flowCL: ontology-based cell population labelling in flow cytometry Query examples

October 30, 2014

1 Create an archive

A pre-loaded archive can be created from a data file in ***flowCL***. (This step can be skipped). To load this archive, enter the following:

```
> flowCL("Archive")
```

2 Querying

flowCL requires an input string representing the markers composing the immunophenotype of the cell type (obtained by manual or automated methods). It then parses and translates this input into appropriate SPARQL queries which are then executed against the CL.

2.1 Basic query

This example illustrates the input, parsing, translation and querying in the simple case in which only one marker constitutes the immunophenotype. In an R console, a user enters:

```
> Res <- flowCL("CCR7-")
```

flowCL will parse this string into the marker "CCR7" and its abundance tag "-" . It will then build the appropriate SPARQL queries and execute them sequentially.

The first step is to retrieve the identifier of the CCR7 marker (including synonym lookup) in the CL:

```
# Common prefix and abbreviation
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix obo: <http://purl.obolibrary.org/obo/>
prefix oboinowl: <http://www.geneontology.org/formats/oboInOwl#>
```

```
select distinct ?x ?label ?synonym
where
        {
        ?x a owl:Class.
        ?x rdfs:label ?label.
        ?x oboinowl:hasExactSynonym ?synonym.
        FILTER regex(?synonym, "^CCR7$", "i")
        }
```

The marker label and the marker ID are both returned:

```
C-C chemokine receptor type 7
obo:PR_000001203
```

flowCL will then use this result to query the CL for all the labels of all cell types that contain that marker:

```
select distinct ?x ?celllabel ?plabel ?marker ?markerlabel
        where
        {
        ?x a owl:Class.
        ?x rdfs:label ?celllabel.
        %?x rdfs:subClassOf ?sub.
        ?x sesame:directSubClassOf ?sub.
        ?sub rdf:type owl:Restriction.
        ?sub owl:onProperty lacks_pmp:.
        ?sub owl:someValuesFrom ?marker.
        ?marker rdfs:label ?markerlabel.
        lacks_pmp: rdfs:label ?plabel.
        FILTER regex(?markerlabel, "C-C chemokine receptor type 7", "i")
        }
```

When the immunophenotype queried consists of only one maker, the sesame:directSubClassOf property is used in the SPARQL query to return only the top-level classes containing that marker rather than the full hierarchy. As all subclasses would inherit this marker, this would significantly increase the computational time required by flowCL without providing useful information to the user. In all other cases (i.e., when multiple makers are queried), the rdfs:subClassOf property is used. ***flowCL*** will then list all the results from the CL, which represent those cell types that contain the queried marker. Each cell type in the list is then ranked, explained below, and the highest ranked cell types are returned.

2.2 Complex query with exact match

In this example, the immunophenotype to be queried for is "CD3+CD4+CD8-CCR7-CD45RA+". The user can invoke the flowCL tool using:

```
> Res <- flowCL("CD3+CD4+CD8-CCR7-CD45RA+")
```

A tree diagram can be obtained by plotting the result in the R console:

```
> tmp <- Res$'CD3+CD4+CD8-CCR7-CD45RA+'
> plot(tmp[[1]], nodeAttrs=tmp[[2]], edgeAttrs=tmp[[3]], attrs=tmp[[4]])
```

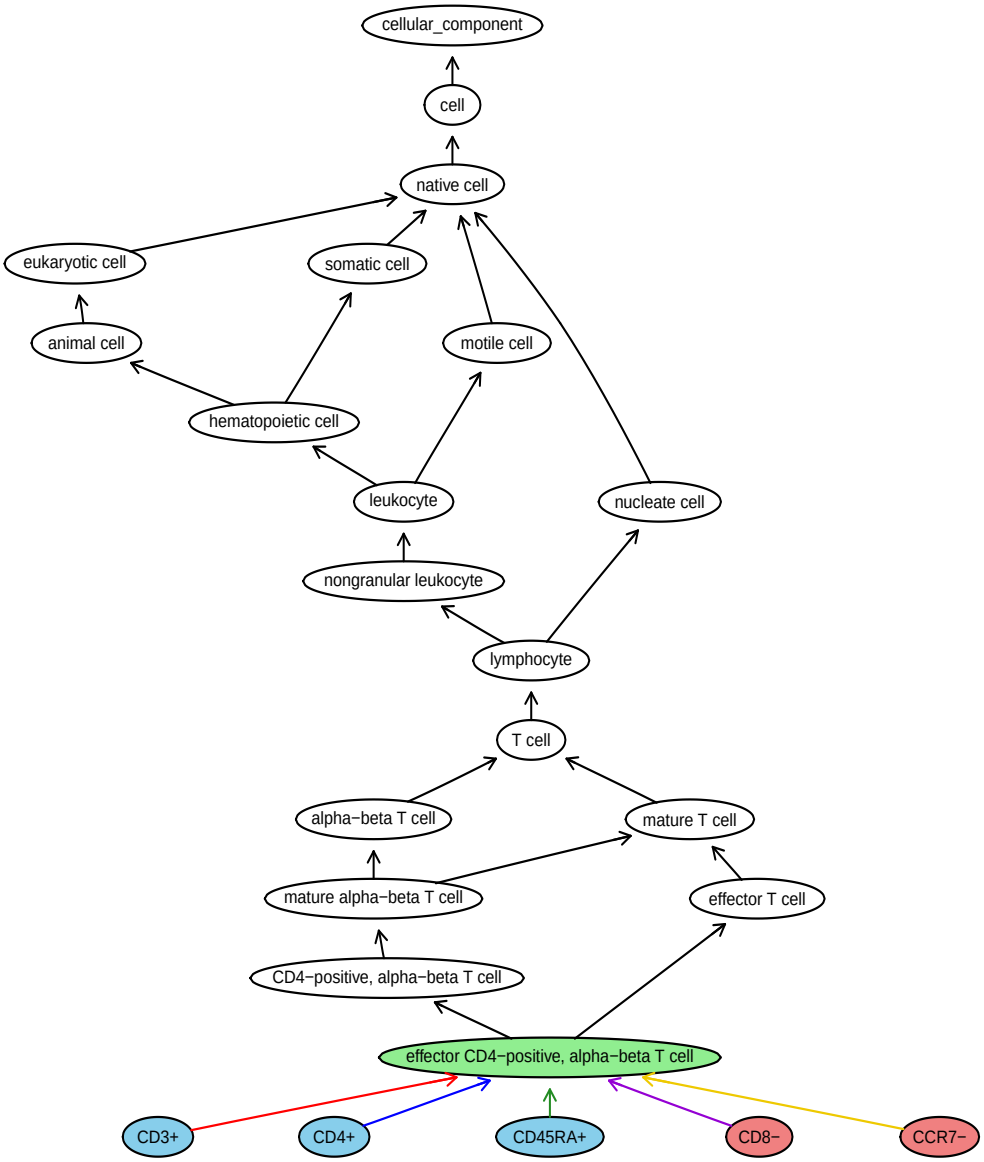

Figure 1: Output tree diagram representing the cell hierarchy when querying with immunophenotype: "CD3+CD4+CD8-CCR7-CD45RA+"

where tmp is the result array returned by flowCL, which is used only for creating the tree diagram: tmp[1] contains objects from the Rgraphviz package that produce the nodes, tmp[2] is the colour of the nodes, tmp[3] is the colour of the arrows and tmp[4] is the font size, node shape and tree orientation.

Figure 1 shows the tree diagram reflecting the cell hierarchy that is dependent on all of CD3+, CD4+, CD8-, CCR7- and CD45RA+ which is produced as output. One cell type is an exact match, i.e., contains all five markers queried. Black arrows indicate the subclass relationship (ex. native cell "is a" cell). Coloured arrows are the inverse of *has/lacks plasma membrane part* or *has high/low plasma membrane amount* (ex. effector CD4-positive, alpha-beta T cell *has plasma membrane part* CD45RA). Each marker is associated with its own coloured arrow, which makes it easier for the user to tell which cell type contains which markers. The nodes can also have different colours. Blue nodes identify the *has plasma membrane part* markers, while red nodes correspond to the *lacks plasma membrane part* markers. Other colours are possible (not represented in Figure 1): navy blue nodes for *has high plasma membrane amount* markers, grey blue nodes for *has low plasma membrane amount* markers. A green node (see Figure 1) is an exact match, while the beige nodes indicate partial matches (see Figure 2).

2.3 Additional information

Information about the query for the cell labels returned, markers matched and rankings is stored in Res and can be viewed calling:

```
> Res$Ranking$'CD3+CD4+CD8-CCR7-CD45RA+'

[1] 1

> Res$Cell_Labels$'CD3+CD4+CD8-CCR7-CD45RA+'

[[1]]
[1] "effector CD4-positive, alpha-beta T cell"

> Res$Markers$'CD3+CD4+CD8-CCR7-CD45RA+'

[[1]]
[1] "{  } CD8-, CCR7-, CD4+, CD3+, CD45RA+ (  ) [  ]"

>
```

2.3.1 Cell Labels

The $Cell_Labels element lists the labels of the cell types found in CL in order of highest score based on their ranking.

2.3.2 Markers

The $Markers element lists all markers that were queried as well as markers that are used to define the certain cell types. Markers are displayed in the form of { **A** } **B** (**C**) [**D**].
A and **B** together make up the markers input for the query.

- **B** are all the markers that are in both the query and the cell type definition,

- **A** lists all the markers in the input for the query that were not required for that particular cell type.

B, **C** and **D** together are the markers that make up the definition of the particular cell type.

- **C** lists the markers that were part of the experiment that were not part of the query,
- **D** lists all other markers that make up the cell type that were not part of the experimental markers.

Additional information, such as cell ID, marker ID or ontology marker names can be viewed using:

```
> Res$Table
```

```
                      [,1]
Short marker names    "CD3+CD4+CD8-CCR7-CD45RA+"
Ontology marker names "alpha-beta T cell receptor complex, CD4 molecule, T cell receptor"
                      "co-receptor CD8, C-C chemokine receptor type 7, receptor-type"
                      "tyrosine-protein phosphatase C isoform CD45RA"
Experiment markers    "CD3,CD4,CD8,CCR7,CD45RA"
Ontology exper. names "alpha-beta T cell receptor complex, CD4 molecule, T cell receptor"
                      "co-receptor CD8, C-C chemokine receptor type 7, receptor-type"
                      "tyrosine-protein phosphatase C isoform CD45RA"
Successful Match?     "Yes"
Marker ID             "1) PR_000001004, PR_000001015, PR_000025402, PR_000001203, GO_0042105"
Marker Label          "1) alpha-beta T cell receptor complex, CD4 molecule, receptor-type"
                      "tyrosine-protein phosphatase C isoform CD45RA, T cell receptor"
                      "co-receptor CD8, C-C chemokine receptor type 7"
Marker Key            "1) { } CD8-, CCR7-, CD4+, CD3+, CD45RA+ ( ) [ ]"
Score (Out of 1)      "1) 1"
Cell ID               "1) CL_0001044"
Cell Label            "1) effector CD4-positive, alpha-beta T cell"
```

2.3.3 Ranking scores

The $Ranking element returns the ranking of each cell label returned (Score out of 1). Scores are calculated by adding all the markers that were queried that are also part of the returned cell type, potentially subtracted by a penalty, and divided by the number of markers that define that cell type:

$$((\mathbf{B} - 2 * \mathbf{A})/(\mathbf{B} + \mathbf{C} + \mathbf{D}))$$

A value of -2 for the penalty was chosen to penalize cell types that had markers in section **A** heavily: section A indicates those markers that were explicitly queried for but are not present in the returned cell type. It becomes apparent why -2 is a better choice than -1 when there are two possible cell types; one having 1 marker in **A**, four markers in **B** and one marker in **D**, while the other has three markers in **B** and three markers in **D**. With a penalty of -2 the first possible cell type has a score of (4-2)/5 = 0.4 while the second one has a score of (3-0)/6 = 0.5. If the penalty was -1, the first possible cell type would have a score of 0.6 and would be ranked higher than the second one. This is frowned upon because queried markers that are in **A** are more incorrect than markers that were left out of the query. The -2 penalty has been empirically determined and worked successfully for all the

HIPC phenotypes, for which the CL includes the immunophenotypes. As part of our future work we plan to augment the CL with an assay-based representation of cell types, thereby ensuring that the number of markers queried for correlates with the number of markers present in the CL.

2.4 Complex query with non-exact match

A query for the immunophenotype "CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi"' can be run using:

```
> Res <- flowCL("CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi")
> tmp <- Res$'CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi'
> plot(tmp[[1]], nodeAttrs=tmp[[2]], edgeAttrs=tmp[[3]], attrs=tmp[[4]])
> Res$Cell_Labels$'CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi'

[[1]]
[1] "CD16-negative, CD56-bright natural killer cell"

[[2]]
[1] "decidual natural killer cell"

> Res$Ranking$'CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi'

[1] 0.8750000 0.5714286

> Res$Markers$'CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi'

[[1]]
[1] "{  } CD19-, CD3-, CD20-, CD14-, CD16-, HLA-DR-, CD56hi (  ) [ SLAM family member 5+ ]"

[[2]]
[1] "{ HLA-DR- } CD19-, CD3-, CD20-, CD14-, CD16-, CD56hi (  ) [ galectin-1+ ]"
```

The query above returns two cell types called "CD16-negative, CD56-bright natural killer cell" and "decidual natural killer cell". The ranking score for both of these cell types are (7-0)/8 = 0.8750000 and (6-2)/7= 0.5714286, respectively using the formula in subsection 2.3.3.

The resulting output tree diagram in Figure 2 shows that no exact matches were identified (no green nodes). The presence of cell type with a "hi" marker is depicted as a navy blue node. Non-exact matches are represented by beige nodes.

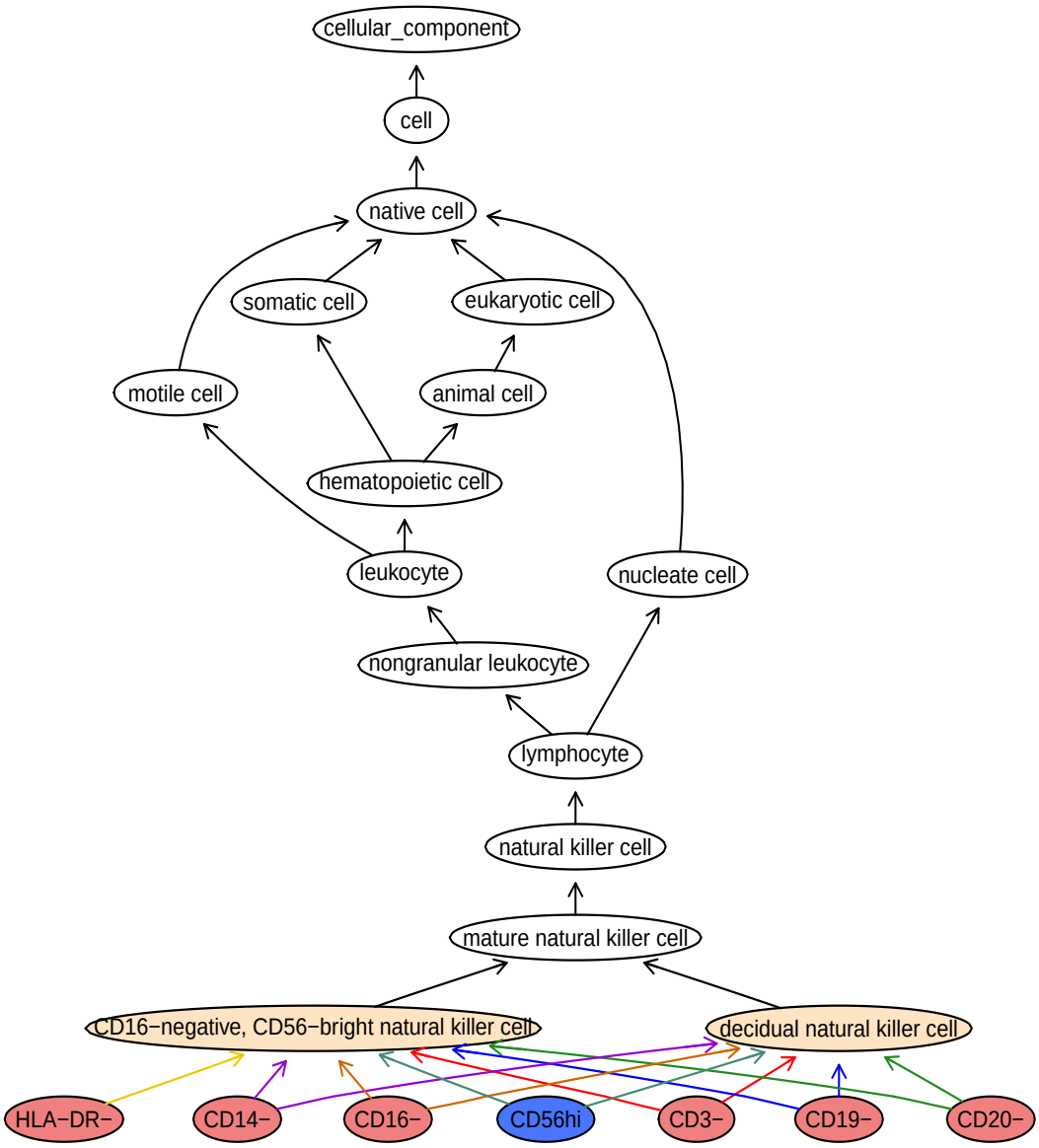

Figure 2: Output tree diagram of the cell hierarchy when querying with immunophenotype: "CD3-CD19-CD20-CD14-HLA-DR-CD16-CD56hi"