


Annotation Comparison Explorer (ACE)

ACE is a shiny app for comparison of annotations such as (i) cell type assignments (e.g., from different mapping/clustering algorithms), (ii) donor metadata (e.g., donor, sex, age), and (iii) cell metadata (e.g., anatomic location, QC metrics). Additionally, this tool can compare results across more than two taxonomies, for example for annotating a novel taxonomy with information from multiple existing taxonomies or for comparison of data from multiple studies of Alzheimer's disease.

- ACE is currently hosted internally at the Allen Institute but will be posted on shinyapps.io in the near future. If you need access, [email Jeremy](#).
- Link to codebase: https://github.com/AllenInstitute/annotation_comparison/tree/dev

Note that this app is under active development and may not match the content of the User Guide exactly. If you run into issues or more generally have any thoughts or opinions about the app, please contact me via the big button on the app or [by clicking here](#).

 [Comments/bugs?](#)

Getting started

If you do not work at the Allen Institute, follow instructions on GitHub https://github.com/AllenInstitute/annotation_comparison/tree/dev.

If you work at the Allen Institute, the easiest way to use this app is to go to the hosted link. This only works if you are on site at the Allen Institute or have VPN access (I think). You also need an account on the relevant server. You need to log-in to use it with your normal login (subtracting the @alleninstitute.org part). If you are an Allen Institute employee and don't have access, you can sign up for an account at the sign in page:

 **posit** Connect

[Sign Up](#)

Log In

Log in with your Posit Connect credentials

Username *

Password *

[Forgot your password?](#)

Log In

Don't have an account? [Sign Up](#)

Data selection

At its core, ACE is a method for comparing different cell-level annotations and (optionally) providing additional information about each annotation. The primary input is a table where rows correspond to cells and columns correspond to metadata such as cluster assignments, mapping results, different levels of the taxonomy hierarchy, donor information, or cell QC metrics:

	A	B	C	D	E	F	G	H
1	sample_name	SEAAD_supertype	subclass	class	GA_cluster	CA_cluster	sex	donor_name
2	AAACCCACAACCTC	Pax6_1	Pax6	GABAergic	Pax6_1	Pax6_1	M	H18.30.002
3	AAACCCACACGGT	L5/6 NP_1	L5/6 NP	Glutamatergic	L5/6 NP_1	L5/6 NP_3	M	H18.30.002
4	AAACCCCACTCT	L5 IT_7	L5 IT	Glutamatergic	L5 IT_7	L5 IT_1	M	H18.30.002
5	AAACCCACATCAG	L6 CT_2	L6 CT	Glutamatergic	L6 CT_2	L6 CT_1	M	H18.30.002
6	AAACCCAGTGTCG	L4 IT_2	L4 IT	Glutamatergic	L4 IT_2	L4 IT_2	M	H18.30.002
7	AAACGAAAGCAC	Astro_1	Astrocyte	Glia	Astro_1	Astro_4	M	H18.30.002
8	AAACGAACACAA	L5 IT_2	L5 IT	Glutamatergic	L5 IT_2	L5 IT_1	M	H18.30.002
9	AAACGAACACGCT	L2/3 IT_5	L2/3 IT	Glutamatergic	L2/3 IT_5	L2/3 IT_3	M	H18.30.002
10	AAACGAAGTACCC	L5 IT_2	L5 IT	Glutamatergic	L5 IT_2	L5 IT_1	M	H18.30.002
11	AAACGAAGTTCCG	L5 IT_2	L5 IT	Glutamatergic	L5 IT_2	L5 IT_1	M	H18.30.002
12	AAACGAAGTTGCC	Vip_9	Vip	GABAergic	Vip_8	Vip_8	M	H18.30.002

Metadata information can be provided by either selecting one of the example tables (via #1 below) or inputting a location to a cell-level annotation file (via #2 below):

Annotation Comparison Explorer (ACE)

Select data set

Select an annotation table:

Alzheimer's cell mapping

1

USER GUIDE

Comments/bugs?

4

Input location of cell-level annotation information

https://raw.githubusercontent.com/AllenInstitute/annotation_comparison/dev/data/DLPFC_

2

Location of metadata (e.g., cluster) information (optional; csv file)

https://raw.githubusercontent.com/AllenInstitute/annotation_comparison/dev/data/AD_stu

3

Dataset description

Data and associated cell type assignments from multiple studies of Alzheimer's disease. All data sets were mapped to SEA-AD data and their mappings as well as original cluster assignments are included in the tables. In addition, each cell type's change in abundance in AD from the original study, as well as some basic information about the cell types are included. Data is subsampled to 100 cells per SEA-AD supertype. The way data is encoded, comparisons between each data set an SEA-AD are valid, but comparisons CANNOT be accurately made between external data sets.

5

Locations can be either to files on GitHub (and probably other URLs, but I haven't checked) or to a location on the server (e.g., “//allen/programs/celltypes/workgroups/...”). At those locations multiple file types are accepted as input:

- A csv file, like the one shown above.
- A gzipped csv file (e.g., XXXXX.csv.gz)
- A directory for visualization of taxonomies on molgen-shiny (the anno.feather file is read)
- An h5ad file in scrattch.taxonomy format (the obs field is read)

The second file (#3 above) is optional, but if provided must a csv file with information about each individual piece of metadata. Specifically, each row corresponds to a piece of metadata (e.g., a specific cluster or subclass) and columns correspond to whatever information about the metadata that you want to share. Here is one example that provides some information about cell sets from SEA-AD:

1	cell_type	level	study	direction	direction_new	description	notes
2	exc	class	SEA-AD	none	not_assessed	All excitatory neurons	
3	glia	class	SEA-AD	none	not_assessed	All non-neuronal cells (glial and non-neural types)	
4	inh	class	SEA-AD	none	not_assessed	All inhibitory neurons	
5	Astro	subclass	SEA-AD	up	up	Astrocytes	
6	Chandelier	subclass	SEA-AD	none	not_assessed	Chandelier MGE interneurons	
7	Endo	subclass	SEA-AD	none	not_assessed	Endothelial cells	
8	L2/3 IT	subclass	SEA-AD	down	down	Layer 2/3 intratelencephalic neurons	
9	L4 IT	subclass	SEA-AD	none	not_assessed	Layer 4 intratelencephalic neurons	
10	L5 ET	subclass	SEA-AD	none	not_assessed	Layer 5 extratelencephalic neurons	

This is currently a work in progress and is quite buggy (e.g., it only works for the SEA-AD example above), and I'd appreciate feedback here (via #4 above)! The only required field is something called “**direction**” which can be “down”, “up”, “none”, or [anything else] and which determines color-coding. I'll improve documentation about this table once I finalize it.

For any pre-loaded annotation tables, a description will pop up describing what is in that data set and what it's general use is (#5). I'm working on making the app a bit more flexible so it will only show relevant visualizations, but for now, all the panels are shown whether or not they make sense.

Filtering your data

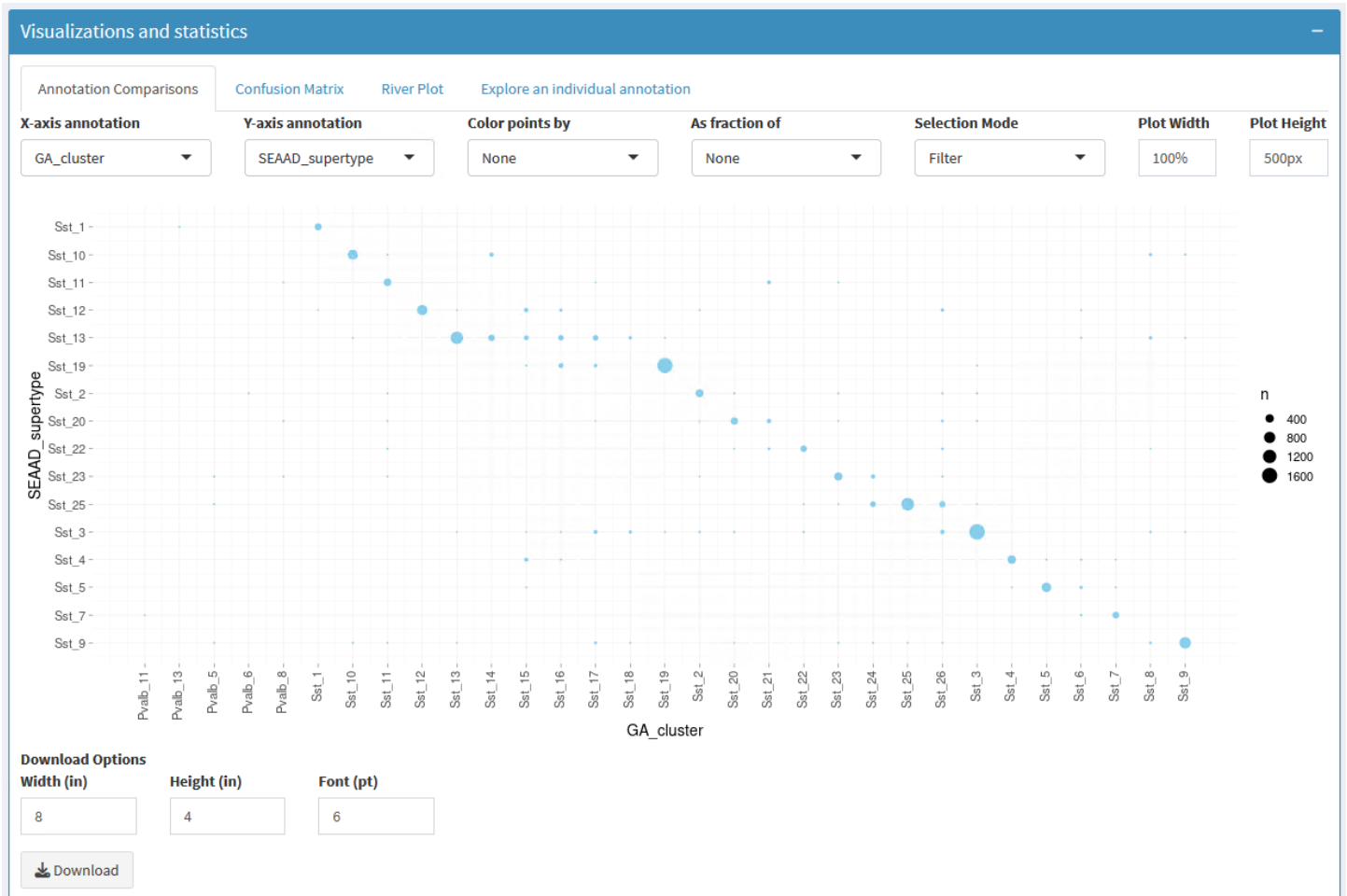
Cell filtering is a critical component of ACE, as it allows you to define the context for all the visualizations and statistics. For example, since there are >5000 cell types, it is not practical to perform comparisons on all of them at once (and at least one of the visualizations will not allow such large amounts of data to load). What is practical, and much more useful, is to see how all the inhibitory cells collected from mouse Basal Ganglia and map to the human whole brain cell types also in basal ganglia. As another example (see below), we can restrict our visualizations to only SST cells collected from the MTG of males, which dramatically decreases the size and scope of the data sets. Here is how that filtering looks:

The screenshot shows the ACE app's filtering interface. It is divided into three main sections. The top-left section, titled "Choose Filter Set", contains two input boxes: "subclass" and "sex". A red circle with the number "1" is drawn around these two boxes. The top-right section, titled "Filter for:", contains two input boxes: "subclass" and "sex". The "subclass" box has a dropdown menu open showing "Sst", and the "sex" box has a dropdown menu open showing "M". A red circle with the number "2" is drawn around these two boxes. The bottom-left section shows the results of the filtering: "subclass: Sst", "sex: M", and "9281 of 137303 samples selected." A red circle with the number "3" is drawn around this section.

The “Choose Filter Set” box (#1) allows you to select one or more metadata columns on which to perform the filtering. This box (and all other boxes in the app where you can type) have autocomplete and will help you select possible options to select. For each chosen variable, the “Filter for” box (#2) allows you to choose how to filter the data. Categorical variables can be filtered by listing all of the values for that variable that you’d like to include in the visualizations for the app. For example, to retain all SST cells, you could either set the “Subclass” filtered to SST (as done in this example) or you could set the “SEAAD_supertype” filter to include every SST cell type (or both). Numeric variables allow you to set a numeric range within which valid cells apply. For example, if there was a column corresponding to number of genes detected, you could set a minimum threshold of 500. Finally, the bottom left corner of this box (#3) shows the filters that have been applied, as well as the number of cells out of the total table that will be included for other app components.

Visualizations and statistics: Annotation Comparisons

The main components of ACE are found in the ‘Visualizations and Statistics’. Each tab within this box corresponds to a different visualization that applies to the filtered data set defined above. An example of the first tab (‘Annotation Comparisons’) is shown below:

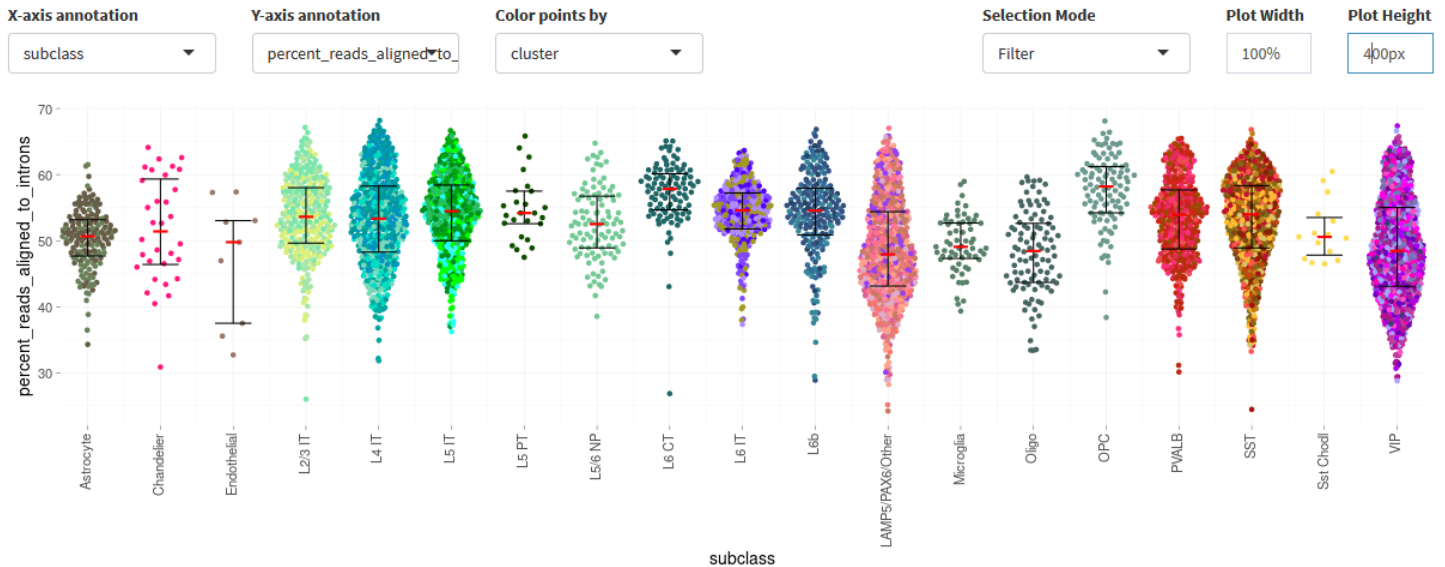


This tab allows comparison of any two pieces of categorical or numeric metadata. Above shows such an example, comparing cell type assignments of neurons defined as SST in SEA-AD with their original cell type assignments in the great ape (GA) study. As you can see, there is high correspondence between them, which makes sense, since SEA-AD supertypes were defined using great ape assignments as a starting point. *Note that this tab and the “Confusion Matrix” tab are highly related when comparing two categorical variables.* All of the tabs are laid out roughly the same way, with controls at the top, the main visualization in the middle, and download or other content on the bottom. This panel includes the following controls:

- **X-axis annotation:** which annotation will be shown on the horizontal axis?
- **Y-axis annotation:** which annotation will be shown on the vertical axis?
- **Color points by:** Which metadata should points be colored by? For categorical visualizations this isn't particularly informative, but for numeric comparisons it can help show structure within a swarm plot (see below).
- **As a fraction of:** By default, this is set to “None”, in which case the size of the points corresponds to the number of cells in that intersection. If changed to rows or columns, this converts the points to fractions where the values row or column sum to 100%.
- **Selection Mode:** By default this ensures the plots only show the filtered data. If set to “highlight,” all data will be shown, with the remaining data not part of the filter colored as grey. I'd recommend leaving this as the default.

- **Plot Width / Plot Height:** Parameters that will change the width (in %) and height (in pixels) of the visualization on your screen.

Here is another example, where in this case the vertical axis corresponds to a numeric variable, with points color-coded by cluster so that different points in the swarm are different colors. By eye I don't see any differences between clusters within a subclass for percent reads aligned to introns, but in theory that could be seen here.



Finally, the bottom of this tab (and several others) includes an option to download the current visualization, along with a few parameters for the size of the images and included text. These defaults are not always reasonable, so it may be worth trying a few downloads to see which one looks best.

Download Options

Width (in)


8

Height (in)

4

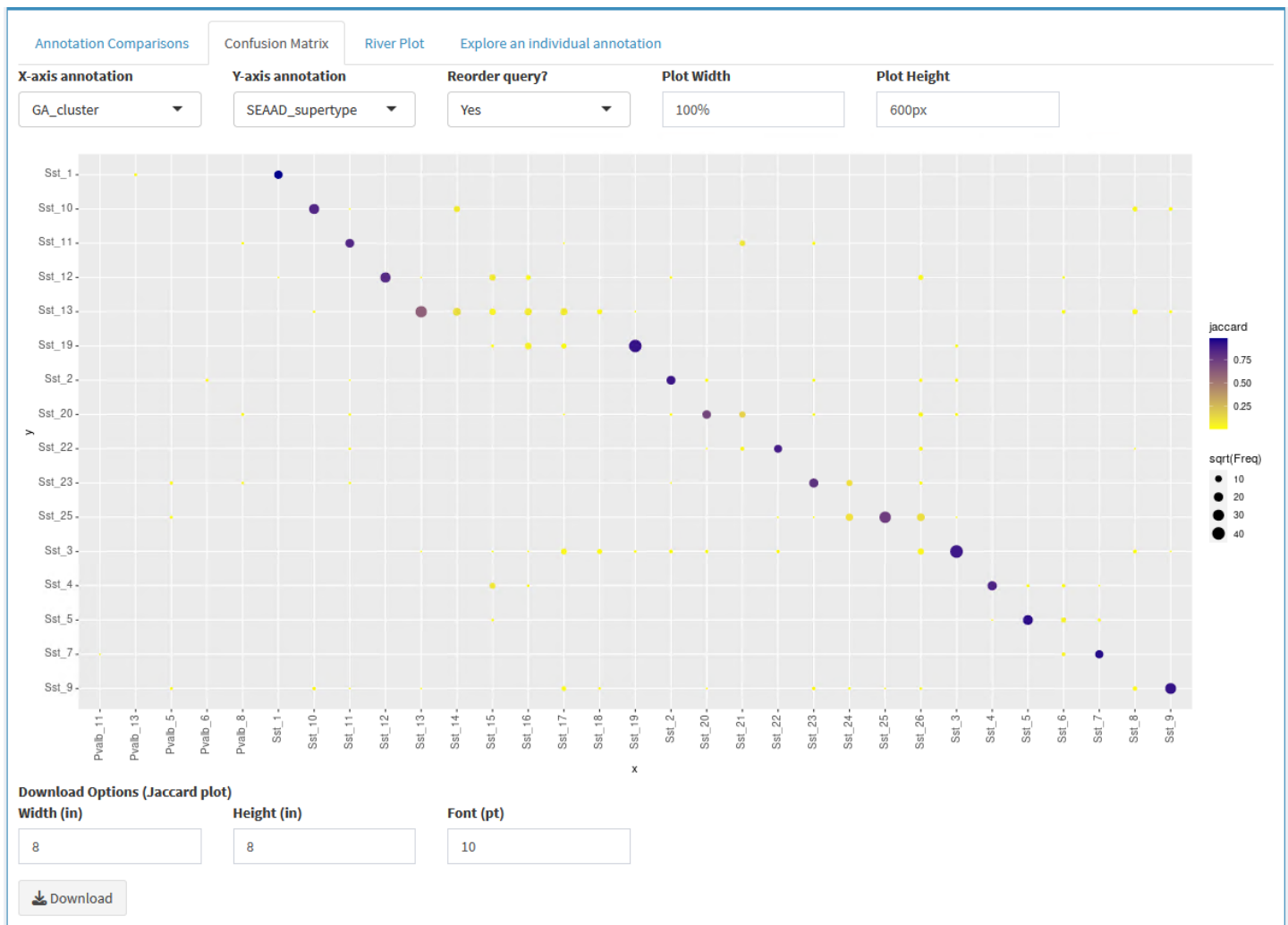
Font (pt)

6

 Download

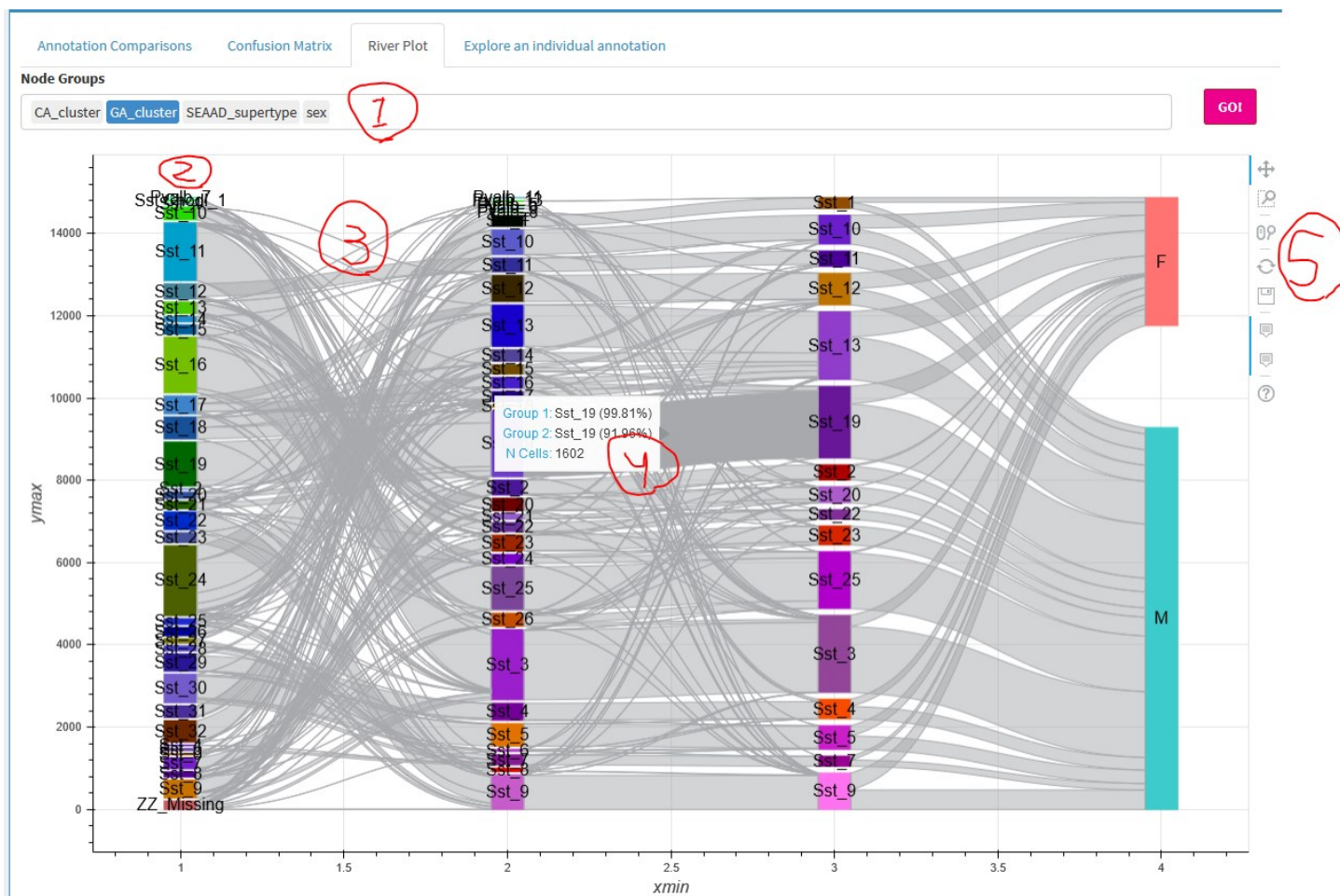
Visualizations and statistics: Confusion Matrix

The second plot shows a confusion matrix between any two categorical variables. This will look VERY similar to the 'Annotation Comparison' tab when the same x and y annotations are included but has two notable differences. First, the "Reorder query?" parameter will attempt to sort the entries within the y-axis metadata to create as good looking of a diagonal as possible. This is extremely useful if the metadata are not sorted in the same way (or at all) and is set to 'Yes' by default. Second, the dots are colored by Jaccard distance, with blue values indicating a better correspondence and yellow values indicating a poorer correspondence. *I plan to merge this tab with the Annotation Comparison tab and am open to suggestions for the best way to do this.*



Visualizations and statistics: River Plot

This panel creates an interactive river plot (also known as a “Sankey diagram”) that allows you to compare two or more pieces of metadata (see below). In the Node Groups box (#1) you can enter any number of metadata files in order (the order matters!), and then click “Go!” to generate the river plot. For each metadata you get a stacked bar plot showing the number of cells that have each value within that metadata, with the value for each label shown (#2). In addition, for each adjacent pair of bar plots, you get “rivers” connecting each pair of metadata values together, where the thickness of each line represents the number of cells sharing the corresponding values from the two metadata fields (#3). Since this plot is interactive, you can hover over any bar or river and see what values and numbers correspond to what you are hovering over (#4). On the right side of the plot (#5) there are some extra controls that allows you to zoom, pan, and interact with the plot in various other ways. Finally, you can download the plot using the buttons on the bottom. Currently there is no way to resize the box or to show the same metadata more than once in the Node Groups sequence, and this may or may not be updated later.



Visualizations and statistics: Explore an individual annotation

As the name suggests this tab is focused on exploring individual annotations (see below). This can be useful for manual annotations of individual clusters (or supertypes, subclasses, etc.) or for understanding how cell types defined in one study compare with cell types defined in other studies. **This tab will only load if an appropriately formatted metadata table is provided** (I plan to relax/clarify this constraint in the future, but for now it needs to be a csv file with at least one column called “direction”, as described above).

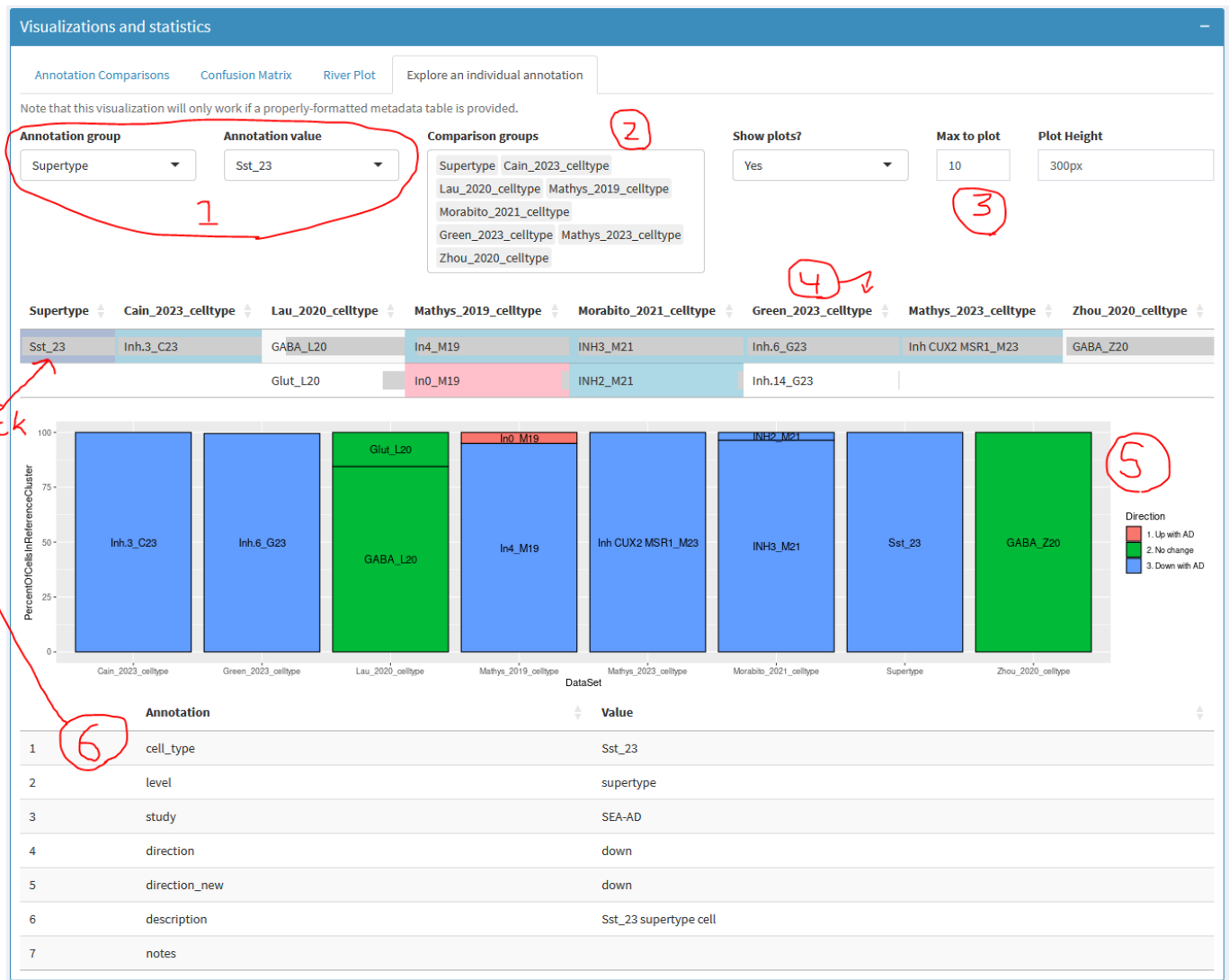
Location of metadata (e.g., cluster) information (optional; csv file)

https://raw.githubusercontent.com/AllenInstitute/annotation_comparison/dev/data/AD_study_cell_types_for_

Unlike the other tabs which compare all values for two or more pieces of metadata, this tab is anchored around a single “**Annotation value**” from a single “**Annotation group**” or metadata column (#1 below). Once this is chosen, the rest of the tab will be performed in comparison to that one annotation value. The other controls at the top indicate which other “**Comparison groups**” or specific metadata columns will be compared against the starting annotation value (#2) and whether/how this information will be displayed as a bar plot below the default display as a chart (#3, more on that shortly). Specifically,

- **Show plots?:** A Yes/No call on whether the barplot should be shown
- **Max to plot:** The maximum number of top overlaps to show in the barplot (for example, if there are 20 SST clusters and “Annotation value” is set to SST subclass, then only the 10 most abundant SST clusters would be shown in the plot).

- **Plot Height:** Height of the plot in pixels.



The next section (#4) shows a chart of the selected annotation data along with the corresponding values from the comparison groups, sorted descending from highest to lowest overlap. Within each table entry, a grey bar indicates the fraction of cells from the selected annotation value that also have the comparison value listed. Boxes are also color-coded with whether they values are up (red), down (blue), or unchanged (white) with Alzheimer's disease (I plan to make this a more generic coloring schema in the future). If shown the bar plot (#5) displays the same information using a different graphic, with each bar indicating the fraction of cells from the selected annotation value that also have the comparison value listed, again colored by change in AD (there are some issues with coloring and ordering of bars that I'm still working out). *Both the chart and the plot are basically river plots without the rivers for a single metadata value.*

The final piece of this table is a table that displays relevant metadata for a given metadata value (#6). If you click on any box in the chart (example above), then a table showing the provided metadata for that specific cluster will be shown below. I'm working with David Osumi-Sutherland on integrating this part with the information available in Taxonomy Development Tools, and more generally would appreciate any suggestions on how to make the information here (or the way to code it in) better.