ontoProc: RDF ontology processing for Bioconductor

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

October 30, 2017

Contents

1 Executive summary

The ontoProc package was developed to facilitate the coding of an ontology-driven visualizer of transcriptomic patterns in single-cell RNA-seq studies (tenXplore).

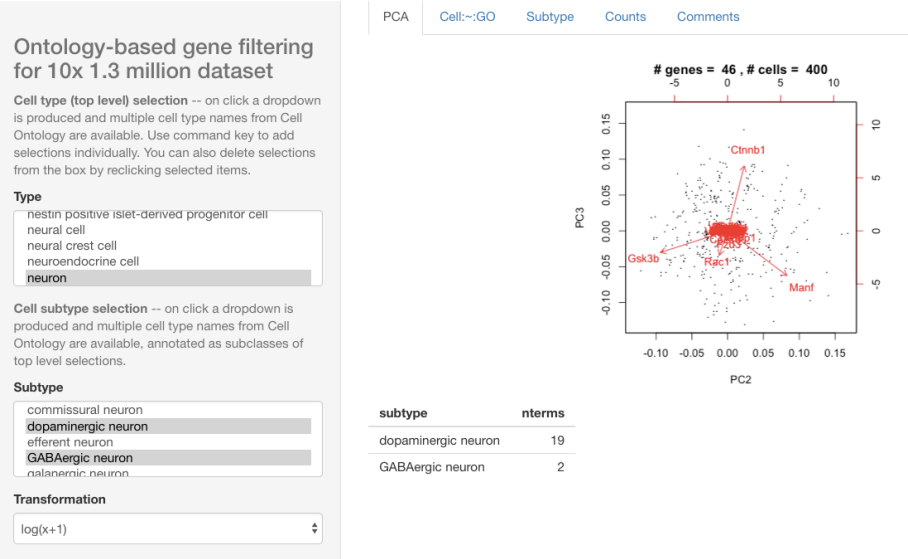

Figure 1: dashsnap

2 Introduction

Our primary objective is facilitating use of ontological metadata to simplify construction of formally annotated hierarchies of samples or features that should be traversed in analysis of complex genomic experiments.

2.1 An enumeration of cell types

We used the Experimental Factor Ontology 'cell type' class (EFO_0000324) to obtain an enumeration of cell types. As of August 22 2017 it is an open question whether Cell Ontology should be used for this purpose. The author's subjective impression is that EFO has a simpler collection of terms for cell types, while Cell Ontology has a better collection of terms for types of neurons.

2.2 Basic operations using ontologyIndex facilities

This package ships with an R serialization of an OBO representation of the Cell Ontology. This is created using `get_OBO` in *ontologyIndex*. (For ontologies only available in OWL format, the python pronto package was used to convert to OBO.)

```
library(ontoProc)
cellOnto = getCellOnto()
cellOnto
## Ontology with 6546 terms
##
##
##
## Properties:
##  id: character
##  name: character
##  parents: list
##  children: list
##  ancestors: list
##  obsolete: logical
## Roots:
##  GO:0008150 - biological_process
##  GO:0005575 - cellular_component
##  UBERON:0001062 - anatomical entity
##  GO:0003674 - molecular_function
##  BFO:0000002 - continuant
##  PATO:0000001 - quality
##  BFO:0000003 - occurrent
##  NCBITaxon:1 - root
##  PR:000018263 - amino acid chain
##  PR:000021082 - bone marrow proteoglycan proteolytic cleavage product
##  ... 113 more
```

At this time, elementary manipulations of the ontology involve collecting the children, siblings, or labels for given URIs.

```
cochil = children_TAG("CL:0000540", cellOnto)
cochil
## TermSet for 34 terms
## GABAergic neuron, adrenergic neuron, ..., spiral ganglion neuron, unipolar neuron
label_TAG("CL:0000540", cellOnto)
## CL:0000540
##   "neuron"
siblings_TAG("CL:0000540", cellOnto)
## TermSet for 22 terms
## abnormal cell, basal cell of olfactory epithelium, ..., retinal cell, smooth muscle cell of the brain vas
```

3 Application: finding genes annotated to neuron subtypes

We focus on mouse. The neuron subtypes identified as OWL subclasses of "neuron" have names

```
cleanNames = function(tset) {
 slot(tset,"cleanFrame")$clean
}
cleanNames(cochil)
##                          CL:0000617                          CL:0000109
##                 "GABAergic neuron"                "adrenergic neuron"
##                          CL:0000526                          CL:0010022
##                  "afferent neuron"                   "cardiac neuron"
##                          CL:2000029                          CL:0000108
##    "central nervous system neuron"               "cholinergic neuron"
##                          CL:0000112                          CL:0000678
##                  "columnar neuron"               "commissural neuron"
##                          CL:0000700                          CL:0000527
##             "dopaminergic neuron"                  "efferent neuron"
##                          CL:0011100                          CL:0000679
##                "galanergic neuron"             "glutamatergic neuron"
##                          CL:1001509                          CL:0011110
##               "glycinergic neuron"             "histaminergic neuron"
##                          CL:0011109                          CL:0000099
##      "hypocretin-secreting neuron"                       "interneuron"
##                          CL:1000606                          CL:2000031
##                "kidney nerve cell"     "lateral line ganglion neuron"
##                          CL:0000104                          CL:0000029
##                "multipolar neuron"      "neuron neural crest derived"
##                          CL:0000528                          CL:0008025
##                 "nitrergic neuron"             "noradrenergic neuron"
##                          CL:0000110                          CL:2000032
##               "peptidergic neuron" "peripheral nervous system neuron"
##                          CL:0000111                          CL:0000116
##                "peripheral neuron"                   "pioneer neuron"
##                          CL:0000102                          CL:0000530
##                 "polymodal neuron"                   "primary neuron"
##                          CL:0000105                          CL:0000535
##            "pseudounipolar neuron"                 "secondary neuron"
##                          CL:0000379                          CL:0000850
##        "sensory processing neuron"              "serotonergic neuron"
##                          CL:0011113                          CL:0000106
##           "spiral ganglion neuron"                  "unipolar neuron"
```

We would like to see if the expression data would allow us to discriminate neurons of these different types.

3.1 Bridging from Cell Ontology to mouse genes

There is no formal linkage at present between terms of Cell Ontology and those of Gene Ontology. Research on inference of tissue of origin from expression signatures has led to accurate classifiers (Lee, Krishnan, Troyanskaya) and applications in cell mixture deconvolution (Houseman).

Formal work in ontology bridging has been described but the specific task of mapping from Cell Ontology terms to Gene Ontology terms has not culminated in any programmatically available resource.

We apply approximate pattern matching (agrep in R) to find gene ontology terms that are apparently relevant to cell type vocabulary terms of interest. These are then mapped to gene annotation. Simple (non-vectorized) functions that accomplish this in an organism-specific are straightforward using the OrgDb packages. We serialized all GO terms for convenience with this package, in the data object `allGOterms`.

```
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", orgDb=org.Mm.eg.db, gotab=allGOterms)
## '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", orgDb=org.Hs.eg.db, gotab=allGOterms)
## '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
```

3.2 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. Thus the best ways to get a feel for it are to use tenXplore() function, and to read the source code. In brief, we often fail to find GO terms that approximately match, as strings, Cell Ontology terms corresponding to cell subtypes. On the other hand, if we match on cell types, we get very large numbers of matches, which, at this time, will need to be filtered to get manageable feature sets. We will introduce tools for generating additional RDF to improve gene harvesting in real time. But the associated statements will need to be curated. The EBI Webulous system should be useful for introducing new terms that facilitate better connections between anatomic structures and sets of genes or other genomic features.