SeqPartitioner Geneious Plugin Manual

Jed Barlow
ejbarlow@ualberta.ca

April 12, 2014

Contents

1 Overview

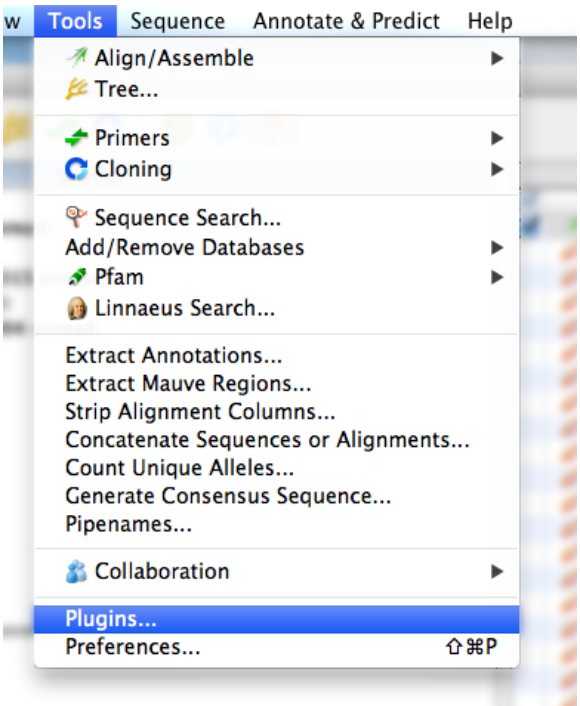
SeqPartitioner is a Geneious plugin for partitioning allele multisets. The plugin operates on alignments to produce various tables expressing the grouping of equivalent aligned nucleotide sequences and the number of differences between strains in terms of non-equivalent alleles.

2 Installation

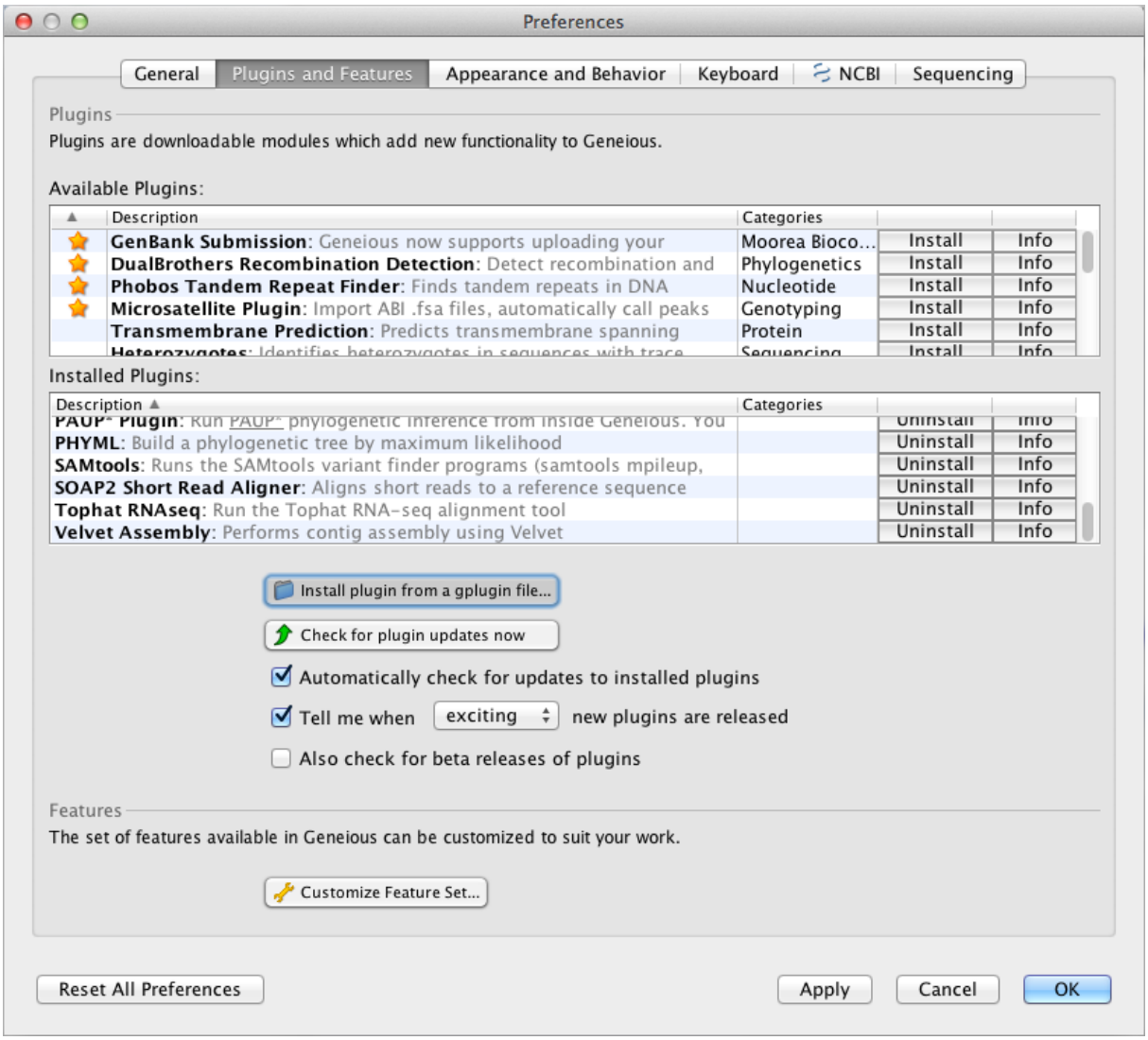
The plugin can be obtained, at the time of the writing of this manual, from two locations.

- In source form: `http://github.com/jedbarlow/biol398_seq_partitioner/`
- As a compiled package: `http://www.ualberta.ca/~ejbarlow/biol398/`

Once obtained, the plugin *SeqPartitioner.gplugin* can be installed by navigating the Geneious menu to `Tools->Plugins`.



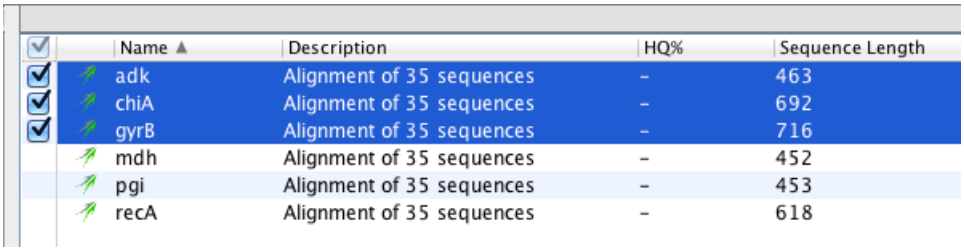
Then pressing the `Install plugin from a gplugin file...` button.



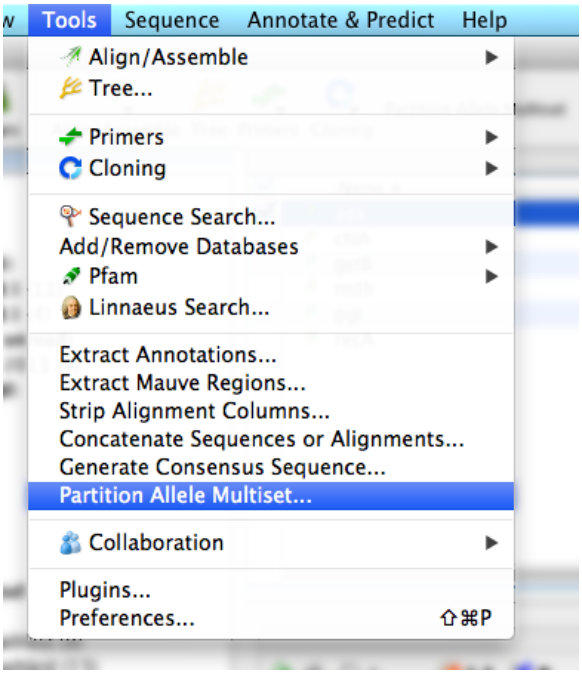
Finally, navigate to and select the plugin package file *SeqPartitioner.gplugin*. Once installed, an entry should appear in the Geneious menu as shown in the second figure below.

3 Operation

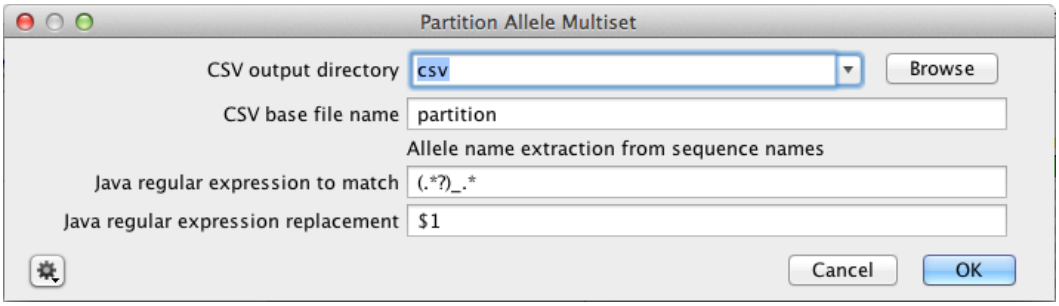
The plugin is used by first selecting a set of alignments.



Then navigating the menu to `Tools->Partition Allele Multiset`.



Then choosing options in the dialog box.



The meanings of the fields are as follows:

- *CSV output directory*
  This is the target directory into which the resulting table files will be written. Note that pre-existing files of the same name as the output files will be overwritten without confirmation.

- *CSV base file name*
  This is a string of text which will be appended with descriptive suffixes for each table. For example the following entry,

  ```
  partition1
  ```

  will result in the production of the following files,

  ```
  partition1_gene_vs_strain.csv
  partition1_strain_vs_strain.csv
  partition1_strain_vs_strain_matrix.csv
  ```

- *Java regular expression pattern match*
  The purpose of these last two fields is to establish a transformation of sequence names to the strain names, so that strains can be correlated between alignment files. The first field is a pattern to match against each sequence name.

- *Java regular expression replacement*
  This field describes a replacement for the text in the sequence name matched by the pattern in the previous field.

3.1 Example Regular Expressions

- If for example, the names of the sequences are patterned as `STRAIN_GENE`, then the default entry will work correctly by reducing sequence names to strain names, which will be consistent across alignment files and allow for correlation.

  Pattern: `(.*?)_.*`

  Replacement: `$1`

  The period matches any character, and the star matches a sequence of characters matching the previous construct (the period in this case). The question mark indicates that the star should only match until the next construct (an underscore in this case) can be matched. The parenthesis in the pattern designate the inner matching text as an identifiable string which can be referred to in the replacement as `$1`. It is important to note that text in the sequence name which does not match the pattern will be left unmodified, so the last part of the pattern `_.*`, which matches the `_GENE` portion of the name, cannot be omitted.

- Sequence names patterned such as `GENE_STRAIN` can be properly grouped by slightly modifying the pattern in the above.

  Pattern: `.*?_(.*)`

  Replacement: `$1`

More information about regular expressions, and in particular Java regular expressions, can be found at the following urls.

- `https://en.wikipedia.org/wiki/Regular_expression`
- `http://docs.oracle.com/javase/tutorial/essential/regex/intro.html`
- `http://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html`

4 Description of Output Tables

Two tables are created when the plugin is run. The tables with the suffixes `strain_vs_strain` and `strain_vs_strain_matrix` describe the number of different alleles (according to the equivalence relation) between any two strains.

The table with the suffix `gene_vs_strain` contains a column for each gene (alignment file), and a row for each strain. The number in a row indicates which group that allele matches with respect to other alleles of strains in the same column. The final column on the right indicates a partitioning of the rows of the other columns, so each matching row of numbers in the other columns will have the same value in the right most column.

4.1 Partition Relation and Errors

The equivalence relation used for partitioning is a comparison of aligned nucleotide sequences where '-' is always considered a match with the corresponding character in the other sequence. As a consequence, transitivity of the equivalence relation can break under certain conditions. So, for example, the following sequences cannot be partitioned:

(1) ATC

(2) A-C

(3) AGC

Since we have (1) = (2) = (3), but (1) $\neq$ (3), therefore no partition can be formed which both groups equal sequences and separates unequal sequences.

When such an error occurs, there will be entries of `-1` in the resulting table. Such values may not be the most effective indicators to use for identifying the source sequences of the error, because the situations of overlapping groups may be very complex and not comprehensively identified by the error markings. No table results should be accepted if the table contains a `-1`.