

Train a cell type classifier for single cell RNA-Seq analysis

March 28, 2022

Sample to Insight -

QIAGEN Aarhus \cdot Silkeborgvej 2 \cdot Prismet \cdot 8000 Aarhus C \cdot Denmark digitalinsights.qiagen.com \cdot ts-bioinformatics@qiagen.com



Train a cell type classifier for single cell RNA-Seq analysis

Introduction

This tutorial uses the CLC Genomics Workbench and CLC Single Cell Analysis Module to focus on one of the main areas when conducting single cell RNA-Seq analysis: performing cell type prediction and overlaying the information on Dimensionality Reduction Plots. This tutorial covers the following:

- Importing Expression Matrices (#) and Cell Clusters ().
- Building a Cell Type Classifier (🍻) from the bottom up.
- Learning how to safely add more cells and new cell types to the classifier.
- Predicting cell types using the classifier.
- Tips and tricks for further analysis.

Prerequisites

For this tutorial, you must be working with CLC Genomics Workbench 22.0 and CLC Single Cell Analysis Module 22.1 or higher. Plugin installation is described in the CLC Genomics Workbench manual.

Importing expression matrices and cell clusters

We start by importing the tutorial data using Import Expression Matrix in Loom Format.

- 1. Download the data and unzip in a location of your choice. The folder contains 5 loom files.
- 2. Start the CLC Genomics Workbench.
- 3. Import the .loom files one by one to a folder named **Samples** via the toolbar:
- Import ((() | Import Expression Matrix () | Import Expression Matrix in Loom Format ().

The loom format contains both the expression matrix and cell clusters that are needed for the analysis. Configure the wizard as shown in figure 1:

- (a) Set Gene or transcript track to 'Homo_sapiens_ensembl_v99_hg38_no_alt_analysis_set_Genes'
- (b) Set Cell format to 'sample-{sample}.{barcode}'.
- (c) Set Expression matrix to the loom file to be imported.
- (d) Add at least 'ClusterCellType' to **Create clusters for**. This contains the assigned cell type for each cell.
- 4. Click Next and tick Create clusters from the output options menu.
- 5. Choose to save the results in a new folder called **Samples**. If needed, create the folder using the **New folder** button near the top. Click **Finish**.



. Choose where to run	Loom import General options				
. Loom import	Gene or transcript tra	ck 🗦 🖁 Homo	_sapiens_ensembl_v99_	hg38_no_alt_analysis_	set_Genes 🙀
. Result handling	Spike-in controls				Ø
	Cell format	sample-{	sample}.{barcode}		
		Press Shift	+ F1 for options		
	Sample				
	Loom options				
	Expression matrix		E-MTAB-6678.loom		Browse
	Cell ID attribute		CellID		~
	Gene or transcript ID a	attribute	Accession		~
	Gene or transcript nam	ne attribute	Gene		~
	Create clusters for		Selected 4 elements.		÷
	Preview cells				
	Input barcode		Parsed sample	Parsed barcode	
	sample-E-MTAB-6678	-	E-MTAB-6678_1	1	Disable
	sample-E-MTAB-6678	_37.3881	E-MTAB-6678_37	3881	

Figure 1: Configuration of the Import Expression Matrix in Loom Format wizard.

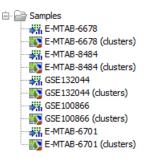


Figure 2: Folder showing the imported data as Expression Matrices and Cell Clusters.

6. Five matrices with accompanying cell clusters should be located in the **Samples** folder after import, see figure 2.

We will build a new Cell Type Classifier using four matrices, out of which one will be used for investigating how incorrect annotations can have a negative impact. We will test the classifier using the fifth matrix, create a UMAP plot, and compare the predicted cell types with the imported ones.

Training a Cell Type Classifier

We will use the four matrices named E-MTAB-6678, E-MTAB-8484, GSE132044 and E-MTAB-6701 for building a classifier. Before starting, we recommend reading the manual section Train Cell Type Classifier and its subsections.



- 1. Start the **Train Cell Type Classifier** tool from the Single Cell Analysis Toolbox: **Gene Expression** () | **Cell Type Classification** () | **Train Cell Type Classifier** ()
- 2. Select E-MTAB-6678 and configure the following wizard as shown in figure 3:

Gx Train Cell Type Classifier		×
1. Choose where to run	Classifier	
2. Select Expression Matrix	Cell type dusters 🚯 E-MTAB-6678 (dusters)	R
3. Classifier	Cell type category ClusterCellType	\sim
4. Result handling	Select cell types (Nothing selected)	4
	Cell type classifier	Ŕ
	 Treat all cells equally 	
	O Use incoming cells first	
	 Use existing cells first 	
	Validation	
	Validation expression matrices	R
	Validation cell type clusters	6
	Validation cell type category -no selection-	~
	Percent regression to report 4.0	
Help Reset	Previous Next Finish C	Cancel

Figure 3: Configuration of the Train Cell Type Classifier tool for training a new classifier.

- Set Cell type clusters to the E-MTAB-6678 cell clusters.
- Set Cell type category to 'ClusterCellType'.
- **Select cell types** can optionally be used for choosing a subset of all available cell types for training. Here we leave it empty, so that all cell types are used.
- 3. Click Next, choose to save the results in a new folder called Classifier, and click Finish.

Inspecting the outputs

A Cell Type Classifier and a report are created, named after the matrix using for training, here E-MTAB-6678. The outputs contain information about the cell types added to the classifier, as shown in figures 4 and 5. For more details, see Interpreting the output of Train Cell Type Classifier.

Extending an existing Cell Type Classifier

The Train Cell Type Classifier tool can extend an existing Cell Type Classifier with additional data. It can also optionally perform validation for the newly trained classifier.



1 Input data cell types			Report Settings		
This section lists the cell types in the in classifier. The cell types are listed in tw whether or not they are in the QIAGEN 1.1 Cell types in the QIAGEN	vo tables depending on Cell Ontology.		1.2 Cell types n	types In the QIAGEN Cell In the QIAGEN	
Cell type	Cells (#)		Text format		
dendritic cells	12				
endothelial cells	76				
epithelial cells	42				
extravillous trophoblast cells	152				
granulocytes	248				
macrophages	1,219				
monocytes	350				
natural killer cells	1,415				
pericytes	94				
plasma cells	15				
stromal cells	468				
T lymphocytes	1,823				
1.2 Cell types not in the QIA All cell types were found in the QIAGEN					
907		20	- 13	Help	Save View

Figure 4: The report generated by Train Cell Type Classifier when training a new classifier.

E-MTAB-6678 (classifie	er) ×					
Rows: 12 Fil	ter to Selection		Filter	₹	I▶ Table Settings Show column =	- ^
Cell type	Number of cells	Number of samples	E-MTAB-6678_1		Cell type	
dendritic cells	12	3			Top features supporting this cell type	
endothelial cells	50	4			Top features supporting another cell	
epithelial cells	42	6				
extravillous trophoblast ce	l <u>s</u> 50	5			✓ Number of cells	
<u>granulocytes</u>	50	19			Number of samples	
macrophages	50	20	4			
monocytes	50	6			E-MTAB-6678_1	
natural killer cells	50	30	2		E-MTAB-6678_10	
pericytes	50	5				
plasma cells	15	5			E-MTAB-6678_11	
stromal cells	50	17	3		E-MTAB-6678_12	
<u>T lymphocytes</u>	50	28	2		E-MTAB-6678 13	V .
III 🖸 🖬					□ □ □ □ □ □ □	

Figure 5: The Cell Type Classifier generated by Train Cell Type Classifier when training a new classifier. Additional columns can be selected in the side panel menu to the right.

1. Start the **Train Cell Type Classifier** tool from the Single Cell Analysis Toolbox:

Gene Expression () | Cell Type Classification () | Train Cell Type Classifier ()

- 2. Select E-MTAB-8484 and configure the following wizard as shown in figure 6:
 - Set Cell type clusters to the E-MTAB-8484 cell clusters.



Gx Train Cell Type Classifier	×
1. Choose where to run	Classifier Training
2. Select Expression Matrix	Cell type clusters 🚯 E-MTAB-8484 (clusters)
3. Classifier	Cell type category ClusterCellType ~
4. Result handling	Select cell types (Nothing selected)
	Cell type classifier 🐉 E-MTAB-6678 (classifier) 😡
	Treat all cells equally
	O Use incoming cells first
	O Use existing cells first
100	
CP.C.	Validation
UST	Validation expression matrices 🟭 E-MTAB-6678
Marine Carlos	Validation cell type clusters 🚯 E-MTAB-6678 (clusters) 👼
19	Validation cell type category ClusterCellType V
10	Percent regression to report 4.0
Contraction of the second	
Help Reset	Previous Next Finish Cancel

Figure 6: Configuration of the Train Cell Type Classifier tool for extending an existing classifier and performing validation.

- Set Cell type category to 'ClusterCellType'.
- Set Cell type classifier to the E-MTAB-6678 classifier you just created.
- Set **Validation expression matrices** and **Validation cell type clusters** to the E-MTAB-6678 expression matrix and cell clusters, which were used for training the previous classifier.
- Set Validation cell type category to 'ClusterCellType'.
- 3. Click Next, choose to save the results in the Classifier folder, and click Finish.

Before looking at the generated outputs, we add another data set to the classifier. Once the previous run is completed, rerun the Train Cell Type Classifier using GSE132044 and configuring the wizard as shown in figure 7.

Inspecting the report when extending a Cell Type Classifier

As we used validation data, the E-MTAB-8484 report contains a section about 'Validation data cell types', see figure 8.

The first table lists the cell types for which validation cannot be performed, and the reason for it. Here, several cell types cannot be validated, as all cells from E-MTAB-6678 annotated with these cell types have been used for training the classifier. The Train Cell Type Classifier tool ensures that validation is not performed for cells that are used for training.

The second table provides a performance summary for the cell types that are present in the validation data. For each cell type, the table contains:



Gx Train Cell Type Classifier	×
1. Choose where to run	Classifier _ Training
2. Select Expression Matrix	Cell type dusters 🛐 GSE132044 (dusters)
3. Classifier	Cell type category ClusterCellType ~
4. Result handling	Select cell types (Nothing selected)
n nesarchanamig	Cell type dassifier 💹 E-MTAB-8484 (dassifier) 🛱
	Treat all cells equally
	O Use incoming cells first
	O Use existing cells first
0	
() ()	Validation
(UST	Validation expression matrices 🚜 E-MTAB-6678, E-MTAB-8484 😡
and the second second	Validation cell type dusters 🛛 🚯 E-MTAB-6678 (dusters), E-MTAB-8484 (dusters)
107	Validation cell type category ClusterCellType ~
10 Tenne	Percent regression to report 4.0
2 O Manualut	
Help Reset	Previous Next Einish Cancel

Figure 7: Configuring the Train Cell Type Classifier tool for adding GSE132044 and performing validation using the previously added matrices.

- Whether the cell type was present in the input data, here E-MTAB-8484.
- How many cells in the validation data, here E-MTAB-6678, are of the given cell type.
- If any regression occurred, for how many matrices there is a regression. A regression is a change in performance ('Change correct (%)') that is negative and its absolute value is larger than the **Percent regression to report**, customized during the execution, see figure 6.
- The number of cells that are correctly predicted with the given cell type, out of the the total number of cells that are annotated with the cell type, for both the existing ('Old correct (%') and new ('New correct (%)') classifiers. The difference between the correct prediction is reported in 'Change correct (%)'.

For most cell types in the E-MTAB-8484 report, there is no change in performance. Prediction worsened slightly, but within the limit of allowed regression, for natural killer cells. For granulocytes and T lymphocytes, prediction improved slightly.

It is difficult to train a classifier that can perfectly predict all cell types, as can be seen in figure 8. Cell types that are related to each other, such as natural killer cells and T lymphocytes, can have expression profiles that are so similar, that it is difficult to accurately distinguish them.

Because no regressions have been identified, the following sections are empty in the E-MTAB-8484 report.

The GSE132044 report indicates regressions for two cell types, highlighted in red, see figure 9.

The following sections give additional details for the identified regressions, see figure **10**.



2.1 Cell type							
	Cell type		In input data?		Reaso	n	
dendritic cells		No			alidation data (afte used to train the o		
epithelial cells		No			alidation data (afte used to train the o		
plasma cells		No			No validation data (after removing any cells used to train the classifier)		
vhose performa natrix are listed	e performance regr ince improved by m in the next tables. ghlighted in red has	essed by more th ore than the sam	an "Percent regre e amount are hig	hlighted in green	are highlighted in r . Details of regres:	ed. Cell types	
vhose performa natrix are listed	e performance regr ince improved by m in the next tables.	essed by more th ore than the sam	an "Percent regre e amount are hig	ession to report" a phlighted in green e-training the clas	are highlighted in r . Details of regres:	red. Cell types sions for each	
vhose performa natrix are listed • If a cell type hi Cell type endothelial	e performance regr ince improved by m in the next tables. ghlighted in red has	essed by more th ore than the sam s "In input data? =	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the clas Change correct	are highlighted in r . Details of regress ssifier without it.	ed. Cell types sions for each Old correct (%)	
whose performa matrix are listed • If a cell type hi Cell type endothelial cells extravillous trophoblast	e performance regr ince improved by m in the next tables. ghlighted in red has In input data?	essed by more th ore than the sam s "In input data? = Cells (#)	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the clas Change correct (%)	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00	ed. Cell types sions for each Old correct (%) 100.00	
vhose performa natrix are listed If a cell type hi Cell type endothelial cells extravillous trophoblast cells	e performance regr ince improved by m in the next tables. ghlighted in red has In input data? No	essed by more th ore than the sam s "In input data? = Cells (#) 26	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the class Change correct (%) 0.00	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00	ed. Cell types sions for each Old correct (%) 100.00	
vhose performa natrix are listed If a cell type hi Cell type endothelial cells extravillous trophoblast cells granulocytes	e performance regr ince improved by m in the next tables. ghlighted in red has In input data? No	essed by more th ore than the sam s "In input data? = Cells (#) 26 102	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the class Change correct (%) 0.00	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00 100.00	ed. Cell types sions for each Old correct (%) 100.00	
vhose performa natrix are listed • If a cell type hi Cell type endothelial cells extravillous trophoblast cells granulocytes macrophages	e performance regr nce improved by m in the next tables. ghlighted in red has In input data? No No No	essed by more th ore than the sam s "In input data? = Cells (#) 26 102 198	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a hlighted in green e-training the class Change correct (%) 0.00 0.00 1.01	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00 100.00 92.93 998.97	ed. Cell types sions for each Old correct (%) 100.00 100.00 91.92 98.89	
vhose performa natrix are listed • If a cell type hi Cell type endothelial cells extravillous trophoblast cells granulocytes macrophages monocytes natural killer	e performance regr ince improved by m in the next tables. ghlighted in red has In input data? No No No No	essed by more th ore than the sam s "In input data? = Cells (#) 26 102 198 1,169	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a hlighted in green e-training the class Change correct (%) 0.00 0.00 1.01	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00 100.00 9 92.93 9 98.97 100.00	ed. Cell types sions for each Old correct (%) 100.00 100.00 91.92 98.89	
vhose performa natrix are listed If a cell type hi Cell type endothelial cells extravillous rophoblast cells granulocytes macrophages monocytes natural killer cells	e performance regr ince improved by m in the next tables. ghlighted in red has In input data? No No No No No No	essed by more th ore than the sam s "In input data? = Cells (#) 26 102 198 1,169 300	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the class Change correct (%) 0.00 0.00 1.01 0.09 0.00	are highlighted in r . Details of regress asifier without it. New correct (%) 100.00 100.00 92.93 9 98.97 9 100.00	ed. Cell types sions for each Old correct (%) 100.00 91.92 98.89 100.00 92.01	
whose performa matrix are listed • If a cell type hi	e performance regr ince improved by m in the next tables. ghlighted in red has In input data? No No No No No No No	essed by more th ore than the sam s "In input data? = Cells (#) 26 102 198 1,169 300 1,365	an "Percent regro e amount are hig Yes", consider r Regressed	ession to report" a phlighted in green e-training the class Change correct (%) 0.00 0.00 1.01 0.00 0.00 0.00 0.00	are highlighted in r . Details of regress ssifier without it. New correct (%) 100.00 9 100.00 9 98.97 9 98.97 9 100.00 9 90.11	ed. Cell types sions for each Old correct (%) 100.00 91.92 98.89 100.00 92.01	

Figure 8: The 'Validation data cell types' section of the report generated by Train Cell Type Classifier when extending an existing Cell Type Classifier and performing validation.

The performance decreased by 7% for the helper T lymphocytes. The cause for this is unknown, but the cell type(s) that these cells are predicted as are either not part of the input data, or the different types of mispredictions (incorrect, less or more specific) are individually below the 4% threshold.

The performance for the T lymphocytes decreased by 4% and the cause is that some of these cells are predicted as helper T lymphocytes, which is a subtype of T lymphocytes.

As these two cell types are so closely related, their annotation is more difficult. It could be that some of the cells that have been annotated as T lymphocytes are in fact helper T lymphocytes, and this leads to the observed regression. As performance for helper T lymphocytes decreases and they also seem to be the cause for the T lymphocytes regression, we choose here to not use



🔄 GSE132044 (report) 🗙

2.2 Performance summary for validation data cell types

This table is sorted alphabetically by cell type. It lists the performance of the new and old classifier on the validation data. Cell types whose performance regressed by more than "Percent regression to report" are highlighted in red. Cell types whose performance improved by more than the same amount are highlighted in green. Details of regressions for each matrix are listed in the next tables.

Cell type	In input data?	Cells (#)	Regressed matrices (#)	Change correct (%)	New correct (%)	Old correct (%)
endothelial cells	No	26		0.00	100.00	100.00
extravillous rophoblast cells	No	102		0.00	100.00	100.00
granulocytes	No	198		-1.52	91.41	92.93
helper T lymphocytes	Yes	158	1 (of 1)	-6.96	86.08	93.04
macrophages	No	1,169		-0.09	98.89	98.97
monocytes	No	300		0.00	100.00	100.00
natural killer cells	Yes	1,365		-2.93	87.18	90.11
pericytes	No	44		0.00	100.00	100.00
stromal cells	No	418		0.00	97.85	97.85
T lymphocytes	No	1,773	1 (of 1)	-4.06	88.16	92.22
Th1 cells	No	263		-1.90	91.25	93.16
Th17 cells	No	115		-0.87	90.43	91.30

• If a cell type highlighted in red has "In input data? = Yes", consider re-training the classifier without it.

Figure 9: The 'Performance summary for validation data cell types' section of the GSE132044 report. Regressions occurred for the the red cell types highlighted in red.

the helper T lymphocytes from the GSE132044 data set.

Removing problematic cell types from the input data

In order to remove the helper T lymphocytes found in the GSE132044 data set from the classifier, we run the Train Cell Type Classifier tool again and configure it to start with as in figure 7. We then use **Select cell types** to select everything but helper T lymphocytes, see figure 11. Save the results in a new subfolder **Classifier / Remove helper T lymphocytes**.

When the execution is completed, open the report to inspect the results. Previously, the performance for natural killer cells decreased by 2.93%, while now the regression increased to 4.25%, so we choose to also remove the natural killer cells.

We run the Train Cell Type Classifier tool again and configure it as before, where we use **Select** cell types to select everything but helper T lymphocytes and natural killer cells. Save the results in a new subfolder Classifier / Remove helper T cells and natural killer cells.

No sufficiently large regressions are found, see figure 12.

^ .



		input data			
lata cell types" section	phabetically by cell type. It li n of this report. For each ma s increased by more than "f	atrix, the % of predictions	that are made to cell ty	pes in the input data is	
 Less specific - the ability to predict the sp More specific - the 	classifier may have gained sum of the % in this column	tated as specifically as t the ability to predict a m	he validation data. The ore specific cell type. Co	onsider re-annotating the	
	Matrix and regression	Incorrect	Less specific	More specific	
Cell type	maan and regression				
helper T lymphocytes	E-MTAB-8484: -7.0%				
helper T lymphocytes 4 Regressions This table is sorted all data cell types" section	E-MTAB-8484: -7.0%	sts all the matrices for ro atrix, the % of predictions	that are made to cell ty	pes in the input data is	
A Regressions A Regressions This table is sorted all fata cell types" section isted when this % has Incorrect - consider Less specific - the ability to predict the sp More specific - the	E-MTAB-8484: -7.0% S for cell types no phabetically by cell type. It li n of this report. For each m s increased by more than "f re-training the classifier wi input data may not be anno vecific cell type. classifier may have gained sum of the % in this column	sts all the matrices for ro atrix, the % of predictions Percent regression to rep thout the cell types in thi stated as specifically as t the ability to predict a m	that are made to cell ty port". These are divided s category. he validation data. The ore specific cell type. Co	rpes in the input data is into three categories: classifier may lose the onsider re-annotating the	
helper T lymphocytes 4 Regressions This table is sorted all lata cell types" section isted when this % has Incorrect - consider Less specific - the ability to predict the sp More specific - the ralidation data. If the s	E-MTAB-8484: -7.0% S for cell types no phabetically by cell type. It li n of this report. For each m s increased by more than "f re-training the classifier wi input data may not be anno vecific cell type. classifier may have gained sum of the % in this column	sts all the matrices for ro atrix, the % of predictions Percent regression to rep thout the cell types in thi stated as specifically as t the ability to predict a m	that are made to cell ty port". These are divided s category. he validation data. The ore specific cell type. Co	rpes in the input data is into three categories: classifier may lose the onsider re-annotating the	

Figure 10: The 'Regressions for cell types (not) in input data' sections of the GSE132044 report.

Gx Select: Select cell type	s		×
Available helper T lymphocytes	₽	Selected B lymphocytes blood platelets CD 14+ cells CD 16 positive monocytes cytotoxic T cells natural killer cells	≜ ∓
			Done

Figure 11: Selecting the cell types from the GSE132044 data set.

Finalizing the Cell Type Classifier

We now extend the GSE132044 classifier from the **Classifier / Remove helper T cells and natural killer cells** using the E-MTAB-6701 data set, and we validate it using E-MTAB-6678, E-MTAB-8484 and GSE132044 data sets.

The report shows a clear reduction in the prediction ability for natural killer cells. The GSE132044



₩ GSE132044 (report) ×

2.2 Performance summary for validation data cell types

This table is sorted alphabetically by cell type. It lists the performance of the new and old classifier on the validation data. Cell types whose performance regressed by more than "Percent regression to report" are highlighted in red. Cell types whose performance improved by more than the same amount are highlighted in green. Details of regressions for each matrix are listed in the next tables.

Cell type	In input data?	Cells (#)	Regressed matrices (#)	Change correct (%)	New correct (%)	Old correct (%)
endothelial cells	No	26		0.00	100.00	100.00
extravillous rophoblast cells	No	102		0.00	100.00	100.00
granulocytes	No	198		-1.01	91.92	92.93
nelper T ymphocytes	No	158		1.27	94.30	93.04
macrophages	No	1,169		-0.17	98.80	98.97
monocytes	No	300		0.00	100.00	100.00
natural killer cells	No	1,365		-2.64	87.47	90.11
pericytes	No	44		0.00	100.00	100.00
stromal cells	No	418		0.00	97.85	97.85
F lymphocytes	No	1,773		0.90	93.12	92.22
Th1 cells	No	263		-0.38	92.78	93.16
Th17 cells	No	115		0.00	91.30	91.30

• If a cell type highlighted in red has "In input data? = Yes", consider re-training the classifier without it.

Figure 12: The 'Performance summary for validation data cell types' section of the GSE132044 report after removing helper T cells and natural killer cells from the GSE132044 data set.

classifier predicted only 80% of the cells correctly ('Old correct (%)'), while the percentage reported when the classifier was produced is 87% (figure 12). This is because we removed the natural killer cells from the GSE132044 data when training, but we did not remove them for the validation. However, this does not explain the observed regression. Investigating the cause, we can see that it occurs in the E-MTAB-6678 data set and the cells are more often predicted to be T lymphocytes. This indicates that annotation of natural killer cells and T lymphocytes is not consistent between the different data sets used, where the two cell types can be mixed. This is expected, as these cell types are closely related to each other and not always easy to tell apart. The cell types are also often located close together in Dimensionality Reduction Plots.

We stop extending the classifier. Rename the E-MTAB-6701 classifier to 'Final Cell Type Classifier'. We will now use it to predict cell types on the fifth data set.

Creating UMAP plots an predicting cell types

In this section we will focus on normalizing the expression matrix, creating a UMAP plot, predicting cell types and overlaying cluster information for visualization.

^ .



1. Start the Normalize Single Cell Data tool from the Single Cell Analysis toolbox:

Gene Expression () | Cell Preparation () Normalize Single Cell Data ()

- 2. Select GSE100866 and configure the following wizard to **Each sample is a batch** in the 'Sample level batch correction' group. This matrix contains two samples, each corresponding to a different tissue.
- 3. Click Next, choose to save the results in a new folder called UMAP, and click Finish.

It will take some time for the tool to finish execution. You can monitor the progress in the 'Processes' tab.

4. Start the UMAP for Single Cell tool from the Single Cell Analysis toolbox:

Dimensionality Reduction (i) | UMAP for Single Cell (

- 5. Select 'GSE100866 (residuals)' from the **UMAP** folder. Use default options. Click **Next**, choose to save the output in the **UMAP** folder, and click **Finish**.
- 6. Start the Predict Cell Types tool from the Single Cell Analysis toolbox:

Gene Expression () | Cell Type Classification () | Predict Cell Types ()

7. Select GSE100866 and set **Cell type classifier** to the 'Final Cell Type Classifier' in the following wizard.

Note that either the original GSE100866 or the 'GSE100866 (residuals)' matrices can be used, the results will be the same.

8. Click **Next**, tick **Output cell annotations** to generate the probability per cell type for all cells in the matrix, choose to save the output in the **UMAP** folder, and click **Finish**.

Comparing imported and predicted cell types

To visualize the clusters, either the imported ones or those produced by the Predict Cell Types tool, open the 'GSE100866 (UMAP)' plot and drag the clusters to the 'Clusters' group in the Side Panel. To change the clusters, click on (g_{\ast}) on the top right of the Side Panel and choose **Clear**. Now different clusters can be chosen.

Note that the same UMAP plot can be opened multiple times and the plots can be rearranged by dragging and dropping so that they are side by side. This way, multiple sources of coloring can be viewed at the same time. To use this functionality, the imported and predicted cell types need to be combined into one Cell Clusters element:

1. Start the **Combine Cell Clusters** tool from the Single Cell Analysis toolbox:

Utility Tools (🔊) | Combine Cell Clusters (🛐)

- 2. Select 'GSE100866 (clusters)' and 'GSE100866 (cell types)'.
- 3. Click **Next**, choose to save the results in the **UMAP** folder, and click **Finish**.

Now we can view both the imported and the predicted cell types by opening the 'GSE100866 (UMAP)' plot twice, dragging the 'GSE100866 (combined clusters)' to the 'Clusters' group and



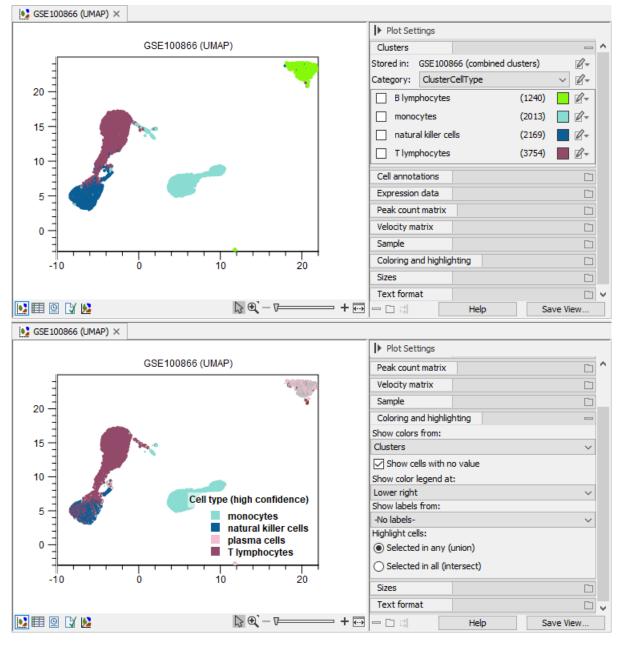


Figure 13: UMAP plot of GSE100866. Top: Cells are colored using the imported cell types. Bottom: Cells are colored using the predicted cell types.

choosing 'ClusterCellType' and 'Cell type (high confidence)', respectively, in the two views (see figure 13).

There is a good general correspondence between the imported and predicted cell types. This can be further investigated by checking the overlap of the different clusters:

- 1. Start the **Create Heat Map for Cell Abundance** tool from the Single Cell Analysis toolbox: **Cell Annotation** (()) | **Create Heat Map for Cell Abundance** (())
- 2. Configure the wizard as shown in figure 14:



Gx Create Heat Map for G	Cell Abundance		×
1. Choose where to run	Parameters Categories		
2. Parameters	Clusters	GSE100866 (combined clusters)	ø
3. Result handling	Cell annotations		ø
	Group by (rows)	Cell type (high confidence)	÷
	Select groups (rows)	(Nothing selected)	÷
	Group by (columns)	ClusterCellType	÷
0%	Select groups (columns)	(Nothing selected)	4
	Scaling		
	🔾 By all		
	Per row		
17011 17011	O Per column		
A MILLING COMPANY			
Help Res	et <u>P</u> revious	Next Einish Cano	:el

Figure 14: Configuration of the Create Heat Map for Cell Abundance wizard for comparing the imported and predicted cell types.

- (a) Set **Clusters** to 'GSE100866 (combined clusters)'.
- (b) Set **Group by (rows)** to 'Cell type (high confidence)'. These are the predicted cell types.
- (c) Set **Group by (columns)** to 'ClusterCellType'. These are the imported cell types.
- 3. Click **Next**, choose to save the results in the **UMAP** folder, and click **Finish**.

The resulting heat map shows the overlap between the different clusters. When the default **Per row** is chosen in the 'Scaling' group, the heat map indicates how the predicted cell types are distributed across the imported ones. Choosing the **Per column** option will instead show how the imported cell types are distributed across the predicted ones (see figure 15). This helps identify that most cells (88%) that were originally annotated as B lymphocytes do not have a predicted cell type (Unknown), while the majority (95%) of the cells that are predicted as plasma cells where annotated as B lymphocytes.

Some of the cells are predicted as either dendritic or endothelial cells, but as shown in the left panel of figure 15, they account to a small percentage of the cells.

Note that the order of the cell types in the heat map can be changed by using **Select groups** (rows) and **Select groups** (columns) when configuring the tool execution (see figure 14).

The B lymphocytes cell type is entirely missing from the predicted cell types, and for those cells were a prediction is made, the predicted cell type is plasma cells. We can investigate if this can be correct by using the QIAGEN Cell Ontology.

Inspecting relations between cell types using the QIAGEN Cell Ontology

When predicting cell types using the Cell Type Classifier, how precise the prediction is depends



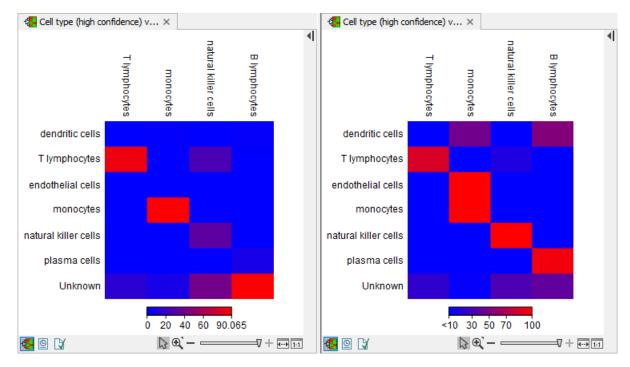


Figure 15: Output of the Create Heat Map for Cell Abundance. Left: Scaling is set Per column. Right: Scaling is set Per row.

on the data used for training. The cell types used here are present in the QIAGEN Cell Ontology, and we can use this to investigate how B lymphocytes and plasma cells relate to each other.

Note that when importing cell clusters, either from .loom files as done here, or from txt files using Import Cell Clusters, clusters can be mapped to the extent possible to the QIAGEN Cell Ontology by using **Map clusters to QIAGEN Cell Ontology**.

There are multiple ways we can find information about cell types that are in the ontology:

- Using the Browse QIAGEN Cell Ontology tool.
- Clicking on the cell type in the Cell Type Classifier, see figure 5.
- Using the Side Panel options in the Dimensionality Reduction Plot, see figure 16.

Using the QIAGEN Cell Ontology, we can see that plasma cells are subtype of B lymphocytes and hence a more specific predicted cell type, see figure 17. Whether this prediction is correct or not needs investigating, but it highlights how the Cell Type Classifier can learn different cell types and lead to more specific cell type predictions than manual annotation.

This tutorial showcases Train Cell Type Classifier and Predict Cell Types tools provided in the CLC Single Cell Analysis Module, it highlights their strengths and demonstrates how the report can help guide training a Cell Type Classifier.



Plot Setting	s					
Clusters					^	
Stored in: GS	E10086	6 (combined clu	isters)	Ø-		
Category: Ce	ell type	(high confidence	~ 🖉 =			
dendritic d	cells		(61)	Ø-		
endothelial cells			(1)	-		
monocyte	(1813)	Ø-				
📃 natural ki	(703)	Ø-				
📄 plasma ce	lls		(110)	-		
T lympho		Show in QIA	GEN Cel	l Ontology		
Unknown		Rename Clus	ter			
Cell annotatio		Delete Cluster				
Expression da		Add to New Cluster				
Stored in: GSE						
Show: Ray		Add to Clust	er		>	
Selected featur		Create Subset				
		Extract to Table				
Peak count matrix					5	
		Help	Sa	ave View		

Figure 16: Opening the QIAGEN Cell Ontology using the Side Panel options in a Dimensionality Reduction Plot.

Gx Browse QIAGEN Cell Ontology				×
Selected: plasma cells				
B lymphocytes	~	Iverties Cell type: Definition:	Is plasma cells A terminally differentiated, post-mitotic, antibody secreting cell of the B cell lineage with the phenotype CD138-positive, surface immunonoglobulin-negative, and MHC Class II-negative. Plasma cells are oval or round with extensive rough endoplasmic reticulum, a well-developed Golgi apparatus, and a round nucleus having a characteristic cartwheel heterochromatin pattern and are devoted to producing large amounts of immunoglobulin. GO_REF:0000031 GOC:add GOC:dsd http://en.wikipedia.org/wiki/Plasma_cell ISBN:0721601464 ISBN:0781735149	~
Help			Close	e

Figure 17: Plasma cells seen in the QIAGEN Cell Ontology. Note that they are subtype of B lymphocytes.



Further analysis of the data

To further explore the data and the alignment between the predicted cell types and expert knowledge, we can investigate the expression of two known marker genes for natural killer cells, NKG7 and GNLY, and if this matches with the predicted natural killer cells cluster, see figure 18.

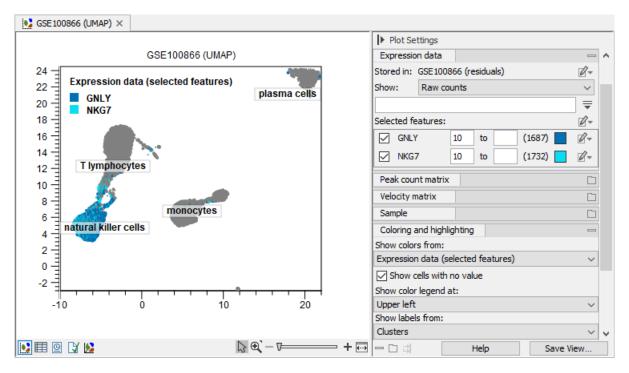


Figure 18: Expression of natural killer marker genes shows overlap with predicted natural killer cells.

We can additionally compare the expression of these two genes across the different cell types by launching the Create Expression Plot tool from the right-click menu.

The probability per cell type generated by the Predict Cell Types when **Output cell annotations** is ticked can also be visualized in the Dimensionality Reduction Plots, guiding manual refinement of the predicted cell types.

For details and more inspiration for manual exploration of the data, see the UMAP and tSNE plot functionality.