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

This manuscript ([permalink](#)) was automatically generated from [lubianat/fcoex_report@6fbbf09](#) on May 4, 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 using some variatino of the following steps (Luecken and Theis 2019):

- Cluster cells
- Find clusters markers.
- Annotate clusters based on known markers.

The annotations are reported in legends alongside tSNE and UMAP plots, and in metadata files in columns that give each cell a single type label.

Two of the tasks that scientists are interested in are (1) mapping cells in a dataset to *known* cell types and (2) discover *new* groupings with biological relevance.

By biological relevance, we mean that similar cells might be found in other studies, and help building predictions about reality.

Here we focus on the second task: the discovery of new groups.

Final cluster annotations might point out rare, uniform population like the newly identified airway ionocytes. [2] [2]

Another approach for proposing new classifications is to use hierarchical clustering, and thus provide a multilevel perspective on cell identity. An example is the description of the human middle temporal gyrus by Hodge et al ([3]), where a single-hierarchy ontology is provided both in a main figure and a supplementary file.

While such works are already groundbreaking, we identified a gap: current works seldom explore the multi-hierarchy clustering.

We are used to tree-like classifications, a natural side-effect of the macroevolutionary process of vertebrates. Cell type classifications, though, are not tree-like and many cell types have more than one direct parent. [4] [5]

```
Figure not a tree ?
```

Towards that goal, we build *fcoex*, an R package that builds coexpression networks as an scaffold for hypothesis generation about cell types, and describe its application to some datasets.

Results

The fcoex method

The *fcoex* tool was built bottom up from first principles, with a goal to avoid complicated algorithms and provide better understandability.

Our first goal was to come up with a smaller set of genes that globally captured the cellular diversity of a dataset.

Instead of using simply the highly variable genes, we decided to explore *symmetrical uncertainty*, a different correlation metric.

Symmetrical uncertainty was serendipitously chosen after initial observations of features selected after applying the FCBF algorithm to scRNA-seq data.

FCBF, an acronym for "Fast Correlation Based Filter", is the name of a popular feature selection algorithm for machine learning (over 2700 Google Scholar citations) with little previous use in biomedical sciences (8 PubMed results for "FCBF" as of April 2021).

Symmetrical uncertainty relies on entropy (in the information theory sense), which relies on categories for calculation.

Thus, we implemented in R a set of heuristics to binarize gene expressions (https://bioconductor.org/packages/release/bioc/html/FCBF.html) and allow processing.

Additionally, mutual information is a supervised method, meaning that before using it, we need to have labels for cells. These labels are obtained after running a standard clustering pipeline, such as Seurat's (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html) and convey information about the distribution of cells in the gene expression manifold.

The first step of the pipeline uses a binarized gene expression matrix and preliminary labels to select a set of genes that *globally* separates cells from each other.

These global markers are not necessarily specific to any cluster; they might be specific to multiple clusters, but still provide information to tell them apart.

The selected features, however, share a degree of redundancy: some pairs of genes have virtually identical expression patterns.

The FCBF feature selection is good at finding and removing that redundancy, what is good for some machine learning algorithms.

We inverted the method to capture redundancy instead of removing so to identify gene coexpression modules and implemented the algorithm in the *fcoex* package (see 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. Moreover, 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, though, is not necessarily the modules, but use them to find biologically relevant populations.

Modules contain correlated and anti-correlated genes, and thus might held signatures for two different populations.

To find those populations, fcoex treats each module as a gene set. It then uses only the expression of genes in the module to re-classify the cells.

After projecting the pipelines, we intende to verify if the modules captured different biological functions, and provide complementary views on cell identities to the ones provided by simply applying the Seurat clustering pipeline.
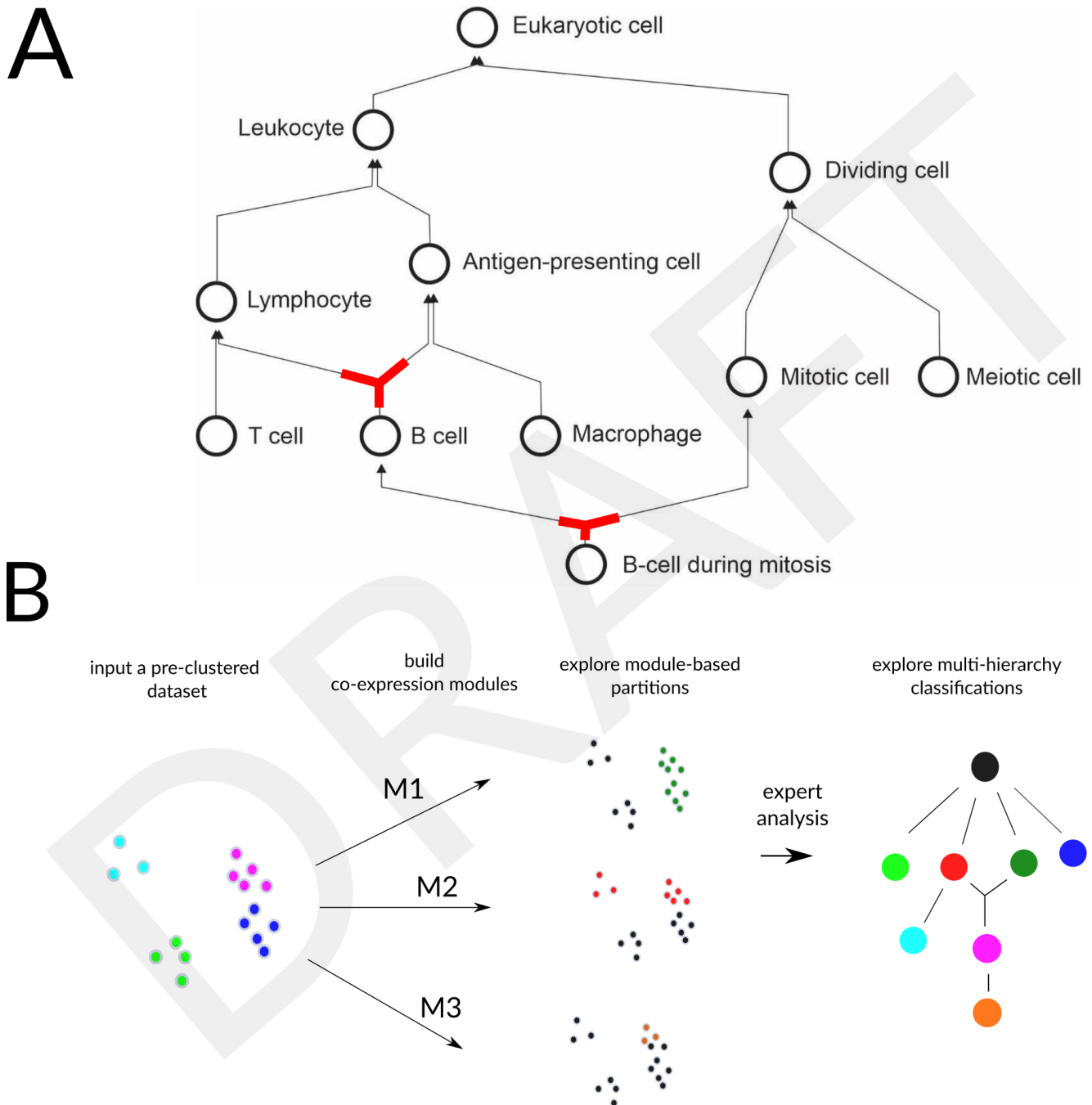

Figure 1: Overview

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 author-defined cluster labels (Figure 2A). The standard fcoex pipeline detected nine modules which, reassureingly, captured transcriptional pathways known to be active in different blood cells (Table A, SUPP FIGURE PBMC 3K). For example, module M8 contained cytotoxicity genes, as PRF1 and GZMA. he reclustering based on M8 split the dataset into cytotoxic (NK + CD8) and non-cytotoxic cells (Figure 2 B-C). M7 (CD79A)

cluster corresponded to B cells, while M5 (HLA-DRB1) grouped monocytes, B cells, and dendritic cells, all known antigen-presenting cells (APC) (https://www.ebi.ac.uk/ols/ontologies/cl/terms?obo_id=CL:0000145). Of note, the APC populations are apart in the UMAP (Uniform Manifold Approximation and Projection for dimensional reduction)(McInnes et al. 2018) representation (Figure 2D), even though the relevance of the class is known acknowledged similarity might have gone unnoticed in the single-cell analysis without fcoex. In general, fcoex clusters are combinations of similar cell types of the original division (Figure 2E), one of which corresponds to a known cell class and the other to the complementary set of cells in the tissue.

fcoex uncovers unknown populations in the zebrafish embryo

After the proof of concept, e explored gene-gene interactions in more depth in a gastrulation dataset of zebrafish cells (75% epiboly, Figure 3 A)(Farrell et al. 2018). The 10 clusters fed to fcoex, though not corresponding to classic types, fed the algorithm with structure from the gene expression space. Using that structure, fcoex identified 8 modules (using default parameters) and the list of modules is displayed in Table B.

The two top ranked modules offer gateways into exploring of zebrafish development biology.

The top-ranked module M1 (MSGN1) harbored the genes msgn1, tbx16, and tbx6, all related to mesoderm development, a core task of gastrulation. In mice, Msgn1 signals via Tbx16(Chalamalasetty et al. 2014) and tbx6l and tbx16 (homologs of Tbx16) play a role together in shaping mesoderm development in zebrafish (Morrow et al. 2017). The population described by this module might be of interest for understanding how mesoderm unfolds.
The second-best modules M2 presented a full pathway - a ligand, apela/Toddler, its receptors, aplnrb, and aplnra, and a putative downstream factor, mespab ((Pauli et al. 2014)(Deshwar et al. 2016). The module contains anticorrelated genes, and the ligand and its receptors are enriched in opposing clusters (TableD, Figure 3C-E).

Discussion

Here we report a pipeline to explore cell type divesity via ranked, lean coexpression modules.

The multilayered view on gene expression is the core conttribution of fcoex to the arsenal of tools.

We note that other methods are increasingly available for co-expression analysis of single cells.

The popular frameworks Scanpy(https://scanpy.readthedocs.io/en/stable/tutorials.html#, https://github.com/theislab/scanpy/issues/72), Seurat(https://satijalab.org/seurat/) and OSCA (from Bioconductor, https://bioconductor.org/books/release/OSCA/#) do not include specific instructions on using co-expressions works or providing multi-layer clustering, but the monocle R package (https://www.nature.com/articles/nbt.2859), widely used for pseudotime analysis, has implemented module-based analysis for its version 3, using the Louvain community detection algorithm to get modules from UMAP dimensions.(https://cole-trapnell-lab.github.io/monocle3/docs/differential/#gene-modules).

A common option is to employ the WGCNA method [6] or its adaptations [7] to identify co-expression modules.

In principle, any of those modules could be used as input of multilayered clustering, and are alternatives for the first step of the fcoex pipeline. We note, though, that in contrast to unranked, large modules found by previous pipelines , fcoex modules makes it easy for domain experts to take

advantage of the data. For example, in this work, a master's student could pinpoint the 4 genes of the apela/Toddler pathway floating on an 18-gene gene set. He would surely have had more trouble finding a biologically meaningful association in a set with hundreds of genes.

Fcoex streamlines the path from dataset to coexpression modules and to multi layered populations.

The "cell type forests" obtained with the parallel reclustering of cells sheds light at higher orders of classifications.

The approach supported our "rediscovery" of B-cells as both a cell type on their own, but also as a subclass of Antigen Presenting Cells.

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. (ADD REFERENCES)

Methods

Data

The pbmc3k dataset version 3.0.0 was downloaded via the SeuratData package (https://github.com/satijalab/seurat-data). The zebrafish development dataset was downloaded from the Broad Single Cell portal (https://singlecell.broadinstitute.org/single_cell/study/SCP162).

Preprocessing

pbmc3k dataset preprocessing

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

Zebrafish dataset preprocessing

The data downloaded as log2 transformed counts were loaded in a Seurat object and preprocessed as in Seurat's 3.0 default tutorial. The resolution of the "FindClusters" function was arbitrarily set to yield 10-20 clusters. The normalized expression matrices, and the labels from the FindClusters function were used as input for the fcoex expression. Precise descriptions of the settings used are available in the source code for this paper on https://github.com/lubianat/fcoex_paper.

Gene expression discretization

As the original Fast Correlation-Based Filter algorithm was constructed to deal with discrete data, we had to discretize gene counts. This was done with the fcoex package, version 1.0.0 (https://bioconductor.org/packages/fcoex/) 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 Seurat's 3.1 FindClusters function). 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

Finding predominantly-correlated module seeds

Modules were built in a bottom-up approach, first selecting genes predominantly correlated to the labels - a higher symmetrical uncertainty score towards the labels than towards any other gene. These genes, that are the output of the Fast Correlation-Based Filter algorithm, are called the module seeds.

Building the coexpression modules/communities

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 other words, a gene Yi from all the genes in the Y universe of all genes in the dataset belongs to a module M headed by gene x if and only if it is more correlated to a gene x (from the set X of module seeds) than to the labels.

```
In practice, the algorithm builds an all x all correlation matrix, 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).
```

1.1 Over-representation analysis We performed an over-representation analysis on the human PBMC dataset by Reactome Pathway gene sets processed locally prior to data analysis ("Reactome - a Curated Knowledgebase of Biological Pathways," n.d.). Visualizations in the fcoex package were adapted from the CEMiTool R package (Russo et al. 2018)

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) . The cluster with a higher expression of the module
seed was labeled as SP (Seed Positive) and the complementary cluster
received, then, the label SN (Seed Negative). Plots were generated via the
DimPlot function of the Seurat package, substituting labels of the Seurat
object for the new ones.
```

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. **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

3. **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

4. **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

5. **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

6. **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

7. **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