###### Data Pre-processing for Distance Matrix Calculation

**Initial filtering:**

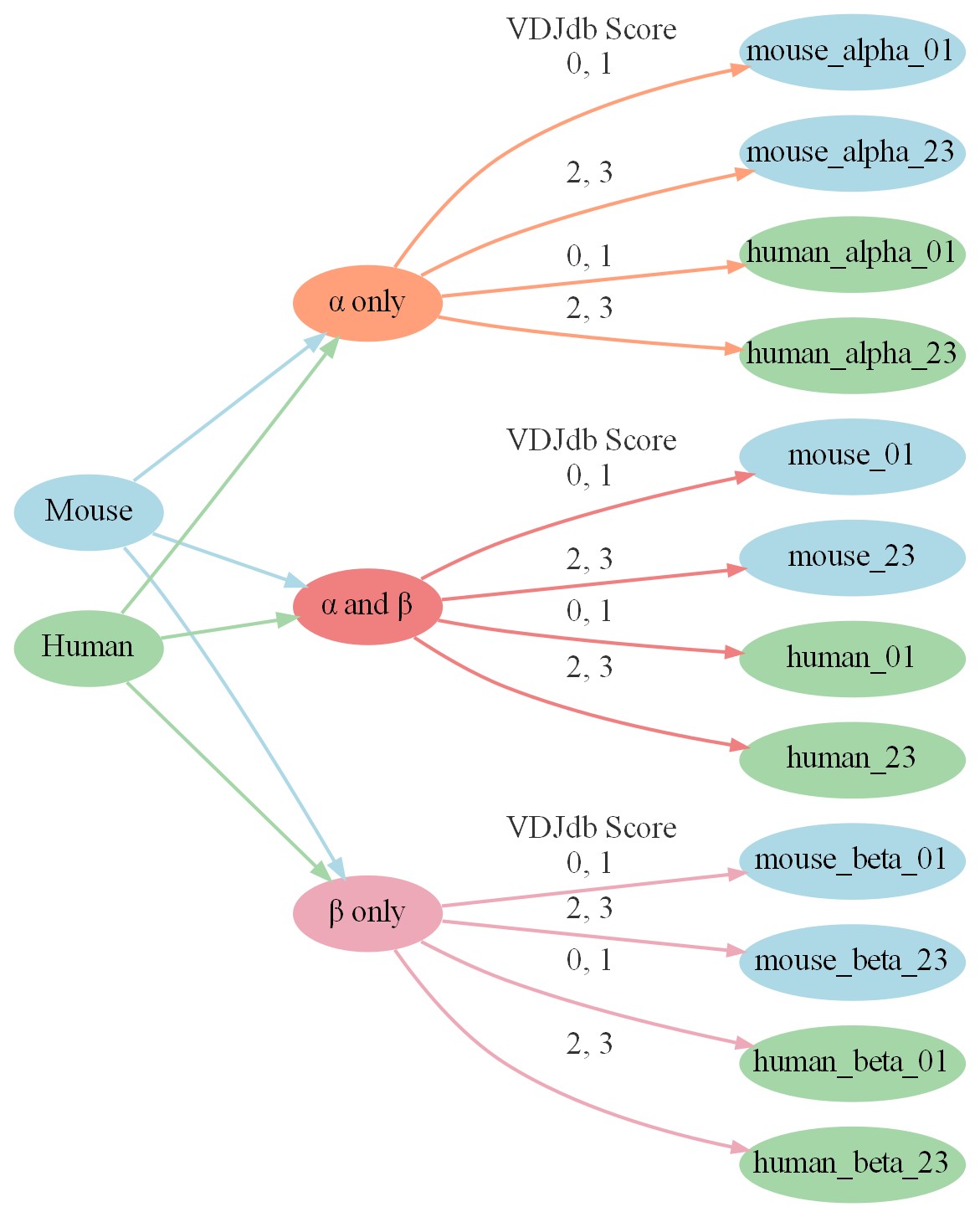
The data contains a total of 21 columns, we can make an initial filter by the type and meaning of each column: Some of the columns contain specific information about the literature sources (reference.id), sequencing methods (method, web.method.seq), collection method (web.method) etc., some of which are related to how vdjdb.score is calculated. Therefore we first remove these columns (reference.id, method, meta, cdr3fix, web.method, web.method.seq, web.cdr3fix.nc, web.cdr3fix.unmp).

**Data Formatting:**

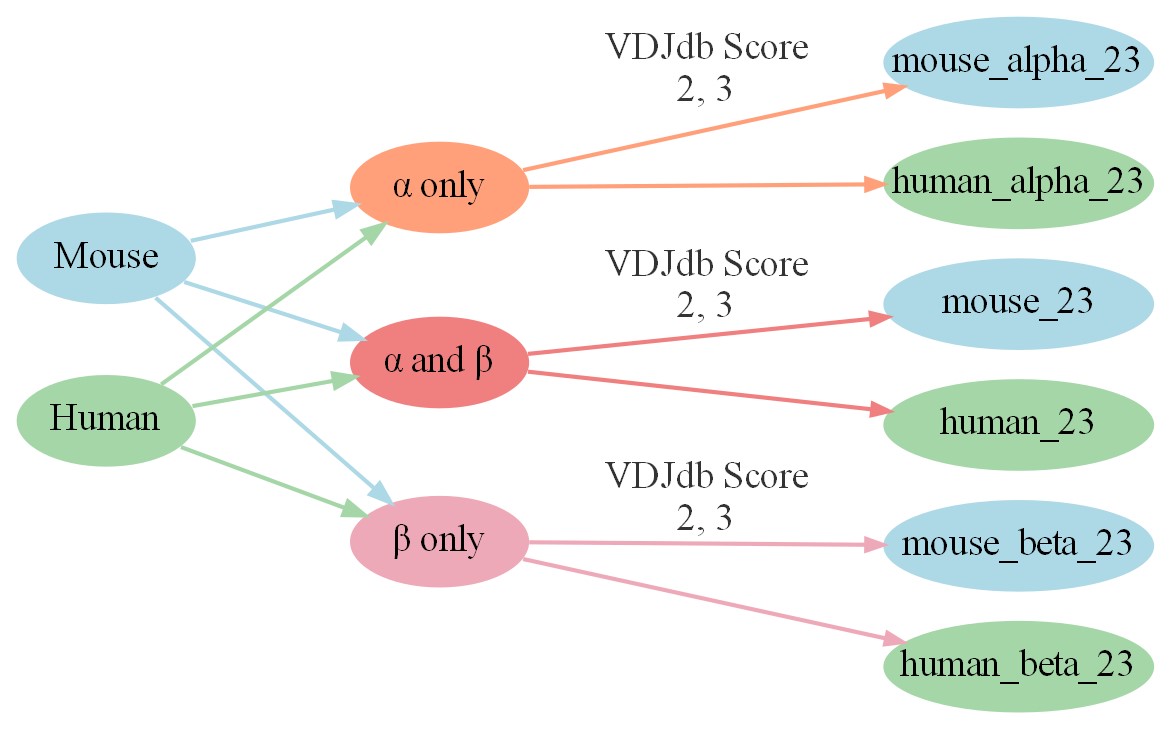
Due to the fact that we are using TCRdist3 to calculate thee distance matrix, we need to format the data as the package requires. Firstly, a new column labeled 'clone\_id' was added to all rows in the data. Subsequently, we replicated the entries with a complex ID of 0 and adjusted the gene columns to fabricate a virtual pairing of chains, preparing the data for the subsequent step of pivoting the data wider. Then, we also set a new column 'count' of all rows to 1. Based on the column 'gene', we pivot the data wider so that it looks like the standard input. Meanwhile, we change the name of the columns as the TCRdist requires.

Most classes and functions in tcrdist3 require specification of an appropriate host organism. Currently only ‘human’ or ‘mouse’ are supported. Consequently, we excluded the line 2073 entry labeled 'Macaca mulatta'. Furthermore, we update the names to 'modern humans' and 'mouse', respectively, to enhance clarity and consistency within the dataset.

Last, we get the formatted input of TCRdist3 and it is arranged in alphabetical order by epitope. Also, we divided the data into different subsets according to different chains, VDJdb scores and species. Since the TCRdist3 will not take the rows of null values into account, we then drop the null values in all of the subsets. Then, we dropped duplicated rows of the formatted inputs.



Due to the low confidence associated with subsets containing VDJdb scores of 0 and 1, we aim to create distance matrices exclusively from subsets with VDJdb scores greater than 1. This decision is motivated by the desire to improve the reliability and accuracy of the distance matrices used in subsequent analyses. By excluding subsets with low VDJdb scores, we aim to mitigate the potential noise and uncertainty. Distance matrices derived from subsets with higher VDJdb scores are expected to better reflect the true similarities and dissimilarities between TCRs, thereby enhancing the robustness of further analyses, such as clustering, dimensionality reduction, and visualization.



###### Calculating the Distance Matrix

**Functions to calculate the distance matrix and access its quality**

Utilizing the functions we created, we get outputs of evaluation results, the distance matrix as well as 3 subplots of Distance Matrix, Epitope Matrix, TCRs in 2 dimensions:

1. Distance Matrix

The distance matrix is a heatmap of the matrix. This coloring helps to quickly identify regions of similarity or difference. Since the diagonal of the matrix represents the distance between an element and itself (which is zero), its darker than the rest of the matrix. This is a visual cue that confirms that each element is closest to itself, as it should be. Darker colors represent smaller distances (or similarities).

2. Epitope matrix

The Epitope matrix heatmap represents an idealized distance matrix, where the rows and columns of the matrix correspond to unique TCR indices within the subset data. The color of each cell reflects the relative relationship between different TCRs based on their epitopes:

- If two TCRs share the same epitope, the cell is dark (high correlation, closer distance).

- If the epitopes of two TCRs are different, the cell is light (low correlation, greater distance).

- If all rows belong to the same epitope, the image would be an entire dark-colored graph.

- Each square on the diagonal represents a unique epitope.

We can evaluate the quality of the distance matrix by comparing the Distance Matrix and the Epitope matrix. The more similar the two images are, the better the distance matrix is.

3. TCRs in 2 dimensions

The TCRs in 2 dimensions draws high-dimensional data points in a 2D space using dimensionality reduction techniques such as PCA, t-SNE, or UMAP. Users specify a range (from rank X to rank Y) for epitopes containing the most points, which are then colored accordingly. Each data point is colored based on its unique epitope, with different epitopes represented by distinct colors and encircled. The size of each point reflects the number of points belonging to its epitope, with larger sizes indicating higher counts. Users can choose to display additional points beyond the specified range (epitopes contain less points than rank Y) in light grey.

**Methods for assessing the quality of distance matrices**

1. Compare the Distance Matrix and the Epitope matrix: The more similar the two images are, the better the distance matrix is.

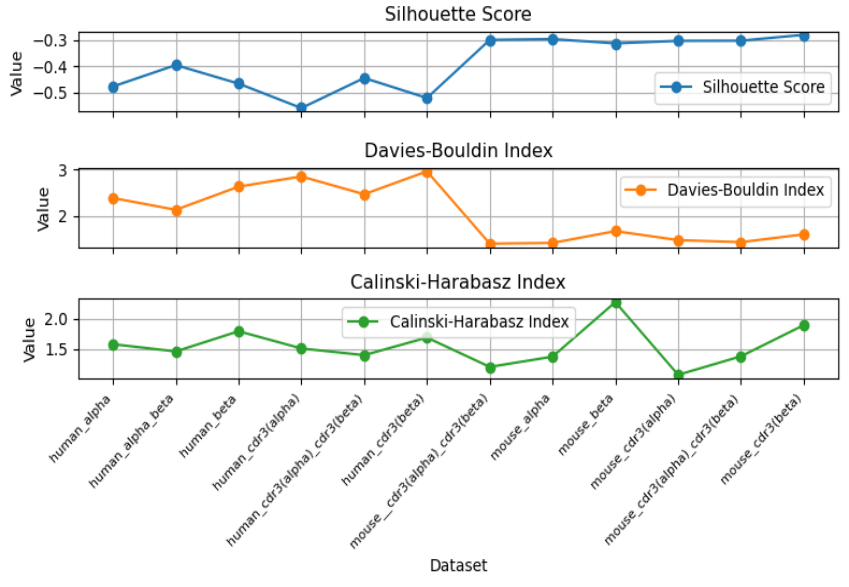
2. TCRs in 2 dimensions: The distance matrix is better if points of the same color (epitope) are closer to each other and further apart from points of other colors (epitope).

3. Evaluation results: A list containing evaluation results for each subset and the corresponding line plots.

- Silhouette score: The closer to 1, the better.

- Davis-Bouldin index: The smaller the index value, the higher the tightness within the cluster and the better the separation between clusters.

- Calinski-Harabasz index: The larger the index value, the better the clustering result.



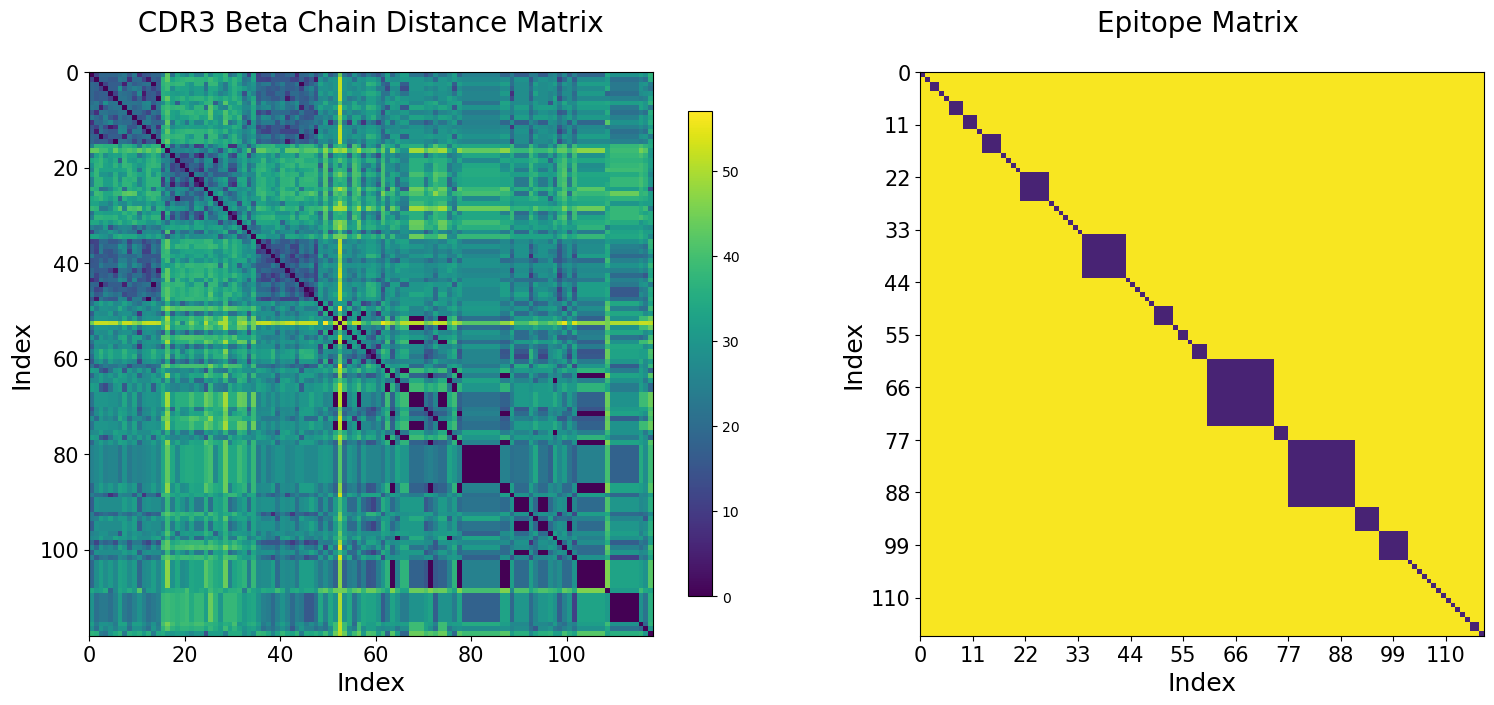
**Findings**

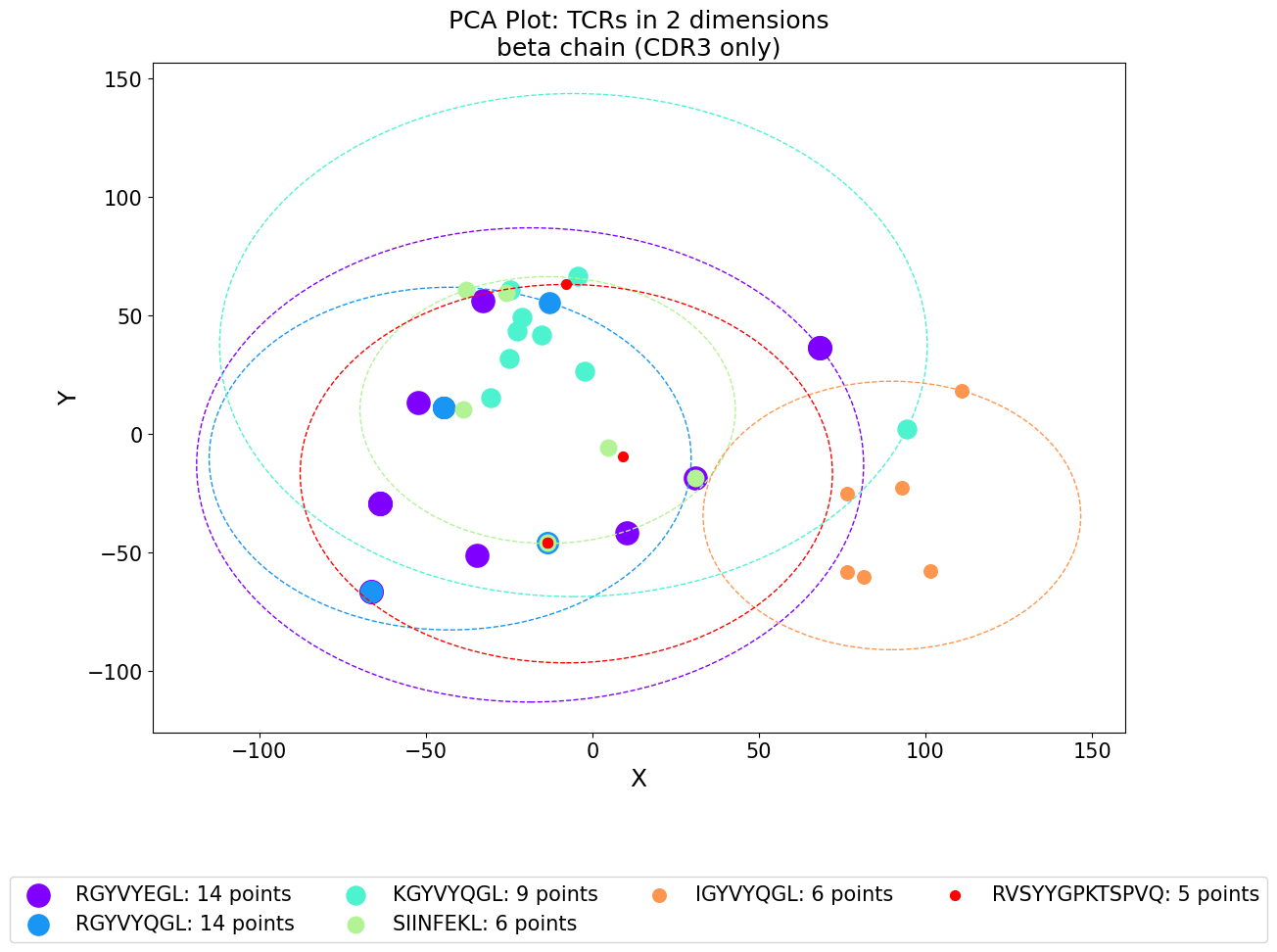
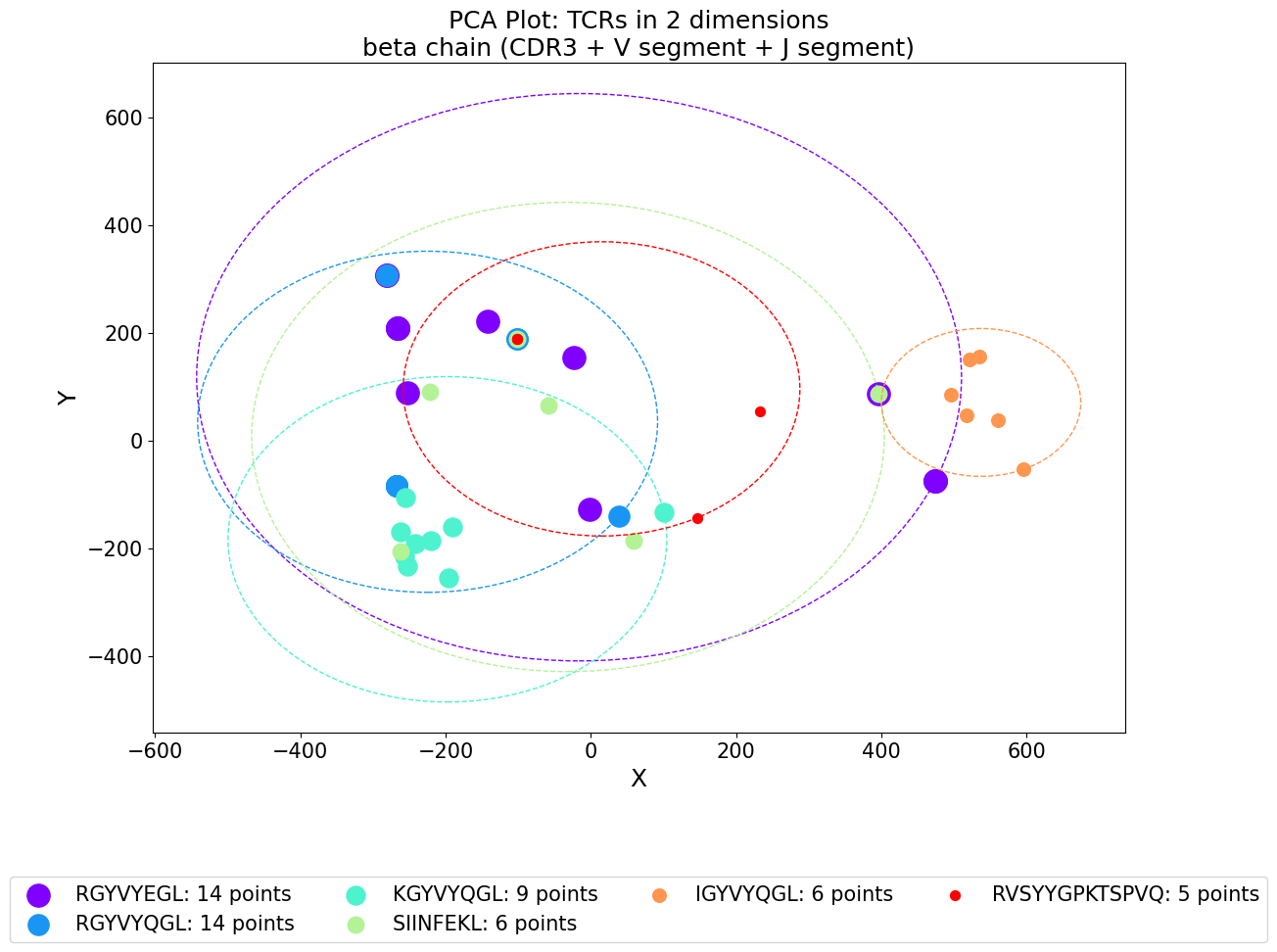
According to different images in different data sets and the evaluation results, we found:

1. In the mouse datasets:

According to the plots, the distance matrix obtained by applying the beta chain (both the entire beta chain and the partial beta chain with only CDR3) calculation works better. Not only the Distance Matrix and the Epitope matrix are very similar, but also for the top 6 epitopes containing the most number of data points plotted, it can be seen that some of the data points of the same color are closer to each other relative to the other colors.

Although the similarity between the Distance Matrix and the Epitope matrix can be observed in the distance matrix computed in other ways, all of the points in the TCRs in 2 dimensions image are clustered and overlapped, and it is difficult to separate the TCRs with different epitopes from each other.





From the evaluation results, the Silhouette scores are similar, and the Davis-Bouldin indecies are also in a similar range. By observing the Calinski-Harabasz index image, we can find large differences. The Calinski-Harabasz indecies of the entire beta chain and the partial beta chain (CDR3 only) are much larger than other categories, indicating that the distance matrix calculated in this way is better.

Combined with the conclusion of evaluation results, the distance matrix calculated in this way also has a better metric score. Therefore, it is highly possible that utilizing the beta chain (both the entire beta chain and the partial beta chain with only CDR3) works better for the mouse datasets.

1. In the human datasets:

Under almost all calculation methods, there is no obvious similarity between Distance Matrix and Epitope matrix. Only in the case of using the entire beta chain (CDR3, V segment and J segment) a slight similarity is observed in the upper left corner of the image.

Since the TCRs in 2 dimensions image of the human dataset has more data points and most of the data points are overlapped, we chose the epitopes with the lower rank (according to the number of data points) for mapping. It can be observed that some of the points of the same color can be clustered together by combining both chains (both the entire chain and the partial chain with only CDR3).

According to the evaluation results, from the Silhouette scores and the Davis-Bouldin index, we can see that distance matrix calculated by the combination of the two chains (the whole chain and the CDR3 only) is better, and they have higher Silhouette scores and a relatively slightly lower Davis-Bouldin index. While there isn't a significant difference in the Calinski-Harabasz index, both the entire beta chain and the partial beta chain (CDR3 only) demonstrate slightly higher Calinski-Harabasz indices. This suggests that the distance matrix calculated in this manner may also be better.

Combined with the conclusion of evaluation results, the above speculations are consistent with each other. So we speculate that beta chain (both the entire chain and the partial chain with only CDR3) and combining both chains (both the entire chain and the partial chain with only CDR3) work slightly better for the human datasets.

