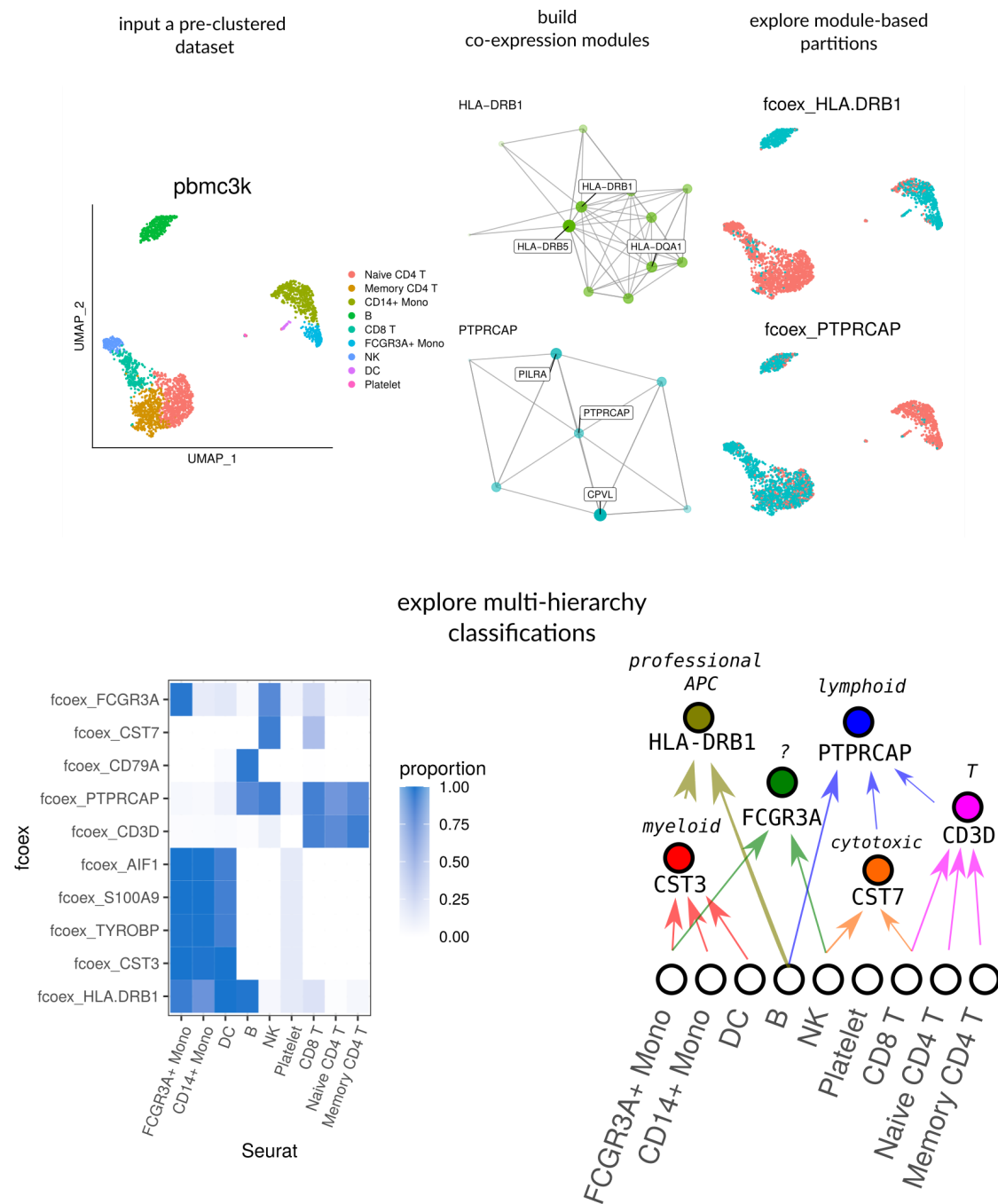
Figure 1: Reclustering based on co-expression modules sheds light on cell diversity. The pre-labeled pbmc3k dataset, with original, SeuratData labels. B)*fcoex* coexpression modules capture patterns across PBMC populations. C-D) Module-based reclustering shows groupings of original clusters, visible both via UMAP projection and cluster x cluster heatmap. E) Upper-level, multilayer classification of the pbmc3k dataset based on_fcoex_ modules.

As mutual information is a supervised method,*fcoex* also needs pre-made cluster assignments obtained after running a standard scRNA-seq clustering pipeline (Fig. 1A). Cluster assignments convey information about the relations between cells and help to guide feature selection.

Fcoex, then, selects genes global markers, which might be specific to 1, 2, or more clusters; the common factor is that they provide information to tell clusters apart. To find the coexpression-module, we inverted the FCBF redundancy removal algorithm as a heuristic to find redundant (co-expressed) gene expression patterns. (see Supplementary Methods for details).

The gene coexpression modules yielded by the pipeline are small by design (10s of genes per module), so to facilitate manual exploration of the coexpression landscape. Each module has one "header" gene, which expression pattern is most representative of the genes in the module.

The ultimate goal of the_fcoex_ pipeline is not necessarily the modules but to find biologically relevant populations._fcoex_ treats each module as a gene set to find cell populations, using only their expression the module to re-classify the cells. The new classifications offer a platform for the biomedical research to gather new insights on the dataset, as exemplified in Figure 1.

fcoex recovers multi-hierarchy of blood types

To validate the_fcoex_ pipeline, we selected the well-known pbmc3k dataset from SeuratData, which contains around 2700 peripheral blood mononuclear cells (PBMC) with previously-defined cluster labels.

The standard_fcoex_ pipeline detected nine modules that capture different parts of the cellular diversity in the dataset. For example, module M8, containing cytotoxicity genes as PRF1 and GZMA, split the dataset into cytotoxic (NK and CD8) and non-cytotoxic cells. M2 (CD3D) split the dataset clearly in T-cells and non-T-cells. M5 (HLA-DRB1) grouped monocytes, B cells, and dendritic cells, all known antigen-presenting cells (APC). (Fig. 1D-E)

In general,*fcoex* clusters combined biologically similar cell types of the original dataset. The clusterings, then, help to explore and classify upper cell classes by function (Fig. 1E). Even in that super well-studied dataset,*fcoex* provided a new light on the shared functionality of some NK cells and macrophages: they both markedly express the CD16-coding gene FCGR3A, whose product is a key player in Antibody-dependent cellular cytotoxicity (ADCC). Thus, a complete functional classification of cells might want to include an 'ADCC-performing cells" class.

Discussion

Here we presented fcoex, a ready-to-use R/Bioconductor package for co-expression-based reclustering of single-cell RNA-seq data. We note that other methods are increasingly available for co-expression analysis of single cells. The monocle R package [8], widely used for pseudotime analysis, has implemented algorithms for detecting co-expression modules (https://cole-trapnell-lab.github.io/monocle3/docs/differential/#gene-modules), and WGCNA, widely used in bulk transcriptomics, has also been applied to scRNA-seq [9] [10].

In principle, any of those algorithms could be used as input for our framework (and we provide code showing how to integrate them to `fcoex`). We note, though, that fcoex modules are generally smaller and provide module header genes, making it a sensible first-pass approach to explore the multi-layered diversity in single-cell transcriptomics datasets. In that way, fcoex offers ways to explore data-driven classifications of cells, aligning itself with the challenges of the Human Cell Atlas and, specifically, of building ontologies of cell types in the single-cell era.

Supplementary Methods

Preprocessing

pbmc3k dataset preprocessing

pbmc3k was loaded as a Seurat object from the SeuratData package. The expression matrix in the "data" slot and the labels in the "Idents" slot as input for creating the *fcoex* object.

Gene expression discretization

As the original Fast Correlation-Based Filter [11] algorithm was constructed to deal with discrete data, we had to discretize gene counts. We chose as a discretization metric a min-max-percent approach. For each gene, we took the lowest and the highest normalized value across the cells. We set a threshold at 25% of the max-min range. All the values below this threshold were considered "OFF," and all above was "ON".

Identification of *fcoex* modules

Filtering genes by correlation to labels

After the discretization step, genes were ranked by their correlation to labels previously assigned by the Seurat Data team. The correlation metric we used was the nonlinear Symmetrical Uncertainty, a variation of mutual information that maps the values between 0 (worst) and 1 (best), and accounts for differences in entropy ranges that arise when variables have a different number of classes (number of labels and number of gene classes). All downstream steps were performed only with the previously filtered genes

Building the coexpression network

Each module M is composed of one module seed (x) predominantly-correlated to the label (L) and all the genes (Yi) more correlated to the seed than to the label. In practice, the algorithm builds an all x all correlation matrix by symmetrical uncetainty, the adjacency matrix of the co-expression network. This adjacency matrix is then trimmed, and edges between nodes Yi and Yj are removed from the network iff SU(Yi, Yj) < SU(Yi, L) or SU(Yi, Yj) < SU(Yj, L). Inferring coexpression modules Modules were built in a bottom-up approach, first selecting genes predominantly correlated to the labels. These genes are the output of the Fast Correlation-Based Filter (FCBF) algorithm for feature selection. Each selected featured forms a new cluster, corresponding to all the genes associated to it in the trimmed network. The modules formed in that way are fuzzy, as a gene might belong to several modules.

Reclustering of cells

To recluster the cells based on each module, we use the "recluster" function of the *fcoex* module. It uses the gene sets in each co-expression community to subset the expression table given originally as input. This reduced table contains the expression values regarding those genes for all the cells in the dataset. The distances between cells in this reduced matrix was calculated by the manhattan distance, and hierarchical clustering was performed. The metric used to calculate the linkage distance between groups was the "ward.D2" metric as implemented in the hclust function of the stats package in R 3.6.1. Two groups of cells were retrieved from each clustering (the k parameter was set to 2) . ## Code availability The *fcoex* package, which performs the coexpression analysis is available at http://bioconductor.org/packages/fcoex/. The discretization and feature selection algorithms are available in a second package, FCBF (http://bioconductor.org/packages/FCBF/). All the analyses performed for this work are available at <https://github.com/lubianat/fcoex paper>.

Acknowledgments

References

1. **The Human Cell Atlas.**
   Aviv Regev, Sarah Teichmann, Eric Lander, Amir Giladi, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna R. Clatworthy, … Human Cell Atlas Meeting Participants
   eLife (2017-12-05) https://www.wikidata.org/wiki/Q46368626
   DOI: 10.7554/elife.27041

2. **Current best practices in single-cell RNA-seq analysis: a tutorial**
   Malte D. Luecken, Fabian J. Theis
   Molecular Systems Biology (2019-06-19) https://www.wikidata.org/wiki/Q64974172
   DOI: 10.15252/msb.20188746

3. **A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte**
   Lindsey W. Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M. Klein, Aron B. Jaffe
   Nature (2018-08-01) https://www.wikidata.org/wiki/Q57318689
   DOI: 10.1038/s41586-018-0394-6

4. **Conserved cell types with divergent features in human versus mouse cortex**
   Rebecca D. Hodge, Rebecca D. Hodge, Trygve E. Bakken, Jeremy A. Miller, Kimberly A. Smith, Eliza R. Barkan, Lucas T. Gray, Jennie L. Close, Brian R. Long, Nelson Johansen, … Ed S. Lein
   Nature (2019-01-01) https://www.wikidata.org/wiki/Q71306466
   DOI: 10.1038/s41586-019-1506-7

5. **What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?**
   Paul Blainey, Hans Clevers, Cole Trapnell, Ed Lein, Emma Lundberg, Alfonso Martinez Arias, Joshua R. Sanes, Jay Shendure, James Eberwine, Junhyong Kim, … Mathias Uhlén
   Cell systems (2017-03-01) https://www.wikidata.org/wiki/Q87649649
   DOI: 10.1016/j.cels.2017.03.006

6. **An ontology for cell types**
   Jonathan Bard, Sue Rhee, Michael Ashburner
   Genome Biology (2005-01-01) https://www.wikidata.org/wiki/Q21184168
   DOI: 10.1186/gb-2005-6-2-r21

7. **A revised airway epithelial hierarchy includes CFTR-expressing ionocytes**
   Daniel T. Montoro, Adam L. Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E. Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, … Jayaraj Rajagopal
   Nature (2018-08-01) https://www.wikidata.org/wiki/Q57318688
   DOI: 10.1038/s41586-018-0393-7

8. **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.**
   Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, John L. Rinn
   Nature Biotechnology (2014-03-23) https://www.wikidata.org/wiki/Q41861103
   DOI: 10.1038/nbt.2859

9. **WGCNA: an R package for weighted correlation network analysis**
   Peter Langfelder, Steve Horvath
   BMC Bioinformatics (2008-01-01) https://www.wikidata.org/wiki/Q21284194
   DOI: 10.1186/1471-2105-9-559

10. **webCEMiTool: Co-expression Modular Analysis Made Easy**
    Lucas E. Cardozo, Pedro S. T. Russo, Bruno Gomes-Correia, Mariana Araujo-Pereira, Gonzalo Sepúlveda-Hermosilla, Vinicius Maracaja-Coutinho, Helder Nakaya
    Frontiers in Genetics (2019-01-01) https://www.wikidata.org/wiki/Q62485122
    DOI: 10.3389/fgene.2019.00146

11. **Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution**
    Lei Yu, Huan Liu
    https://www.wikidata.org/wiki/Q106704674