flowCL: Semantic labelling of flow cytometric cell populations

Justin Meskas

April 12, 2014

jmeskas@bccrc.ca

Contents

1 Licensing

Under the Artistic License, you are free to use and redistribute this software.

2 Loading the Library

To install ***flowCL*** type *source("http://bioconductor.org/biocLite.R")* and then type *biocLite("flowCL")* into R. For more information on installation guidelines see the Bioconductor and the CRAN websites.

Once installed, to load the library, type the following into R:

```
> library("flowCL")
```

3 Running *flowCL*

3.1 Introduction

Given a cell type, one can use a cell ontology to discover which markers are used to uniquely identify that particular cell type. In the reverse situation, how can one find an appropriate cell type with a given phenotype (or list of markers)? ***flowCL*** matches a phenotype to a cell type from the cell ontology. If the match is not unique, then the best alternative is returned. Markers are input followed with a "+" or a "-". The "+" stands for "has plasma membrane part", while the "-" stands for "lacks plasma membrane part". For example, "CD8+" will search for cell types that have a plasma membrane part of CD8.

flowCL executes queries against the Cell Ontology (CL), available at http://cellontology.org. The CL file is hosted on a triplestore, i.e., a database for storage and retrieval of Resource Description Framework (RDF) triples. The SPARQL endpoint at http://cell.ctde.net:8080/openrdf-sesame/repositories/CL is used to execute the SPARQL queries retrieving the correct matches from the CL. While other SPARQL endpoints can be used, users should be aware that in our case the CL file has been reasoned upon, and resulting extra inferred axioms have been added to the triplestore, providing a more complete result set.

3.2 Archive

A folder called "flowCL_results" will be created in the current directory. In this folder a file called "listPhenotypes.csv" is created. This file lists the results of ***flowCL***, and is the same as what is returned from the function in $Table. In addition, the tree diagrams are created and saved as ".pdf" files and these show the cell dependency of the matched cell types. The code will slowly build an archive of information in "[current directory]/flowCL_results/" inside the folders of "parents", "parents_query" and "results". Accessing this archive is much faster than querying every time, however, if the ontology gets updated the code will use the now old data from these folders. To make sure the code is as up to date as possible the archive should be reset every once in a while.

For the sake of time, a pre-loaded archive can be created from two data files in ***flowCL***. To load this archive, enter the following:

```
> data(Parents_query_archive, Parents_Names)
> dir.create ( paste(getwd(),"/flowCL_results/parents_query",sep=""),
```

```
+     showWarnings=FALSE, recursive=TRUE )
> for (j in 1:length(Parents_Names))
+     write.table(Parents_query_archive[[j]],paste(getwd(),"/flowCL_results/parents_query/",
+     Parents_Names[[j]], sep=""), sep=",", row.names = FALSE)
```

This pre-loaded archive has the possibility of being outdated. This archive will be updated by the maintainer, along with the vignette, whenever the ontology is updated. More discussion of this is in the "Last Comments" section. The current version was updated last on March 18^{th} 2014. To check that the version is up to date run:

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

If the returned date matches the one above, then the pre-loaded archive is the correct one. Please note that the user can skip this section altogether and it will always give the most current archive. The down side is the code will take up to 40 minutes to run the first query because it must build an archive.

3.3 Simple Example

The simplest example of ***flowCL*** is to enter in one phenotype:

```
> Res <- flowCL("CCR7+CD45RA+")
> Res$Table
```

```
                      [,1]
Short marker names    "CCR7+CD45RA+"
Ontology marker names "C-C chemokine receptor type 7, receptor-type tyrosine-protein"
                      "phosphatase C isoform CD45RA"
Successful Match?     "Yes - 5 hits"
Marker ID             "1) PR_000001203, PR_000001015 2) PR_000001203, PR_000001015 3)"
                      "PR_000001203, PR_000001015 4) PR_000001203, PR_000001015 5)"
                      "PR_000001203, PR_000001015"
Marker Label          "1) C-C chemokine receptor type 7, receptor-type tyrosine-protein"
                      "phosphatase C isoform CD45RA 2) C-C chemokine receptor type 7,"
                      "receptor-type tyrosine-protein phosphatase C isoform CD45RA 3) C-C"
                      "chemokine receptor type 7, receptor-type tyrosine-protein phosphatase"
                      "C isoform CD45RA 4) C-C chemokine receptor type 7, receptor-type"
                      "tyrosine-protein phosphatase C isoform CD45RA 5) C-C chemokine"
                      "receptor type 7, receptor-type tyrosine-protein phosphatase C isoform"
                      "CD45RA"
Cell ID               "1) CL_0000895 2) CL_0000898 3) CL_0000900 4) CL_0001045 5) CL_0002677"
Cell Label            "1) naive thymus-derived CD4-positive, alpha-beta T cell 2) naive T"
                      "cell 3) naive thymus-derived CD8-positive, alpha-beta T cell 4) naive"
                      "CCR4-positive regulatory T cell 5) naive regulatory T cell"
```

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

The output from *Res*$*Table* shows many properties of the phenotype queried. The name of the markers in the ontology are shown, in this case "C-C chemokine receptor type 7" and "receptor-type tyrosine-protein phosphatase C isoform CD45RA" for CCR7 and CD45RA, respectively. A successful match is also specified given that there are cell types that contain all the markers queried. The 1) -

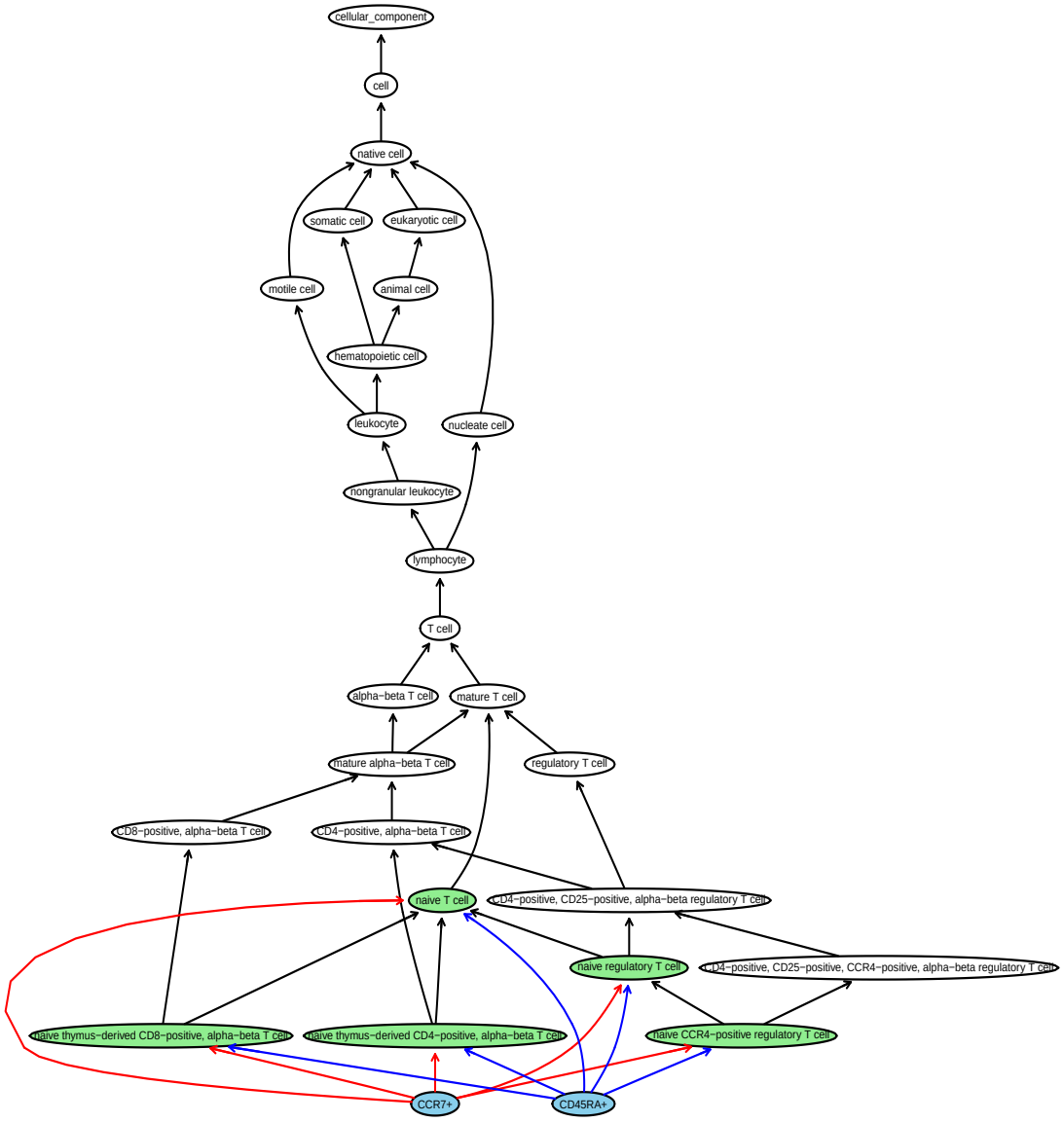

Figure 1: Tree diagram of the cell hierarchy when querying with phenotype: “CCR7+CD45RA+”

5) represent the five cell types that are considered exact matches. In Figure 1, a tree diagram shows the cell hierarchy that is dependent on both CCR7+ and CD45RA+. There are five cell types that contain both markers.

The black arrows are defined as "is a" (ex. native cell "is a" cell). The coloured arrows are the inverse of "has/lacks plasma membrane part" (ex. native regulatory T cell "has plasma membrane part" CD45RA). Each marker is associated with its own colour, which makes it easier for the user to tell which cell type contains which markers. The blue nodes are the "has" markers, while the red nodes are the "lacks" markers. The green nodes are the exact matches, while the beige nodes are the partial matches.

3.4 Visual-Skip, Reset-Archive and Keep-Archive Example

The "VisualSkip" argument in ***flowCL*** can be used to if the user does not want the visual results.

```
> Res <- flowCL("CCR7+CD45RA+", VisualSkip = TRUE)
```

This can reduce the computational time. There is an argument called "ResetArch", which is not shown in this vignette. If it is set to TRUE, it will first delete the entire archive and start adding data to a new one every time something is queried. Therefore, the code is slower after the archive is reset, however, the code will become faster on average with every query. There is also an argument called "KeepArch", also not shown, which allows the user to remove the archive at the end of the simulation if it is set to FALSE. This is useful when the user rather not have files stored on the hard drive.

3.5 Perfect-Match with Ontology-Names Example

An example of a perfect match is shown with "CCR7+CD45RA+CD8+".

```
> Res <- flowCL("CCR7+CD45RA+CD8+", CompInfo = TRUE, OntolNamesTD = TRUE)

The phenotype of interest is CCR7+CD45RA+CD8+
Marker CCR7 is called C-C chemokine receptor type 7
Marker CD45RA is called receptor-type tyrosine-protein phosphatase C isoform CD45RA
Marker CD8 has been FORCED to update to "T cell receptor co-receptor CD8"
At least one marker was not previously queried. Querying all.
Locating marker C-C chemokine receptor type 7
Locating marker receptor-type tyrosine-protein phosphatase C isoform CD45RA
Locating marker T cell receptor co-receptor CD8
Initial query results saved in flowCL_results/results/results_CCR7+CD45RA+CD8+.csv
Perfect match(es) found.
Parent information saved in flowCL_results/parents/parent.resCCR7+CD45RA+CD8+.csv
Time elapsed: 00:00:01
Iterations at 1 out of 1

Total time was:  00:00:01
Archive saved in "[current directory]/flowCL_results/"

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

The tree diagram in Figure 2 only has one perfect match. The "OntolNamesTD" option allows for the marker nodes to display their ontology names in the tree diagrams instead of their short names (ex.

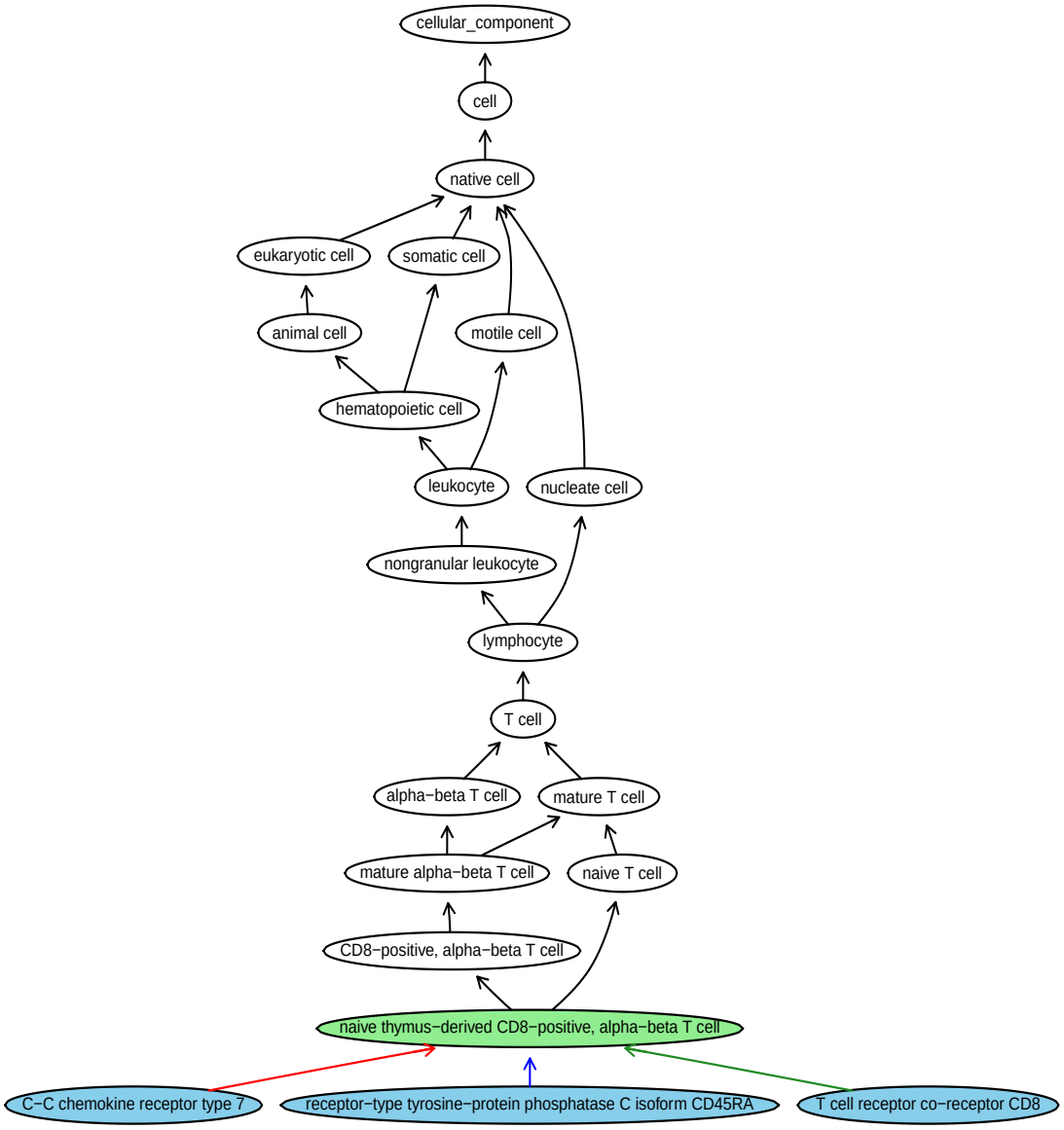

Figure 2: Tree diagram of the cell hierarchy when querying with phenotype: "CCR7+CD45RA+CD8+"

"T cell receptor co-receptor CD8" is displayed instead of "CD8"). The default for "OntolNamesTD" is FALSE. The "CompInfo" option allows the user to view the computational updates while the code is in progress. The default for "CompInfo" is FALSE.

3.6 HIPC Example

There is one pre-set input for a list of phenotypes. If the user types "HIPC" as an argument, then this pre-set list is used. "HIPC" first lists all the markers that make up the HIPC phenotypes and then lists all the common HIPC phenotypes.

```
> Res <- flowCL("HIPC", Indices=c(73,54,50), MaxHitsPht=7)
> tmp <- Res$'CD3+CD4+CD127-CD25+'
> plot(tmp[[1]], nodeAttrs=tmp[[2]], edgeAttrs=tmp[[3]], attrs=tmp[[4]])

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

> tmp <- Res$'CD3-CD19+CD20+CD38+CD24+'
> plot(tmp[[1]], nodeAttrs=tmp[[2]], edgeAttrs=tmp[[3]], attrs=tmp[[4]])
> Res$Table

$`CD3+CD4+CD127-CD25+`
                      [,1]
Short marker names    "CD3+CD4+CD127-CD25+"
Ontology marker names "alpha-beta T cell receptor complex, CD4 molecule, interleukin-7"
                      "receptor subunit alpha, interleukin-2 receptor subunit alpha"
Successful Match?     "Yes - 2 hits"
Marker ID             "1) PR_000001004, PR_000001380, PR_000001869, GO_0042105 2)"
                      "PR_000001004, PR_000001380, PR_000001869, GO_0042105"
Marker Label          "1) alpha-beta T cell receptor complex, CD4 molecule, interleukin-2"
                      "receptor subunit alpha, interleukin-7 receptor subunit alpha 2)"
                      "alpha-beta T cell receptor complex, CD4 molecule, interleukin-2"
                      "receptor subunit alpha, interleukin-7 receptor subunit alpha"
Cell ID               "1) CL_0000896 2) CL_0001043"
Cell Label            "1) activated CD4-positive, alpha-beta T cell 2) activated"
                      "CD4-positive, alpha-beta T cell, human"

$`CD3+CD4+CD8-CCR7-CD45RA-`
                      [,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"
Successful Match?     "Yes"
Marker ID             "1) PR_000001004, PR_000025402, PR_000001203, PR_000001015, GO_0042105"
Marker Label          "1) 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"
Cell ID               "1) CL_0000905"
Cell Label            "1) effector memory CD4-positive, alpha-beta T cell"
```

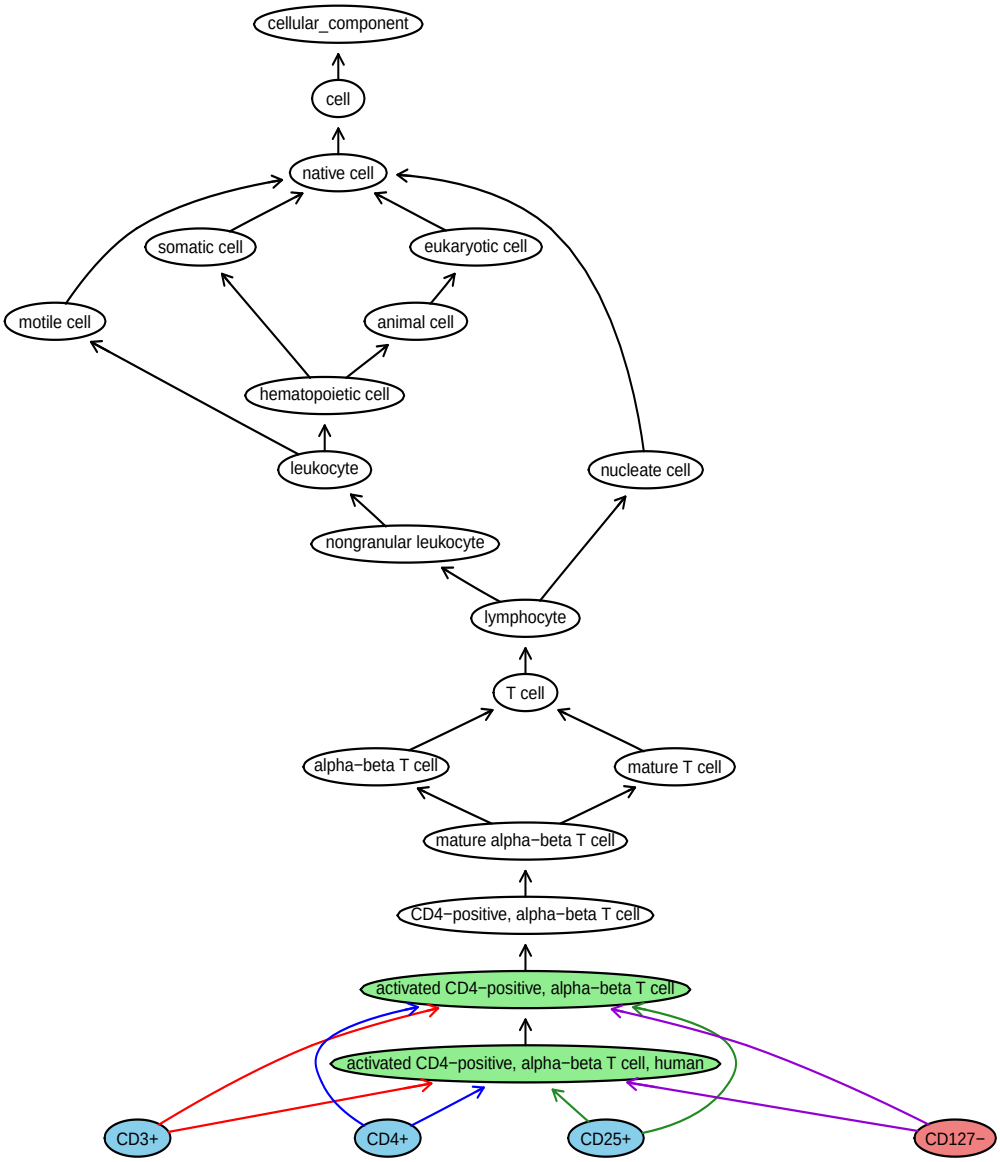

Figure 3: Tree diagram of the cell hierarchy when querying with phenotype: "CD3+CD4+CD127-CD25+"

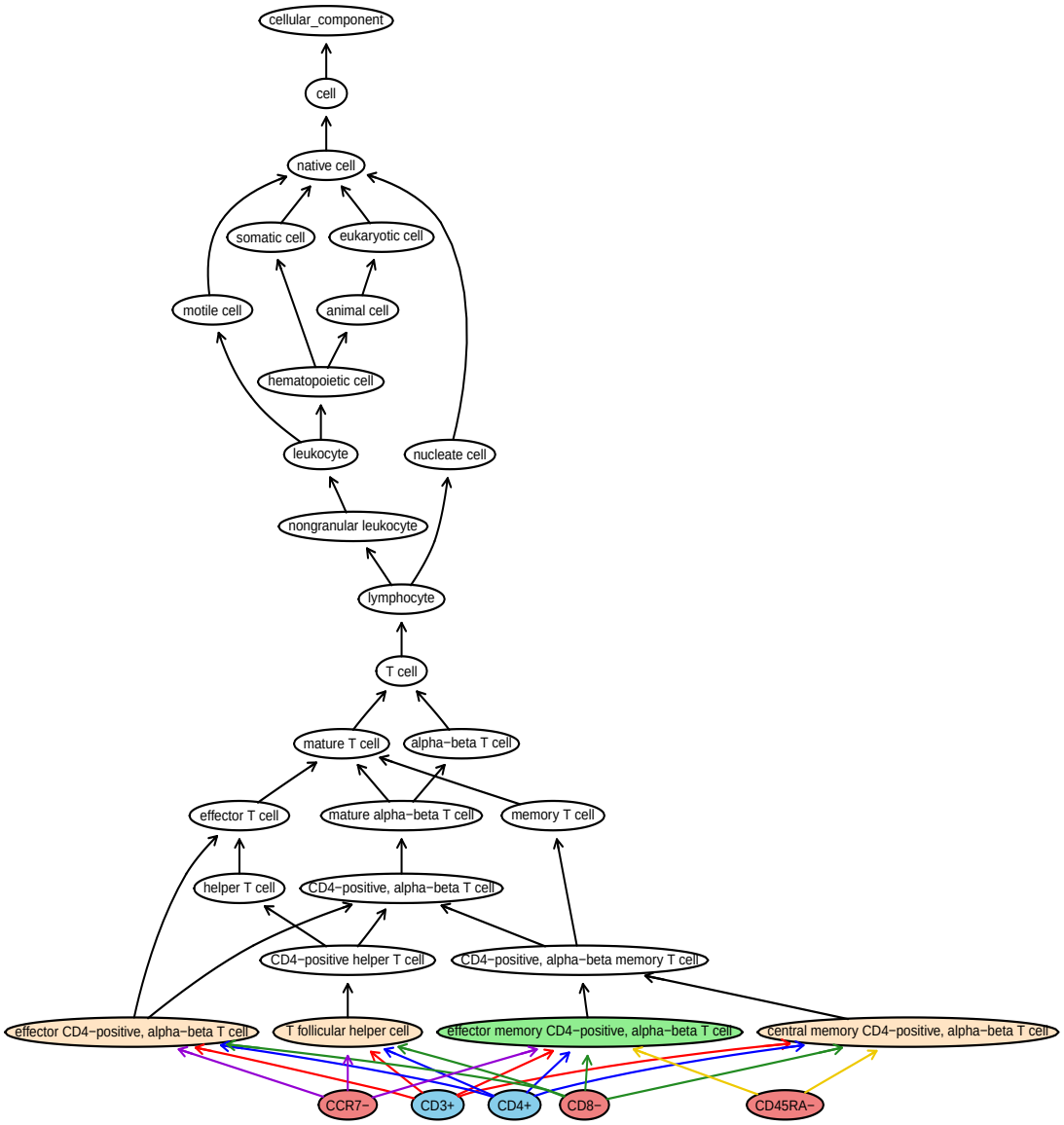

Figure 4: Tree diagram of the cell hierarchy when querying with phenotype: "CD3+CD4+CD8-CCR7-CD45RA-"

```
$`CD3-CD19+CD20+CD38+CD24+`
                        [,1]
Short marker names      "CD3-CD19+CD20+CD38+CD24+"
Ontology marker names   "CD3 epsilon, CD19 molecule, membrane-spanning 4-domains subfamily A"
                        "member 1, ADP-ribosyl cyclase 1, signal transducer CD24"
Successful Match?       "No"
Marker ID               "1) PR_000001002, PR_000001289, PR_000001408, PR_000001020 2)"
                        "PR_000001002, PR_000001289, PR_000001408, PR_000001020 3)"
                        "PR_000001002, PR_000001289, PR_000001408, PR_000001020 4)"
                        "PR_000001002, PR_000001289, PR_000001408, PR_000001020 5)"
                        "PR_000001002, PR_000001289, PR_000001408, PR_000001020 6)"
                        "PR_000001002, PR_000001289, PR_000001932, PR_000001020 7)"
                        "PR_000001002, PR_000001289, PR_000001408, PR_000001020 + more"
Marker Label            "1) CD19 molecule, membrane-spanning 4-domains subfamily A member 1,"
                        "ADP-ribosyl cyclase 1, CD3 epsilon 2) CD19 molecule,"
                        "membrane-spanning 4-domains subfamily A member 1, ADP-ribosyl cyclase"
                        "1, CD3 epsilon 3) CD19 molecule, membrane-spanning 4-domains"
                        "subfamily A member 1, ADP-ribosyl cyclase 1, CD3 epsilon 4) CD19"
                        "molecule, membrane-spanning 4-domains subfamily A member 1,"
                        "ADP-ribosyl cyclase 1, CD3 epsilon 5) CD19 molecule,"
                        "membrane-spanning 4-domains subfamily A member 1, ADP-ribosyl cyclase"
                        "1, CD3 epsilon 6) CD19 molecule, membrane-spanning 4-domains"
                        "subfamily A member 1, signal transducer CD24, CD3 epsilon 7) CD19"
                        "molecule, membrane-spanning 4-domains subfamily A member 1,"
                        "ADP-ribosyl cyclase 1, CD3 epsilon + more"
Cell ID                 "1) CL_0000962 2) CL_0000963 3) CL_0000965 4) CL_0000964 5) CL_0000966"
                        "6) CL_0002054 7) CL_0002101 + more"
Cell Label              "1) Bm2 B cell 2) Bm3-delta B cell 3) Bm3 B cell 4) Bm2' B cell 5) Bm4"
                        "B cell 6) Fraction E immature B cell 7) CD38-positive naive B cell +"
                        "more"
```

The "Indices" option allows the user to choose which one(s) to query from the list of phenotypes. The order that the values are input into "Indices" will be the order of the results. The 73^{th}, 54^{nd} and 50^{th} elements of the "HIPC" phenotype list are queried. These three were chosen to showcase some of the differences in results the user can receive.

The first case, 73^{rd} element with phenotype "CD3+CD4+CD127-CD25+", is a double perfect match, as seen in Figure 3. However, this phenotype is defined by HIPC as a "CD4-positive, CD25-positive, alpha-beta regulatory T cell". Therefore, these are incorrect matches. This occurred because the correct marker to be searched would have been "CD127lo" instead of "CD127-". The current version of ***flowCL*** does not handle this type of input. A future version of ***flowCL*** will hopefully address this. Therefore, the user must be careful relying on the results of ***flowCL***, especially when the desired cell populations are dim, lo, bright or hi instead of the more common "+" or "-".

The second case, 54^{th} element with phenotype "CD3+CD4+CD8-CCR7-CD45RA-", has one perfect match. The tree diagram in Figure 4 also shows two more possible cell types that match 4 out of 5. This was done to give the user more information on possible alternative matches. If there were more than a total of four that were 4 or 5 out of 5, then the 4 out of 5 cases would not be shown on the tree diagram. This value of four is arbitrary. It was chosen to make sure the tree diagrams did not get cluttered. This will be explained more with the "cut-off score", introduced shortly.

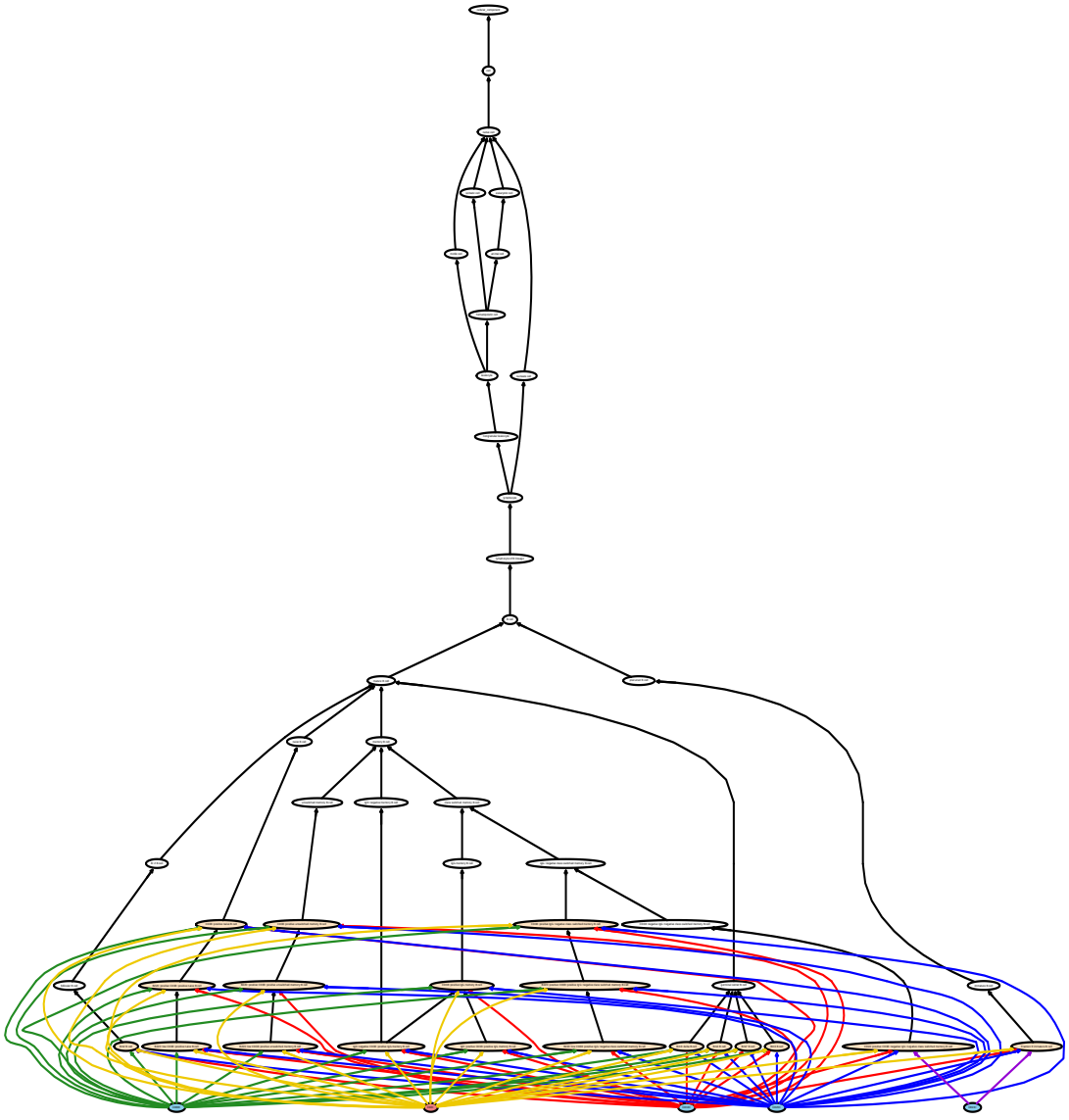

Figure 5: Tree diagram of the cell hierarchy when querying with phenotype: "CD3-CD19+CD20+CD38+CD24+"

The third case, 50^{th} element with phenotype "CD3-CD19+CD20+CD38+CD24+", there were no perfect matches. There are many 4 out of 5 matches and "MaxHitsPht" dictates how many of these are shown in $\$Table$; the default is 5 and in this example it is set to 7. The tree diagram, however, shows all of the 4 out of 5 cases as seen in Figure 5. This tree diagram is influenced by a "cut-off score" and not by "MaxHitsPht".

The "cut-off score" is either equal to the number of matched markers on the highest matched cell type or one less than that. The latter case only occurs provided there are at most four cell types selected. For example, Figure 4 shows where the "cut-off score" was lowered from 5 to 4, while Figure 5 shows where the "cut-off score" was set at 4 and not lowered. This current way of calculating the cut-off has its complications. First, if there are too many non perfect cell type matches, then the tree diagram will become cluttered, as seen in Figure 5. And second, if only one marker is queried, then the tree diagram that is created is usually completely unclear. Since there is more interest in phenotypes being queried than individual markers, this second complication is acceptable. The first, however, is unavoidable.

3.7 Cell Label Example

In many cases ***flowCL*** will only be used to extract the best possible cell label given a set of markers. The following code shows an example of this:

```
> x <-"CD3+CD4+CD8-CCR7-CD45RA-"
> Res <- flowCL(x)
> Res$Cell_Label[[x]][[1]]
```

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

The user should note that the returned value could be one of many possible best cell labels. To access the other possible best cell labels the "1" in $Res\$Cell\_Label[[x]][[1]]$ can be changed to a higher index. Also, the cell labels with a lower number of marker matches compared to the best match will not be accessible with this method.

3.8 Last Comments

- If the user wants all possible results and all tree diagrams, then the command of "flowCL()" will achieve this. All the defaults are set to run the HIPC phenotypes and their individual markers. Running the full code with no archive can take up to 40 minutes.
- There are two packages that ***flowCL*** depends upon. The first is ***SPARQL***, which is used when retrieving cell ontology data. The other is ***Rgraphviz***, which is used when creating the tree diagrams.
- The user should note that since the ontology can be and will be updated, the R results shown in this vignette, both the figures and the static text, may not match a live run of flowCL. Also the data files containing a pre-loaded archive can be outdated. The maintainer will try to keep the vignette and data file as up to date as possible. The data file as well as text and the figures in the examples from this vignette describe results from the cell ontology that was last updated on March 18^{th} 2014.