tenXplore: ontology for scRNA-seq, applied to 10x 1.3 million neurons

Vincent J. Carey, stvjc at channing.harvard.edu

October 30, 2017

Contents

1 Introduction/Executive summary

The tenXplore package includes prototypical code to facilitate the coding of an ontology-driven visualizer of transcriptomic patterns in single-cell RNA-seq studies.

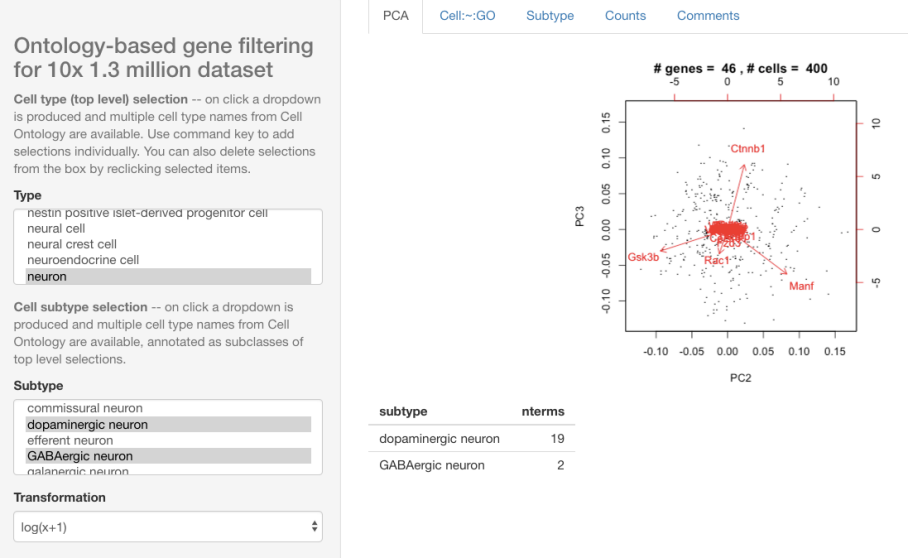

Figure 1: dashsnap

This package is intended to illustrate the role of formal ontology in the analysis of single-cell RNA-seq experiments.

We anticipate that both unsupervised and supervised methods of statistical learning will be useful.

- The 10x 1.3 million neuron dataset primarily invites unsupervised analysis, as no anatomical or other structural details are provided along with single-cell expression measures.
- Other studies that include details on anatomic source or other non-transcriptomic features of cellular identity would be amenable to supervised analysis, which will benefit from standardization of metadata about experimental factors and anatomic structures.

1.1 A challenge: finding expression signatures of anatomic structures or cell types

Gene Ontology and Gene Ontology Annotation are the primary resources at present for enumerating gene signatures for cell types. The *ontoProc* package assists in providing some mappings, but much work is needed.

```
library(ontoProc)
## Loading required package: ontologyIndex
data(allGOterms)
cellTypeToGO("serotonergic neuron", gotab=allGOterms)
```

```
##             GOID                                        TERM
## 18623 GO:0036515           serotonergic neuron axon guidance
## 18625 GO:0036517 chemoattraction of serotonergic neuron axon
## 18627 GO:0036519  chemorepulsion of serotonergic neuron axon
cellTypeToGenes("serotonergic neuron", gotab=allGOterms, orgDb=org.Mm.eg.db)
## 'select()' returned 1:many mapping between keys and columns
##           GO EVIDENCE ONTOLOGY            ENSEMBL SYMBOL
## 1 GO:0036515      IMP       BP ENSMUSG00000007989   Fzd3
## 2 GO:0036515      IMP       BP ENSMUSG00000026556 Vangl2
## 3 GO:0036515      IMP       BP ENSMUSG00000023473 Celsr3
## 4 GO:0036515      IMP       BP ENSMUSG00000107269 Celsr3
## 5 GO:0036517      IDA       BP ENSMUSG00000021994  Wnt5a
cellTypeToGenes("serotonergic neuron", gotab=allGOterms, orgDb=org.Hs.eg.db)
## 'select()' returned 1:many mapping between keys and columns
##           GO EVIDENCE ONTOLOGY         ENSEMBL SYMBOL
## 1 GO:0036515      IEA       BP ENSG00000008300 CELSR3
## 2 GO:0036515      IEA       BP ENSG00000104290   FZD3
## 3 GO:0036515      IEA       BP ENSG00000162738 VANGL2
## 4 GO:0036517      IEA       BP ENSG00000114251  WNT5A
```

1.2 Discrimination of neuron types: exploratory multivariate analysis

At this point the API for selecting cell types, bridging to gene sets, and acquiring expression data, is not well-modularized, but extensive activity in these areas in new Bioconductor packages will foster enhancements for this application.

In brief, we often fail to find GO terms that approximately match, as strings, Cell Ontology terms corresponding to cell types and subtypes. Thus the cell type to gene mapping is very spotty and the app has nothing to show.

On the other hand, if we match on cell types, we get very large numbers of matches, which will need to be filtered. We will introduce tools for generating additional type to gene harvesting/filtering in real time.

2 Next steps

The *ontoProc* package includes facilities for working with a variety of ontologies, and tenXplore will evolve to improve flexibility of selections and visualizations afforded by the *ontologyIndex* and *ontologyPlot* packages.