

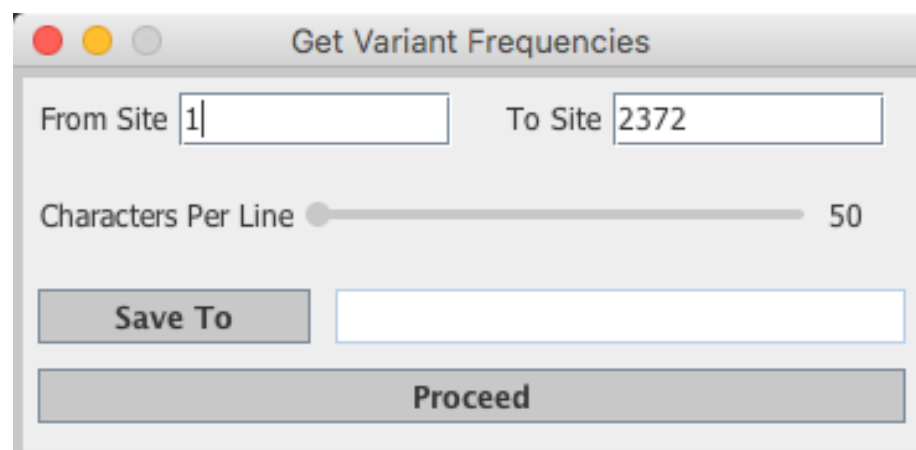
## Identifying variants using CView

This PDF follows outlines how a list of variants can be obtained using CView. Details about the software can be obtained from a preprint version of the manuscript, titled “CView: A network based tool for enhanced alignment visualization”, (@: <https://www.biorxiv.org/content/10.1101/2022.01.17.476623v1>) and the CView wiki (@: <https://sourceforge.net/projects/cview/>).

Once a fasta formatted alignment has been loaded into CView using the “**All Sequences**” -> “Load fasta” menu option unique variants can be obtained that represent: (i) all sequences within the alignment, (ii) sequences matching a user search criteria, (iii) sequences passing through a selected node on the network or (iv) a user specified set of sequences. Here we will briefly go through each of these options using the use-case scenario data described in the CView manuscript and that is available from the Zenodo repository (<https://doi.org/10.5281/zenodo.6475666>).

### (i) Variants from all sequences within the alignment

To obtain a list of unique variants from the aligned sequences select the “**All Sequences**” -> “Variant Frequencies” menu option. This will open the popup window depicted in figure 1.



**Figure 1:** Popup window for “**All Sequences**” -> “Variant Frequencies” menu option.

Select the sites of the alignment, using the “From Site” and “To Site” text boxes that you wish the variants to represent. By default this is set to the entire length of the alignment, but in the majority of user cases it is likely that a more localized region will be of interest.

The “characters per line” slider defines the number of characters per line within the output file for the fasta formatted variant sequences. Sequences longer than this will be split into multiple lines (identical to that of fasta formatted sequences).

The “Save To” button allows the user to select the output file to where variant information will be placed.

Once the “Proceed” button is selected CView will extract all unique character permutations from between the specified co-ordinates, along with their frequency of occurrence, and output this information to the output file. For each variant present the format is as follows:

```
>VARIANT_1_FREQUENCY_11
ATGAGAGTGAAGGGGATCA---GGAGGA-----ATTA
TCA---GCGC---TT---ATGGACATG-----GGGCACCT
```

The title indicates that the sequence permutation associated with Variant 1 was observed 11 within the sequences (between the specified co-ordinates). Below this fasta-formatted list of all variants, the titles of the individual sequences represented by this residue permutation are provided:

```
>VARIANT_1_FREQUENCY_11
  B.US.-.101_TP3_14b.FJ798323
  B.US.-.101_TP3_18.FJ798326
  B.US.-.101_TP3_23.FJ798328
  B.US.-.101_TP1_18b.FJ798341
  B.US.-.101_TP1_25b.FJ798347
  B.US.-.101_TP1_42.FJ798356
  B.US.-.101_TP1_6.FJ798362
  B.US.-.101_TP1_9.FJ798365
  B.US.-.101_TP1_22.FJ798570
  B.US.-.101_TP1_33.FJ798571
  B.US.-.101_TP3_36.FJ798575
```

More generally information within the output file is presented as follows:



user wishes to replace a bar with a white space, it makes it possible to produce an image that visually highlights any individual characters that were different from the character present within the most frequent variant. For example with some simple manipulation (mainly the removal of the last 14 character of each line to fit on the width of this document line and replacing “|” with “ ”) the previous example can be displayed as:

```
>V_1 ATGAGAGTGAAGGGGATCA---GGAGGA-----ATTATCA---GCGC---TT---ATGGACATG---
>V_2      T  A              A              G      A      GGTGG      A
>V_3      A      A  C      A              T  G      A      T      G
```

**Note:** In each case the variants outputted follow the same format that will be described for the first example only.

## (ii) Variants from sequences matching a user search criteria

To obtain a list of unique variants from the aligned sequences whose titles match a user search criteria, select the “**Title Search**” -> “Variant Frequencies” menu option. This will open the popup window depicted in figure 2 that is almost identical to that in figure 1 but where there is an extra textbox field where the user can enter a search criteria such as a year, a location or a subtype.

**Figure 2:** Popup window for “**Title Search**” -> “Variant Frequencies” menu option where “Sequences where title contains” textbox is present.

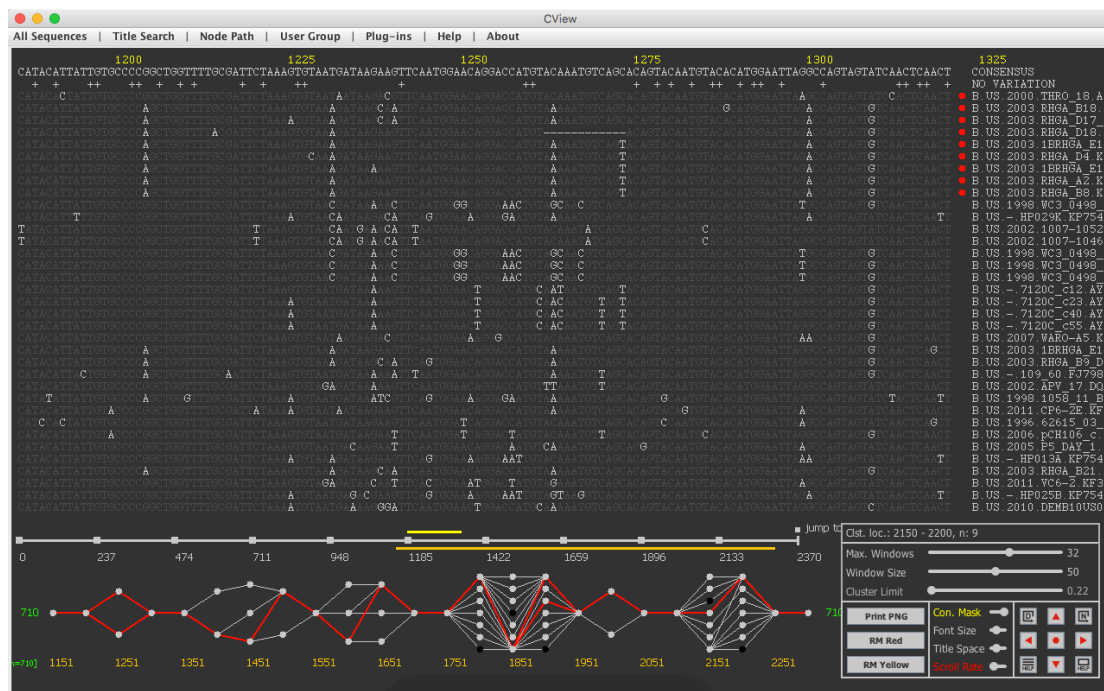
The other fields within this popup window are identical as before and when the “Proceed” button is pressed variants will once again be identified by CView, but this time only taking sequences into account that matched the search term specified by the user.

Given an alignment that contains sequences identified within different years, locations or some other factor, this option makes it easy to compare the most frequent variant without the need to produce separate alignments for each factor level. The latter would lead to complications such increased manual labour as well as sites within each individual alignment not being guaranteed to be compatible with each other.

### **(iii) Variants from sequences passing through a user selected node**

A third way of identifying variants within CView is to apply the method to sequences that pass through a user-selected node on the network. This is more exploratory than the previous two approaches, but depending on the specific interests of the user in question, this feature, in conjunction with the network parameter sliders, was included to allow for the rapid visual and intuitive exploration portions of diversity across the alignment.

To identify variants from sequences passing through a selected node, the user first must simply click on a node of the network that they wish to explore and this will automatically highlight the sequences passing through (figure 3).



**Figure 3:** A node of interest was clicked on using the mouse pointer to highlight the sequences passing through.

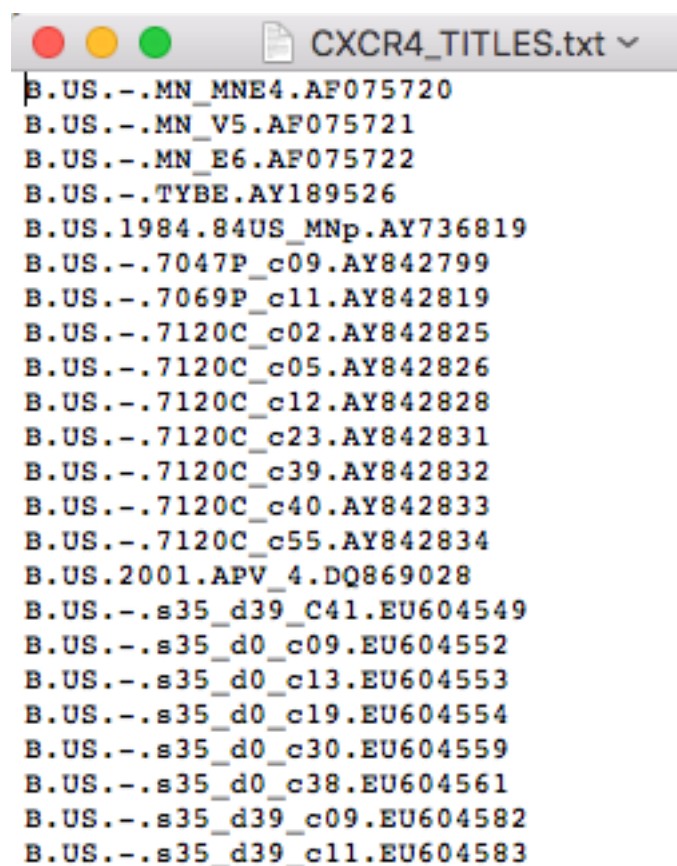
Following which the “**Node Path**” -> “**Variant Frequencies**” menu must be selected. This will result in a popup window identical to that of figure 1. When the “**Proceed**” button is pressed CView will identify variants in a similar manner as before but where only sequences passing through the selected node will be taken into account. These sequences will have a red circle placed next to their titles and will follow paths on the network that are highlighted in red.

It should be noted that when looking at variants in this manner the network itself can also be printed (using the “**print PNG**” button) and with third party image drawing tool, and some user manipulation of the variant formatting such as that at the end of section (i), an image of such variants can be produced that encompasses the background diversity network containing the highlighted paths.

#### (iv) Variants from a list of user specified sequences

This approach is employed using the “**User Group**” -> “**Variant Frequencies**” option and works in a similar way to the other three

described but with the exception that this time the popup window that appears allows the user to supply a list of titles to which the method will be applied to, i.e. variant frequencies will be calculated from sequences from within the alignment that match the list supplied by the user. The supplied title list should be in a simple text file of the format shown in figure 4.



```
B.US.--MN_MNE4.AF075720
B.US.--MN_V5.AF075721
B.US.--MN_E6.AF075722
B.US.--TYBE.AY189526
B.US.1984.84US_MNp.AY736819
B.US.--.7047P_c09.AY842799
B.US.--.7069P_c11.AY842819
B.US.--.7120C_c02.AY842825
B.US.--.7120C_c05.AY842826
B.US.--.7120C_c12.AY842828
B.US.--.7120C_c23.AY842831
B.US.--.7120C_c39.AY842832
B.US.--.7120C_c40.AY842833
B.US.--.7120C_c55.AY842834
B.US.2001.APV_4.DQ869028
B.US.--.s35_d39_C41.EU604549
B.US.--.s35_d0_c09.EU604552
B.US.--.s35_d0_c13.EU604553
B.US.--.s35_d0_c19.EU604554
B.US.--.s35_d0_c30.EU604559
B.US.--.s35_d0_c38.EU604561
B.US.--.s35_d39_c09.EU604582
B.US.--.s35_d39_c11.EU604583
```

**Figure 4:** Example of a list of titles that can be specified by the user for use with methods available under the “**User Group**” top level menu option.

**Note:** With the exception of methods designed for dissecting the alignment or extracting titles and those within the “**Plug In**” section, most implemented methods follow a similar type of user interaction as that described for the four top-level menu options used here (“**All Sequences**”, “**Title Search**”, “**Node Path**” and “**User Group**”), for example those involved in generating tables of residue frequencies, kmers hamming distances and clustering.