

How-to-DataExtract

2/16/2020

Introduction

What is This Thing?

The Data Extractor, or simply “DataExtract” for short, is a utility plugin for SeqGeq, to automatically de-convolute annotations saved in a CSV file for data loaded into SeqGeq.

Use Case

Say a researcher has annotated individual single cells in terms of their class (or “phenotype”). This type of categorical information can be imported into SeqGeq by dragging and dropping an appropriately formatted CSV file containing the classes per cell, onto the underlying data matrix within a SeqGeq workspace. However, the annotation labels will be lost during this conversion, and the categories will be encoded into integer values within a derived parameter.

DataExtract performs this merging operation for the categorical annotations automatically, and preserves the annotation information as populations within the sample.

Workflow

SetUp

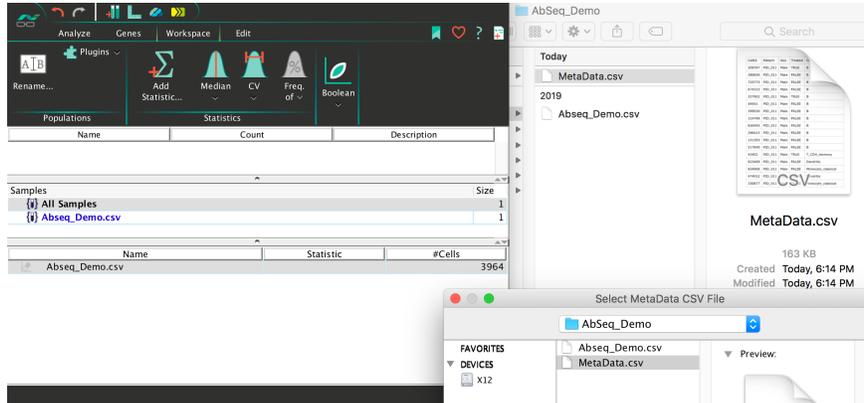
Load the plugin into your plugins folder, and restart SeqGeq to make the option available within the plugins dropdown.

Make sure the annotations CSV file (or “MetaData” file) is formatted with a “CellId” column that contains CellIds matching those in the corresponding expression matrix:

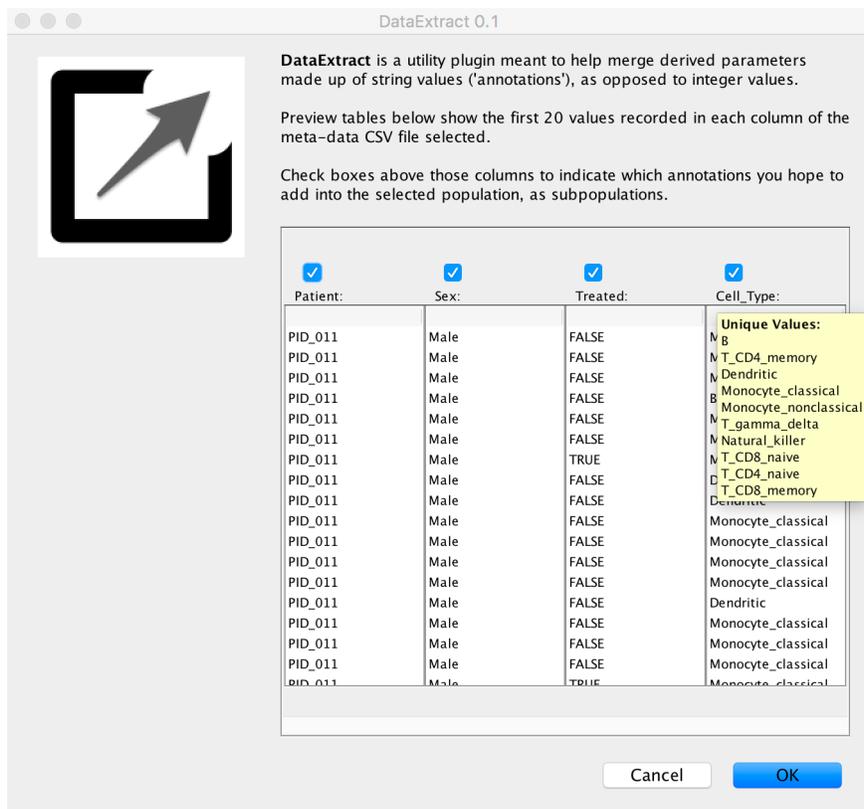
	A	B	C	D	E
1	CellId	Patient	Sex	Treated	Cell_Type
2	309787	PID_011	Male	TRUE	B
3	380656	PID_011	Male	FALSE	B
4	725775	PID_011	Male	FALSE	B
5	674315	PID_011	Male	FALSE	B
6	257902	PID_011	Male	TRUE	B
7	49541	PID_011	Male	FALSE	B
8	399530	PID_011	Male	FALSE	B
9	124788	PID_011	Male	FALSE	B
10	820495	PID_011	Male	FALSE	B
11	296413	PID_011	Male	FALSE	B
12	131355	PID_011	Male	FALSE	B
13	517849	PID_011	Male	FALSE	B
14	43452	PID_011	Male	TRUE	T_CD4_memory
15	823468	PID_011	Male	FALSE	Dendritic
16	834906	PID_011	Male	FALSE	Monocyte_classical
17	474012	PID_011	Male	FALSE	Dendritic
18	330877	PID_011	Male	FALSE	Monocyte_classical
19	186434	PID_011	Male	FALSE	Monocyte_classical
20	876162	PID_011	Male	FALSE	Monocyte_nonclassical
21	112410	PID_011	Male	FALSE	B
22	320244	PID_011	Male	FALSE	Dendritic
23	748463	PID_011	Male	FALSE	Monocyte_classical
24	850073	PID_011	Male	FALSE	Monocyte_classical
25	728345	PID_011	Male	FALSE	Monocyte_classical
26	808964	PID_011	Male	FALSE	Dendritic
27	272628	PID_011	Male	FALSE	Dendritic
28	708113	PID_011	Male	FALSE	Dendritic

Merging MetaData

Select the population of interest; where the merge should take place. Run the plugin and select the MetaData CSV file:

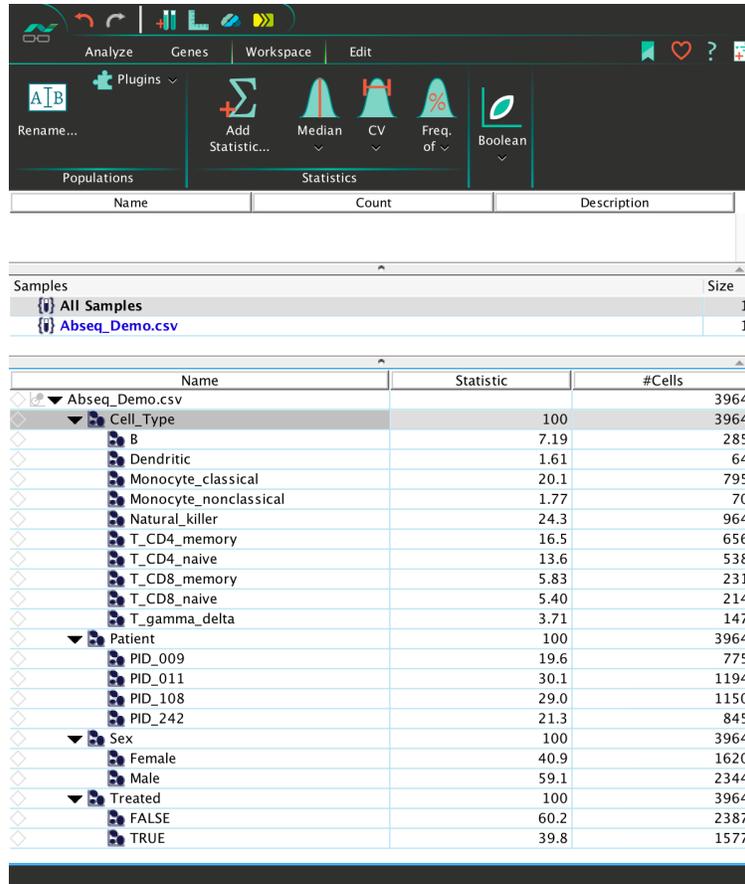


In the next dialog select columns of meta-data to be merged via check-box. Mouse over column headers to see the unique values within any given column (up to the first 24 values). Note, tables within the dialog will illustrate the first 20 values discovered within each column (including blank rows).



Outputs

When the plugin is finished running, parent populations for each derived parameter added will be created in the selected population, below which child populations representing each category will identify the subsets classified.



The screenshot shows a software interface with a menu bar (Analyze, Genes, Workspace, Edit) and a toolbar with icons for 'Add Statistic...', 'Median', 'CV', 'Freq. of', and 'Boolean'. Below the toolbar is a table with columns 'Name', 'Count', and 'Description'. Underneath is a 'Samples' table with columns 'Name' and 'Size'. The main part of the screenshot is a large table with columns 'Name', 'Statistic', and '#Cells', showing a hierarchical view of data for 'Abseq_Demo.csv'.

Name	Statistic	#Cells
Abseq_Demo.csv		3964
Cell_Type	100	3964
B	7.19	285
Dendritic	1.61	64
Monocyte_classical	20.1	795
Monocyte_nonclassical	1.77	70
Natural_killer	24.3	964
T_CD4_memory	16.5	656
T_CD4_naive	13.6	538
T_CD8_memory	5.83	231
T_CD8_naive	5.40	214
T_gamma_delta	3.71	147
Patient	100	3964
PID_009	19.6	775
PID_011	30.1	1194
PID_108	29.0	1150
PID_242	21.3	845
Sex	100	3964
Female	40.9	1620
Male	59.1	2344
Treated	100	3964
FALSE	60.2	2387
TRUE	39.8	1577

Trouble-Shooting

- Some parameter combinations may not be amenable to merging simultaneously. If you see an error for particular combination of derived parameters, try selecting fewer parameters to merge.
- Make sure the MetaData CSV file contains a CellId column (first column) with CellId values that exactly match the CellIDs in your corresponding expression matrix.
- If the MetaData file contains header information in rows above the derived parameters, try removing that information before merging.

If you have additional questions don't hesitate to reach out: seqgeq@bd.com