MacVector 17.5

for Mac OS X

Align to Reference: Sequence Confirmation Tutorial

MacVector, Inc.
Software for Scientists

Copyright statement

This version of the sequence confirmation tutorial was published in January 2020.

Contents

Introduction

MacVector has a unique sequence confirmation functionality that lets you align one or more chromatogram sample files against a reference sequence. It has many similarities to sequence assembly, except that it uses a known reference sequence as a scaffold.

You can use this functionality to help you solve a number of typical laboratory problems:

- Confirming the sequence of a cloned fragment
- Sequencing across the ends of a cloned fragment to confirm the junction sequence
- Screening clones from a site-specific mutagenesis experiment to identify successful mutations
- Screening related clones for single nucleotide polymorphisms
- Validating whole genome sequencing consensus sequences

The main limitation of this implementation is that you must have a reference sequence to act as a scaffold against which the sample sequences can be aligned. You therefore cannot use this to assemble trace files or fastq files from *de novo* sequencing projects. If you need full "contig assembly" capability, you should explore the MacVector Assembler module that is available as an add-on component.

Sample Files

After installing MacVector, you will find example files for this tutorial in the folder;

`/Applications/MacVector/Tutorial Files/Align To Reference/Sequence Confirmation/`

This contains an annotated MacVector format file called `SequenceSample` along with a `Trace Files` folder containing 32 chromatogram files in SCF format.

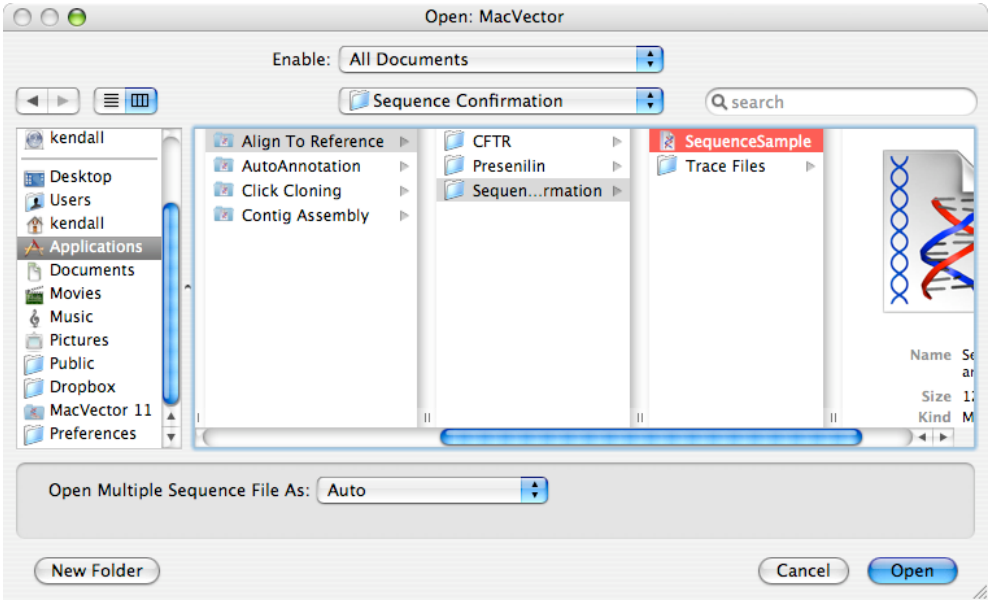
Tutorial

Opening SequenceSample

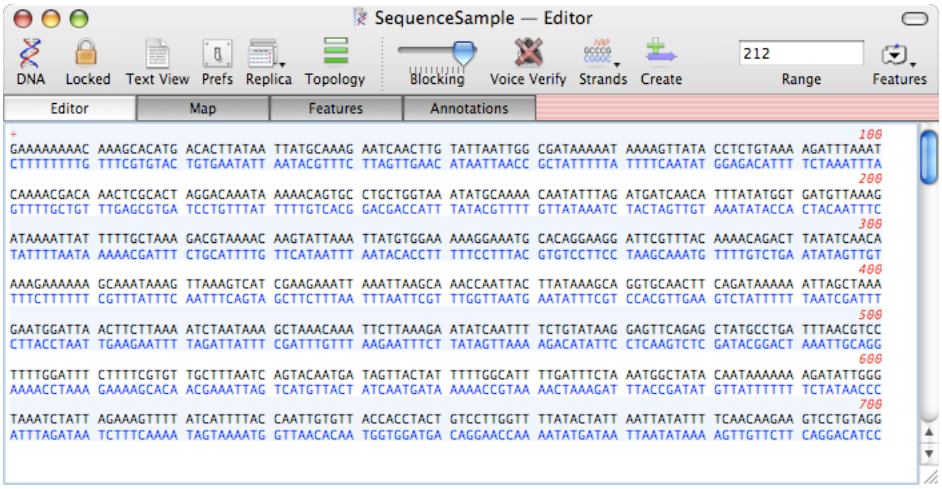
The first step in the tutorial is to open and explore the sample sequence to be used in the analysis

Select **File | Open** and navigate to the /`Tutorial Files/Align To Reference/Sequence Confirmation/` folder.

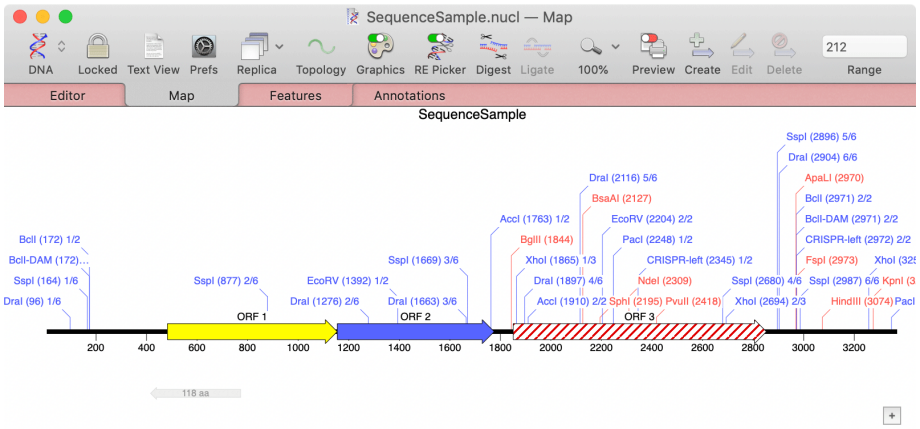
Select the sequence entitled `SequenceSample` and click on the **Open** button.



A standard MacVector nucleic acid sequence window will open;



Click on the **Map** tab to display the graphical view

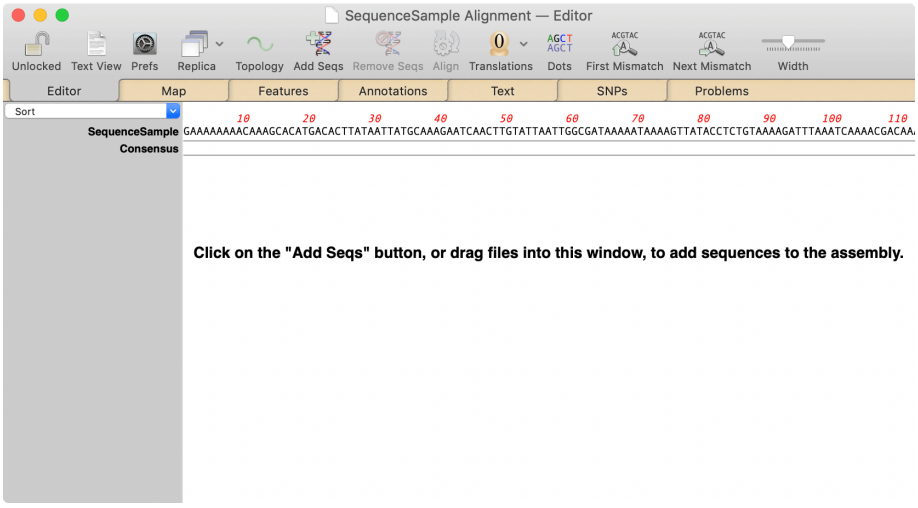


Note that the sequence has already been annotated with three open reading frames. The tutorial will demonstrate how this information is retained throughout the sequence confirmation analysis.

Opening the Align to Reference Window

Choose **Analyze | Align to Reference**.

The assembly window will open;



This window contains a ***copy*** of the original `SequenceSample`. The actual sequence residues are displayed along the top line of the window. You can click on the standard **Map**, **Features** and **Annotations** tabs to see exactly the same information that was present in the original sequence.

The assembly window is split into an upper and a lower pane. We will see that the lower pane is initially hidden but is used to display chromatogram information if one or more chromatogram files are added to the assembly. You can adjust the width of the title pane on the left-hand side by clicking

on and dragging the vertical separator bar. You can also hide or modify the toolbar buttons by `<ctrl>` – clicking on toolbar and selecting items in the dropdown menu.

Close the original `SequenceSample` window

The window closes, but the assembly window stays open. This reinforces the fact that the align to reference window contains a ***copy*** of the original sequence. This is different from many other MacVector analysis windows, but is necessary to simplify editing and saving in the window. Adding Sequences to the Assembly

Either (a) click on the **Add Seqs** toolbar button or (b) select **Edit | Add Sequences From File**.

In the file open dialog, navigate to the folder;
`Applications/MacVector/Tutorial Files/Align To Reference/Sequence Confirmation/TraceFiles/`

Click on the first file in the folder, scroll to the bottom of the list, then hold down the <**shift**> key and select the last file in the folder.

Finally, click on the **Open** button to import the entire list of files

The assembly window becomes populated with the imported trace files and the lower pane becomes visible. You can import trace files in ABI, SCF and ALF formats. In addition, you can import plain sequence files in any supported MacVector format, including large fasta and fastq files from NGS projects. The relative size of the upper and lower panes can be adjusted by clicking and dragging on the horizontal separator bar.

The imported sample sequences (also known as “Read Sequences” or simply “Reads”) appear in the upper pane, with their titles displayed as buttons in the left hand pane, along with an arrowhead indicating the direction of assembly. You can click on the title buttons to select the entire Read sequence.

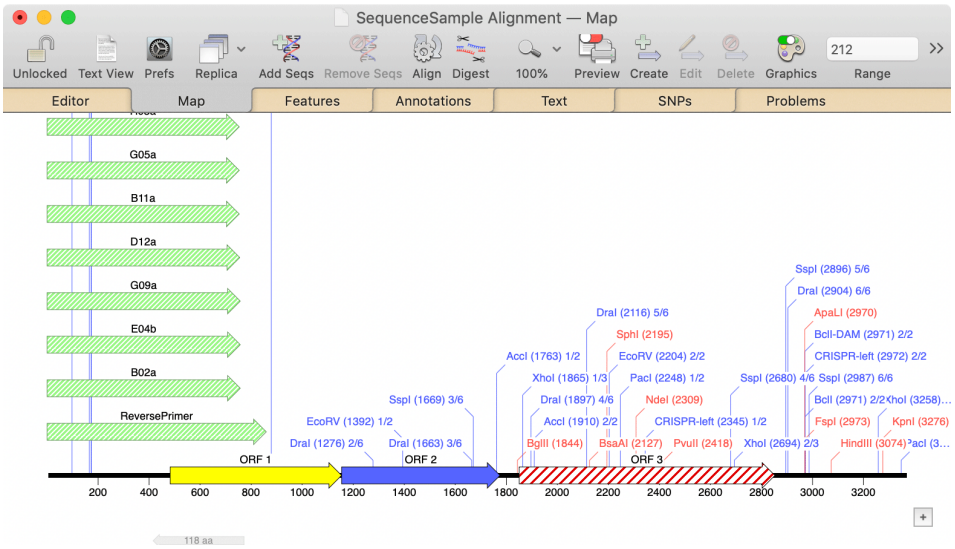
The imported Read sequences are initially shown in *italics* to indicate that they are unaligned. They are also inserted into the assembly at the left-most position.

At this stage the consensus sequence is blank because the consensus calculation only considers assembled sequences. We will see the consensus become updated later when we assemble the sequences.

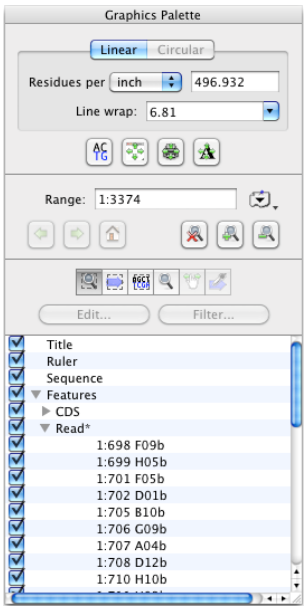
The lower pane shows the actual chromatogram trace displays of the imported sample sequences. If any MacVector plain sequence files were imported, these are not shown in the pane. We will look at these in more detail later.

Click on the **Map** tab.

The map view is displayed, showing each of the added sample sequences as a pale green arrow at the beginning of the sequence.



Each sample sequence is displayed as type “Read*”. You can show/hide these on an individual or group basis using the floating features palette window. If you don’t see this window, make it visible by choosing the menu option **Windows | Show Graphics Palette**



Move back to the **Editor** tab before continuing with the assembly

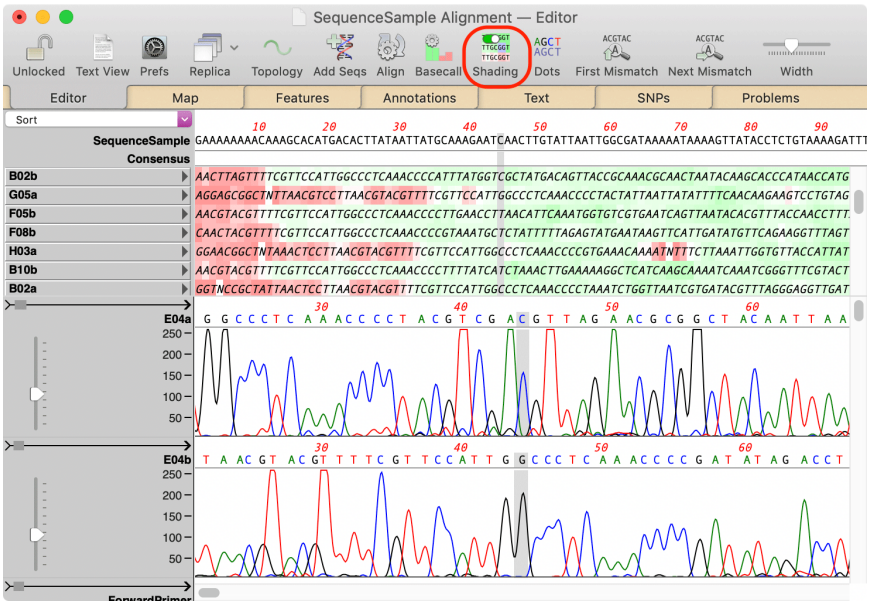
Base Calling with Phred

MacVector 17.5 added the ability to basecall chromatogram files using the popular phred algorithm. This not only re-interprates the peaks, often generating a more accurate sequence than the original basecall from the machine, but it also generates a "quality" value, estimating the likelihood that the basecall is in error. It uses a logarithmic scale from 0-99, where a value of 10 means that there is a 1 in 10 chance that the basecall is in error, a value of 20 is a 1 in 100 chance of error, 30 is a 1 in 1,000 chance of error etc. The scale maxes out at 95, with the values 98 and 99 reserved for edited residues.

Select one or more sequence by clicking on their names (hold down the `<shift>` key and scroll to select multiple) in the left hand pane an click on the **Basecalls** button.

Phred should complete within a few seconds, then the display will refresh.

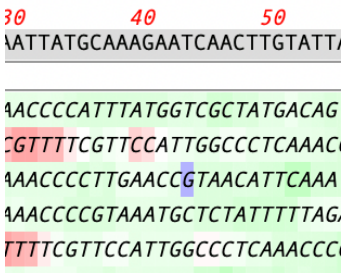
If not already selected, click on the **Shading** toolbar item to turn on colored background shading



You can see the residues now have colored backgrounds ranging from a dark red through white to a dark green. The scale runs from 1 (dark red) through increasing light red to 20 (white), then through increasingly darker shades of green to 40. All values above 40 are colored the same shade of green. This corresponds to the accepted "good" value for *phred* calls being 20 – a 1 in 100 chance of error. Gaps and residues with quality 0 are always shown with a white background.

Select any random residue and type "G"

The new overwritten residue is given a quality value of 99 and is shown with a blue background.

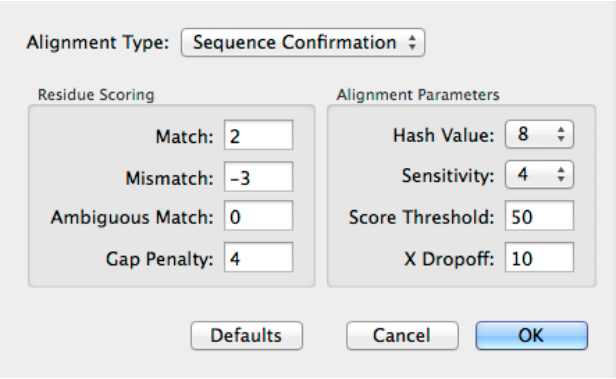


Choose **Edit | Undo** to revert the change before continuing.

Assembling Sequences

Click on the **Align** toolbar button

The *Assembly Parameters Dialog* is displayed. Make sure you have **Alignment Type:** set to **Sequence Confirmation**.



The **cDNA Alignment** algorithm is covered in a different tutorial. For most assemblies where the sample files are of good quality and are closely related to the reference, the default parameters are usually ideal. If you want to try other combinations of parameters, you can always revert to the recommended parameters by clicking on the **Defaults** button.

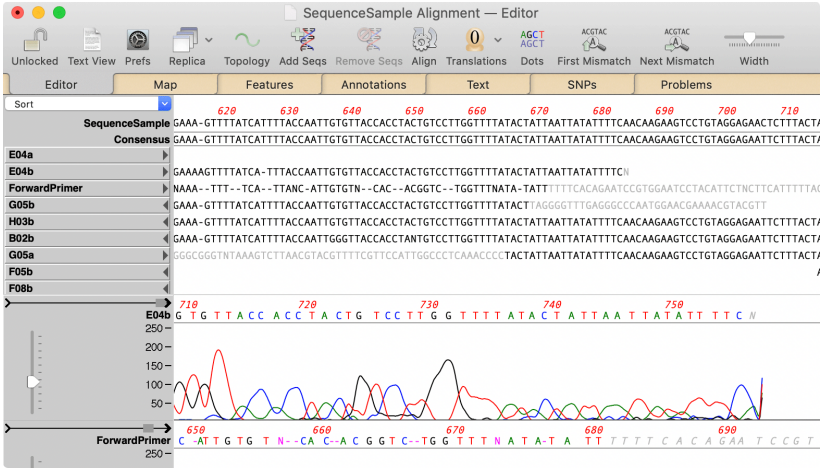
Click on the **OK** button to initiate the assembly calculation.

Once complete, the sequence confirmation window refreshes to display the aligned sequences. Any sequences that were not assembled remain italicized and positioned at the beginning of the assembly.

During the assembly, gaps become inserted into the reference sequence. However, these are for display purposes only so that the assembled sample sequences can align appropriately – behind the scenes, the reference sequence is unchanged after the assembly. The only way the reference sequence will change is if you make edits to the sequence directly.

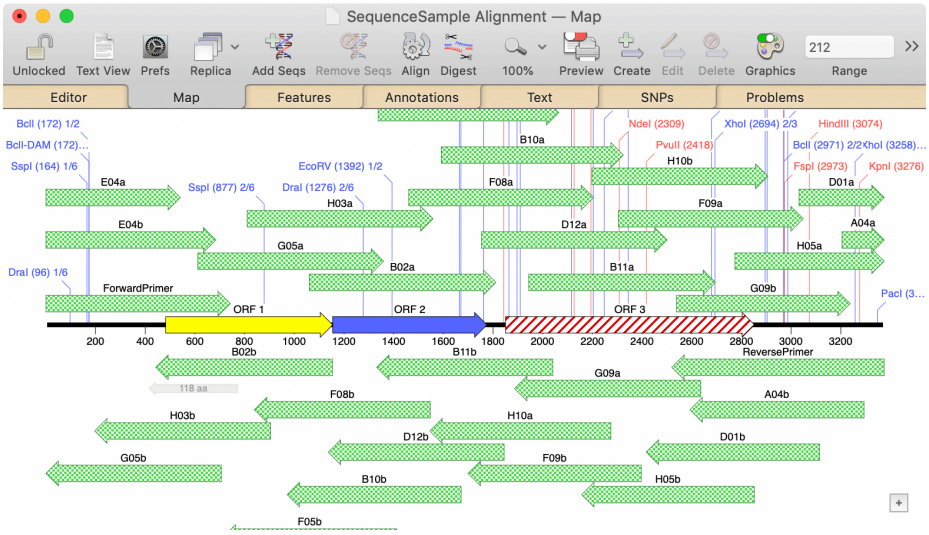
Note that the consensus line also has quality coloring. In this case, in accordance with the *phred* standard for assembled sequences, the scale runs from 1 (red) through 40 (white) to 80 (green) i.e. twice the values for single sequences.

The assembly algorithm understands that the sample sequences may have vector sequence at the ends, or poor quality sequence that interferes with the assembly. To account for this, it will terminate the alignment extension when it encounters poor quality matches. The display indicates those residues that are "masked" (i.e. not included in the assembly) by drawing them in gray rather than black.



Click once more on the **Map** tab.

The graphic tab is redisplayed, but this time the sample sequences are shown in their new location on the assembly, and in addition they are shown in a darker green color. This is because assembled sequences are given the type “Read” rather than the “Read*” type assigned to unassembled sequences.



Navigating in the Assembly Window

The sequence confirmation window provides a number of different functions for exploring and selecting sequences and residues in the assembly. You can use these tools to quickly locate differences between the assembled sample files and the reference sequence.

Move back to the **Editor** tab. Click on a base in the reference sequence.

The base highlights as expected, but, in addition, the equivalent bases in the consensus sequence and aligned sample sequences highlight in gray. The same bases are highlighted in the lower multiple trace pane, and the chromatograms scroll so that the highlighted bases are all aligned at the center of the screen.

Click on a base in the consensus sequence then drag-select in the consensus to highlight multiple residues.

This has a very similar effect to clicking on the reference sequence except that now the consensus sequence gets the primary highlight and the reference sequence shows no selection. However, the aligned sample sequences become highlighted and the chromatograms scroll just as with the reference sequence.

You can use this feature at any time to center the chromatogram display to any location in the reference or consensus sequence. When drag-selecting, all of the corresponding residues in the assembled sample sequences

become highlighted and the chromatograms scroll to the center residue of the selection.

Use the horizontal scroll bar to scroll through the sequence.

Use the upper vertical scroll bar to scroll down through the assembled sample sequences to keep the appropriate sequences in view as you scroll horizontally through the assembly.

Each assembled sample sequence gets its own line in the upper pane, so you need to scroll vertically to keep them in view. You can also use the vertical splitter control to view more of the sequences in the upper pane.

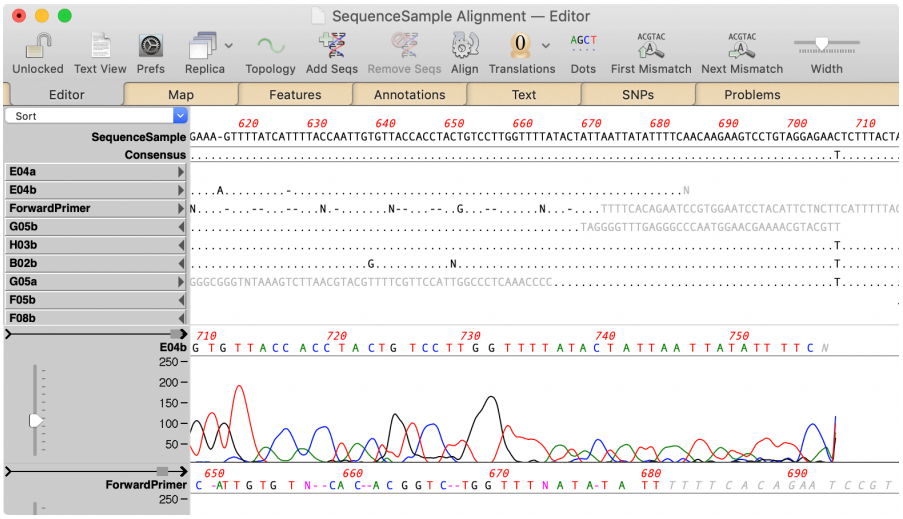
As you scroll through the assembly, the chromatograms in the lower pane scroll appropriately. They key in on the center base in the visible portion of the reference sequence, although the selection does not change and will scroll off the screen as you navigate through the assembly.

Click on a base in the one of the assembled sample sequences.

This selects the residue in the sample sequence and also in the chromatogram. However, the chromatograms do not re-center – this lets you edit the sample sequences without the chromatograms constantly moving around.

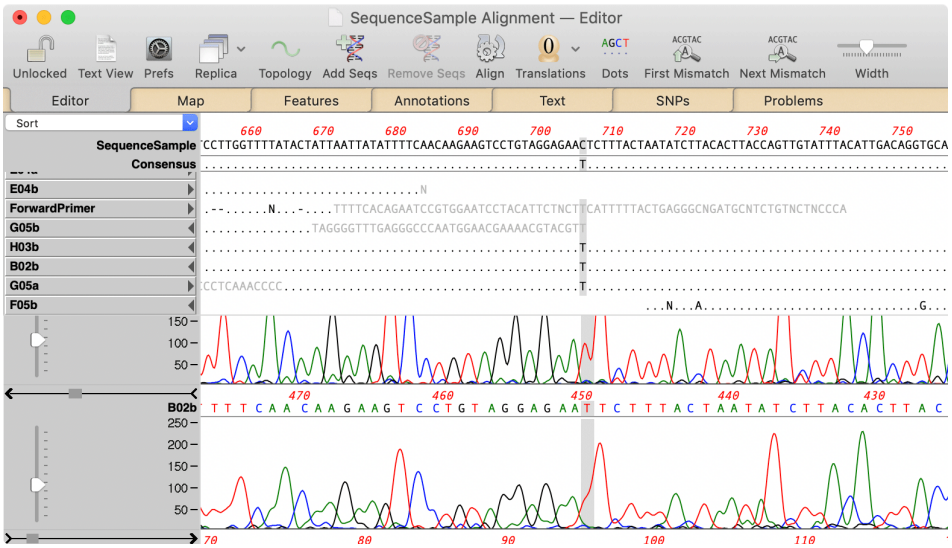
Identifying Mismatches

Click on the **Dots** toolbar icon.



The upper pane redraws so that any residues in the consensus or assembled sample sequences that match the reference are shown as dots.

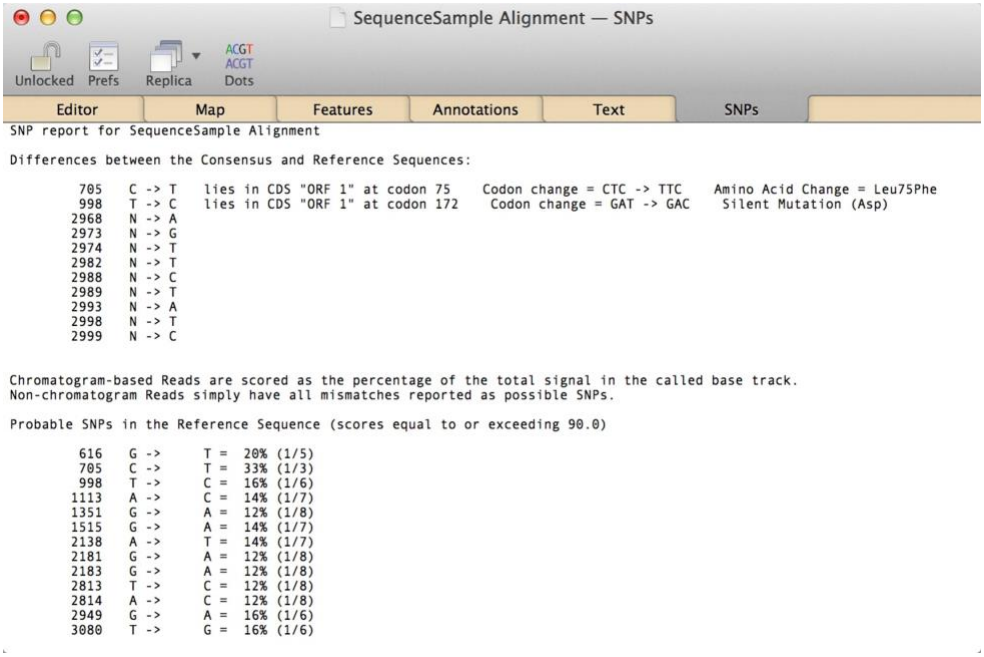
Click on the **First Mismatch** toolbar icon. If the icon is not visible in the toolbar, either increase the size of the window, or click on the **>>** icon at the right edge to display the additional icons.



This searches for the first mismatch between the consensus sequence and the reference sequence. The display refreshes as if you had clicked on that residue in the reference sequence i.e. the residue becomes selected and the chromatograms scroll to that position. You can click on the **Next Mismatch** toolbar icon to move to the next mismatch in the assembly.

Identifying Single Nucleotide Polymorphisms

Click on the SNP tab.



This tab lists the possible Single Nucleotide Polymorphisms in the aligned sequences. It is split into several sections;

SNPs in the Consensus

```
Differences between the Consensus and Reference Sequences:

         705   C -> T   lies in CDS "ORF 1" at codon 75    Codon change = CTC -> TTC    Amino Acid Change = Leu75Phe
         998   T -> C   lies in CDS "ORF 1" at codon 172    Codon change = GAT -> GAC    Silent Mutation (Asp)
        2968   N -> A
        2973   N -> G
        2974   N -> T
        2982   N -> T
        2988   N -> C
        2989   N -> T
        2993   N -> A
        2998   N -> T
        2999   N -> C
```

This section lists any differences between the consensus and the reference sequence. As well as list the actual base change, it also indicates if the change is within a coding sequence and, if it is, what change (if any) the change would have on the encoded protein.

Common SNPs in the Reads

```
Probable SNPs in the Reference Sequence (scores equal to or exceeding 90.0)

         616   G ->     T =  20% (1/5)
         705   C ->     T =  33% (1/3)
         998   T ->     C =  16% (1/6)
        1113   A ->     C =  14% (1/7)
        1351   G ->     A =  12% (1/8)
        1515   G ->     A =  14% (1/7)
        2138   A ->     T =  14% (1/7)
        2181   G ->     A =  12% (1/8)
        2183   G ->     A =  12% (1/8)
        2813   T ->     C =  12% (1/8)
        2814   A ->     C =  12% (1/8)
        2949   G ->     A =  16% (1/6)
        3080   T ->     G =  16% (1/6)
```

MacVector calculates the area under the curve for each trace. If over 90% of the signal is due to a residue that is different from the reference sequence, this is considered a “probable” SNP. If it exceeds 75% of the signal, this is considered a “possible” SNP. Each position that contains a probable or possible SNP is listed, along with the number of reads that contains that SNP.

SNPs in Individual Reads

```
ForwardPrimer     length 784  1 probable SNPs, 2 possible SNPs

     Probable SNPs
         616   G -> T   (94.2)

     Possible SNPs
         456   A -> C   (78.5)
         554   T -> G   (88.3)

G05b     length 731  0 probable SNPs, 2 possible SNPs

     Possible SNPs
           9   A -> C   (77.5)
          13   A -> G   (81.3)

H03b     length 733  1 probable SNPs, 0 possible SNPs

     Probable SNPs
         705   C -> T   (90.3)
```

Finally, the probable and possible SNPs for each Read are listed. The number in parentheses e.g. (94.2) is percentage of total trace signal accounted for by the SNP residue.

Editing Assemblies

You can edit the reference sequence directly, or any of the assembled sample sequences. You CANNOT edit the consensus sequence directly.

This is always calculated dynamically from the overlapping sample sequences.

Click on the **First Mismatch** toolbar icon to reset the assembly to the first consensus/reference mismatch. This should be around residue 705.

Examine the chromatograms – clearly the reference sequence is incorrect: it has a "C" where most of the sample sequences indicate the residue should be a "T". Type a "T".

The residue changes as you would expect. If you have **Show Dots** set, the consensus and sample sequence residues will change to dots to indicate that they now match the reference.

Click on the **Next Mismatch** toolbar icon to locate the next consensus/reference mismatch. This should occur at residue 998.

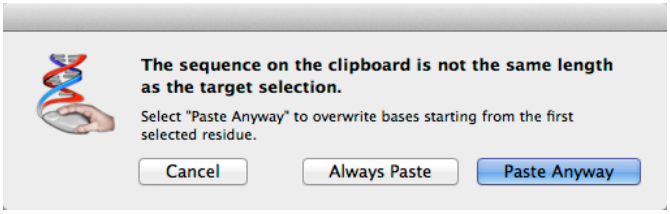
Examine the chromatograms – again the reference sequence is incorrect: in this case it has a "T" instead of a "C".

Edit the residue, then click on **Next Mismatch** once more. This time there are a number of mismatches in the region immediately after the highlighted residue.

Click on the first mismatched residue in the consensus sequence. Drag-select across the entire region of mismatched bases – about 40 bases or so.

Choose **Edit | Copy**, then click on the first mismatched residue in the reference sequence. Choose **Edit | Paste.**

A dialog appears alerting you to the fact that you are attempting to overwrite a selection with a different length sequence from the pasteboard;



Click on **Paste Anyway** to paste the selected residues over the reference sequence.

The reference sequence is overwritten with the copied residues. You should use this feature with care; it is designed to simplify copying the consensus sequence to the reference, but if you make a mistake, the reference can become badly corrupted. However, you can always choose **Edit | Undo** to revert the last editing action.

Choose **Edit | Undo**, then repeat the last copy/paste operation with the "Show Dots" mode enabled.

Note how, although the consensus sequence contains dots where it matches the reference, when you copy the consensus, the original residues are copied to the clipboard, not the dots. You can confirm this by pasting into an external text editing application if you wish.

You can also paste the residues copied from the consensus line into a new MacVector nucleic acid sequence window. Note once again that when you copy and paste the consensus, any gaps are carried along with the sequence. When you copy and paste the reference sequence, gaps are removed.

Saving Assemblies

Choose **File | Save**. A file save dialog appears. Choose a suitable location for the file and click on the **Save** button.

This saves the assembly in MacVector's assembly format. This is an XML format based on the BSML standard. You can view the contents of the file using a standard text editor (e.g. TextEdit) if you want to see how this is formatted.

Choose **File | Export**. This time, choose **MacVector NA Sequence File** from the Format popup menu before choosing a file name and location.

This saves a copy of the reference sequence, complete with all features and any edits you have made.

Open the file you have just created. Hint: the easiest way to do this is to select the file name from the **File | Recent Files** menu.

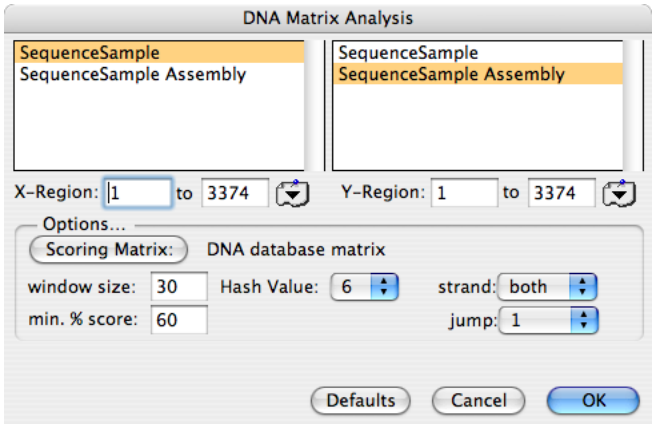
The file opens as a standard MacVector nucleic acid sequence file. If you look at the feature list, you will see that the locations of the assembled chromatogram files have been saved in the file as "Read" features. This gives you a record of the traces you used in the sequence confirmation assembly.

Analyzing the Reference Sequence

You can perform any MacVector nucleic acid sequence analysis directly on the sequence confirmation assembly. There is no need to save the file in MacVector format before running an analysis.

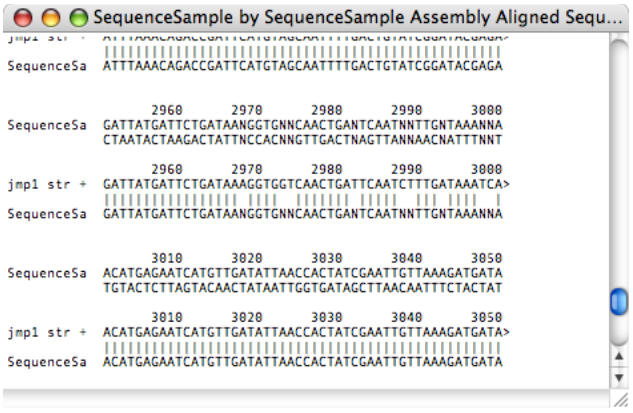
If you do not still have it open, open the original `SequenceSample` nucleic acid file you started with.

Choose **Analyze | Create Dot Plot | Pustell DNA Matrix**. Make sure `SequenceSample` is selected in the left pane and `SequenceSample Assembly` in the right pane.



Click on **OK**. When the calculation is complete, make sure the **Aligned Sequence** option is selected in the Result filter dialog and click on the **OK** button.

As you scroll through the **Aligned Sequence** tab in the results window, you should be able to spot the mismatches in the regions where you made the edits in the reference sequence.



Now try an **Analyze | Restriction Enzyme…** analysis on the assembled sequence and compare the results to an analysis on the original `SampleSequence`. Make sure you are using a restriction enzyme file such as `Common Enzymes` or `New England Biolabs` and that you are using either **All Enzymes** or **Selected Enzymes** where you have selected *Eco*RI. Select *Show Restriction Map* in the display **Options** dialog. You should see that the edit that you made at residue 705 created an *Eco*RI restriction site (at position 702) that is not present in the original sequence.

Reference Manual

Creating and Populating a New Assembly

The **Align to Reference** menu item in the **Analyze** menu is active whenever a single nucleic acid sequence window is active. When chosen, an "empty" Assembly window is opened, titled "<sequence name> Assembly" by default. The "empty" window actually contains the reference sequence displayed along the top of the alignment pane.

You can then add sample sequences to the window by either (a) selecting the **Add Seqs** toolbar button or (b) selecting **Edit | Add Sequences from File**. Either option will bring up a standard multiple file add dialog.

You can select one or more files to be added to the alignment. These can be any type of sequence data that MacVector can recognize, including NGS data in fasta or fastq format, even if they are gzipped. Note that at this stage, there is no filtering for suitable files. The **Enable** menu contains the standard MacVector filters, but most users should select **All Documents** as trace files sometimes do not have a standard Macintosh file type and thus can only be identified when MacVector attempts to actually load the file.

MacVector will attempt to import all of the selected files into the assembly project. Any trace or nucleotide sequence that can be opened by MacVector should be capable of import into the window. An error message should be displayed listing the names of any files that could not be imported. The assembly window updates to display the sequences in the order in which they were imported. The sequence of each sample is displayed in the upper Alignment Pane and the chromatogram traces (if present) are displayed in the lower Multiple Trace Pane. Note: Unaligned sequences are displayed in italics starting at residue "0" to alert you to the fact that they have yet to be aligned (or could not be aligned after automated assembly). Italicized sequences are NOT included in consensus calculations.

Automated Assembly

Interface

The user can click on the **Align** button to bring up the alignment dialog.

Parameters

The parameters have the following impact on the assembly algorithm;

Match

(Valid range -100 to 100, default 2). This is the value the algorithm assigns to a match between a sample residue and the reference residue. It should typically be a positive value.

Mismatch

(Valid range -100 to 100, default -3). This is the value the algorithm assigns to a mismatch between a sample residue and the reference residue. It should typically be a negative value so that it reduces the cumulative match value as the algorithm extends the match between the sequences.

Ambiguous Match

(Valid range -100 to 100, default 0). This is the value the algorithm assigns to an ambiguous match between the sample and reference residues. The default is a neutral value – you can increase this to the match value if you want ambiguous matches to be treated exactly the same as matches. Alternatively, assigning it the same value as the mismatch parameter treats ambiguous matches as full mismatches.

Gap Penalty

(Valid range -100 to 100, default 4). This is the value the algorithm subtracts from the cumulative match score whenever it has to insert a gap character. Unlike some other algorithms, the SNP Assembly algorithm does not distinguish between gap insertions and gap extensions – all gaps are treated as a gap insertion.

Hash Value

(Valid range 1 to 10, default 8). This is the number of bases MacVector uses for the hashing algorithm. A value of 4 (the default) means that MacVector initially only searches for perfect 4 base matches between the reference and sample sequences. Larger values lead to faster searches, but with reduced sensitivity.

Sensitivity

(Valid range 1 to 10, default 4). This value affects what MacVector does when it is extending a match and encounters a mismatch. The value determines how far ahead the algorithm should look to determine whether to insert a gap in either of the sequences or to accept the mismatch. A value of 4 means it looks ahead in all directions until 4 additional mismatches (or the end of the sequence) has been encountered. The larger the value, the more likely the algorithm can handle longer regions of poor quality sequence, but performance will be much slower (the algorithm takes approximately three times longer to run for each unit increment in this value). The default (or even a smaller value) is quite adequate for most sequencing projects. Only increase this if you have long regions of poor quality sequence that is being poorly aligned using the default

parameter.

Score Threshold

(Valid range 1 to 1000, default 50). This value controls how MacVector determines that a match is significant. After finding an initial match, MacVector attempts to extend the match in each direction using the match/mismatch and gap penalty scoring parameters. It retains the extended segment that gives the highest score. If the best score exceeds the score threshold, then MacVector considers this to be a significant match and includes the sample sequence in the alignment. If no individual match segment exceeds this score, the sample is treated as "unaligned" and will appear in the alignment view in italics. If you dramatically change the values for match/mismatch, you should alter this to keep it approximately in sync.

X Dropoff

(Valid range 1 to 1000, default 10). This value is used by MacVector to tell it when to give up extending a match. When extending, it keeps track of the best possible match score. It continues to extend the match in each direction, only giving up when the cumulative score falls to less than "X Dropoff" from the best score. If you were to reduce this to a low value (e.g. "0") only perfect matches would be generated in the final alignment. You should typically set this to a value that will permit a few mismatches to be incorporated into the alignment to allow for sequencing errors. A smaller value will speed up calculations, but may not correctly align longer regions of poor quality matches.

Algorithm

The assembly algorithm is essentially a pairwise alignment algorithm, tuned for small insertions/deletions. Each sample sequence is independently aligned against the reference sequence as follows;

- A hash table is first created from the reference sequence using the hash value parameter set by the user. This is basically a "shortcut" table that points to the first occurrence of a short sequence in the reference. The default value of 4 sets up a table of the location of all 4 nucleotide sequences in the reference.

- The sample sequence is then scanned for matches to the reference as follows - the first "hash" residues of the top strand are used to lookup the locations of matches in the reference. If matches are present, each is extended in the forward and reverse direction until the first mismatch is encountered. The ungapped perfect alignment is scored using the "match" parameter value (e.g. if the match was 10 residues and the "match" parameter was 4, the score would be 40). This is repeated for all residues in the sample sequence and

the top ~20 matches are saved. The algorithm automatically discards matches that are subsets of already calculated matches to speed up calculations.

- After the initial scan is complete, the best scoring "perfect" match is then extended in both the forwards and backwards directions to find the optimum gapped match;

  - When a mismatch is encountered, the algorithm enters a "recursive" comparison to determine the best course of action – (a) accept the mismatch (or ambiguous match) penalty and continue to the next pair of residues, (b) accept the gap penalty and continue after inserting a gap into the reference sequence or (c) accept the gap penalty and continue after inserting a gap into the sample sequence.
  - The recursive algorithm works by extending the match until "sensitivity" number of mismatches/gaps have been encountered in each direction. If sensitivity is 1, then it simply looks ahead until the next mismatch after trying (1) no insertion, (2) insertion in the reference, and (3) insertion in the sample. It calculates the score for each "path" by adding together the match scores along with the mismatch score for the mismatch. This results in a total of 3 extended comparisons. With a sensitivity of 2, the algorithm iterates to a depth of "2", for a total of 9 alignments and saved scores. A sensitivity of 3 leads to 27 alignments etc.
  - The algorithm remembers the best score it found, and the locations on the sequences that gave that score, as it extends in each direction.
  - The extension is continued in each direction with the "current" score being adjusted appropriately with mismatch/ambiguous match/gap penalty variables. The extension is terminated when the "current" score falls more then "x-dropoff" below the best score. This parameter stops the algorithm from wasting time on the wrong "path". After the recursion level ("sensitivity"), it is probably the most important parameter that affects computation time.
  - Once all extensions and matches have been exhausted, the alignment search is repeated on the opposite strand.
  - The best score is kept as the optimal alignment for that sequence, with both orientations considered. If the optimal score falls below the "Score Threshold", the sample sequence is considered to not match the reference.
  - If the comparison is considered to be a match, the algorithm then adds any gaps that needed to be inserted into the master copy of the reference sequence and assigns appropriate offsets (and

gaps) to the sample sequence.

Editing the Assembly

Colors and Fonts

MacVector shows all residues in the currently selected editor font. Certain font variations are used to indicate status;

- Italics – unassembled sequences are shown in italics in the alignment pane. They are always positioned at "0" in the alignment. All sequences that are added to an alignment are initially unassembled and will be shown in italics. After assembly, only those sequences that do not have significant alignments with the reference sequence are shown in italics. NOTE: when italics are used, they should always apply to the entire sequence. There is no case where just a subset of a sequence would be in italics.

- Gray Text – This is used to indicate the regions where a sequence could not be aligned with the reference. This is currently determined only by the automated assembly algorithm. There is no way the user can directly change this. Grayed residues are NOT considered in consensus calculations.

Navigation

Scroll Bars

There are three scroll bars in the assembly window. Each will respond interactively to drags on the "thumb" and will scroll the appropriate display incrementally when clicked on the arrows or in the "page scroll" region.

- Alignment Vertical – this scrollbar lets users scroll vertically through the sample sequences that have been added to an assembly. The reference and consensus sequences always stay at the top of the alignment pane. Because each sample sequence has its own line in the display, you may see a lot of "white space" while scrolling through a large assembly.

- Multiple Trace Vertical – this scrollbar lets users scroll vertically through the trace sample sequences that overlap the current Alignment view. This is typically only a subset of the available sample sequences.

- Horizontal – this scrollbar lets users scroll horizontally through the entire alignment.

Scaling Controls

These controls affect the way the trace panes are displayed. There are two types;

- Width Control – this is a single control on the toolbar. When dragged, it interactively adjusts the horizontal scaling factor of the trace displays. The default value displays the traces at a scaling factor of one pixel per sample point.

- Vertical Gain – each individual trace pane has a vertical slider in the left hand pane that controls the vertical scaling of the traces. The default is that the tallest peak in the trace is "100%". When dragged, it interactively adjust the vertical scaling factor of the trace so that users can "zoom in" to more closely examine areas of ambiguity.

Mouse Selections

Much of the interaction with the window is driven by mouse clicks.

Alignment Pane

- Reference Sequence – if the user clicks on the reference sequence, the selected residue is highlighted. In addition, the multiple trace view will scroll so that all sample sequences that overlap the selected residue are centered in the multiple trace pane. In addition, each trace will have the corresponding residue selected. The user can drag-select in the reference sequence to highlight multiple residues – in this case, the corresponding residues will be selected in all overlapping traces.

- Consensus Sequence – clicking in this region is similar to clicking in the reference sequence except that the consensus cannot be edited. The consensus sequence is an algorithmically generated sequence that the user cannot directly interact with.

- Sample Sequence – clicking on a sample sequence deselects any other selections and selects the chosen residue in that sequence. If that sequence and residue is visible in the multiple trace pane, it will be highlighted. However, selecting a sample sequence does not cause any scrolling of the display. The user can drag-select in the reference sequence to highlight multiple residues – in this case, the corresponding residues will be selected in the multiple trace pane.

Multiple Trace Pane

In the current implementation, the only region that responds to mouse clicks is the actual sequence residue line. Clicking elsewhere may deselect any current selection but will not have any other effect.

- Sample Sequence – clicking in the editable region deselects any other selection and selects the chosen residue in that sequence. If that sequence and residue is visible in the alignment pane, it will be highlighted. However, selecting a sample sequence does not cause any scrolling of the display. The user can drag-select in the reference sequence to highlight multiple residues – in this case, the corresponding residues will be selected in the alignment pane.

Splitter Bars

The assembly window has two splitter bars that control the size and location of the panes; Vertical Splitter - This splitter lets users adjust the relative size of the upper (alignment) pane and the lower (multiple trace) pane. Horizontal Splitter - This splitter lets users adjust the relative size of the left hand “title” pane.

Editing

IUPAC Keys

Edits can be made in;

- The reference sequence
- Sample sequences in the alignment pane
- Sample sequences in the multiple trace pane

There are a number of restrictions on editing;

- Only standard IUPC DNA characters can be used.
- Editing is only available when a single residue is selected.
- Editing usually replaces a selected character. When an edit is made, the following actions occur;

  - The alignment is made “dirty”, meaning that **Save** will be enabled and the user will be prompted to save the alignment if the main window is closed.

  - If the edit is in a sample sequence in the alignment pane, the corresponding sequence in the multiple trace pane is updated and the overall consensus sequence is recalculated.

  - If the edit is in a sample sequence in the multiple trace pane, the corresponding sequence in the alignment pane is updated and the overall consensus sequence is

recalculated.

- If the edit is in the reference sequence, the location of any features is updated if the edit turns a gap into a residue, or if a residue is deleted.

• To insert a residue, hold down the <option> key while typing a base.

Space Key

This can only be used if a single residue selection is made in a reference or sample sequence. In this case a gap character will replace the residue. If the change means that all aligned Reads and the reference sequence contains a gap at that position, the gap will be deleted from all sequences and will "close up"

Delete or Backspace Key

This deletes the residue under the cursor. In addition, if an entire sample sequence is selected (e.g. by clicking on the title button) that sequence will be removed from the alignment.

Option Key

Hold down the `<option>` key when typing a character or a gap and the typed residue will be inserted immediately before the selected base.

Miscellaneous

ShowAsDots Mode

The user can toggle the **Show Dots** button to switch between two different display modes. In normal mode, all sequences are displayed as expected. When **Show Dots** is enabled, residues in the consensus and sample sequences that exactly match the reference sequence are shown as dots. This includes gap characters, but not unaligned sample sequences, spaces or "masked" regions where a sample sequence is not considered to be aligned to the reference. The user can edit "dots" as with any normal residue – if they type a residue that matches the reference, that residue will be displayed as a dot.

Cut/Copy/Paste

Cut is always disabled. **Copy** is active only when a selection is present and acts on the currently selected sequence as follows;

• If the reference sequence is highlighted, the selected residues in the reference are copied to the clipboard as a nucleic acid sequence. Any overlapping features are NOT copied.

• If the consensus sequence is highlighted, the selected residues in the

consensus are copied to the clipboard as a nucleic acid sequence. Any overlapping features are NOT copied.

- If a sample sequence is highlighted (and the reference is not), the selection in that sample sequence is copied to the clipboard as a nucleic acid sequence.

Paste is enabled only when the reference sequence is selected. When you paste into the reference, the residues on the clipboard will *overwrite* the reference sequence, NOT *insert*. You can use this feature to copy and paste the consensus sequence into the reference.

Undo

Most operations in the assembly window can be undone, as with other MacVector windows. In particular, you can undo an assembly operation, and also adding/deleting sequences.

Export

The **Text** tab displays a text representation of the alignment that updates in real time when the alignment is edited in the **Editor** tab. With this tab selected, you can save the alignment as a text file by choosing **File | Export**.

Using the Find Options

Standard Find

The user can select the **Edit | Find** menu to bring up the standard Find dialog. This lets you find matching residues in the reference sequence. Note that gaps are ignored during the search.

Mismatch Find

The dialog lets the user find mismatches between the reference sequence and the consensus. It is invoked by clicking on the Find button on the toolbar (the image with a pair of binoculars and a "dotted" mismatched base). When the "Find" button is selected, the display changes so that the first mismatch between the reference and consensus is selected and displayed. When “Find Next” is chosen, the display centers on the next mismatch between reference and consensus. If you have multiple assembly windows open, the find dialog is always associated with the one at the front. When a different type of window is selected, the dialog is hidden to help reduce screen clutter.

Saving/Loading Alignments

File Format Details

Alignment files have a file type of 'AXML' with the standard MacVector creator type ('MVTR'). They are actually BSML (Biological Sequence Markup Language) XML format files, but are given the file extension ".axml". The file contents can be viewed and edited with any standard text editor. Because MacVector uses the Macintosh operating system to actually read in the file and parse the contents, you should even be able to use TextEdit to save modified files in rich text format (the TextEdit default)(this doesn't work for other MacVector text formats!). The contents of the file largely follow the standard BSML format with DS Gene variations as appropriate. There have been a few changes to support sequence assembly;

- The "model-type" is "assembly" rather than "sequence"

- Each sequence in the alignment is stored in standard BSML/DS Gene format – however, the "ID" field of each sequence has specific meaning. The edited reference sequence has ID "REFERENCE_SEQ", the consensus is "CONSENSUS_SEQ" and each component is labeled "SEQ_0", "SEQ_1" etc. A copy of the original sequence, along with all features, annotations and feature appearance information is stored in "ORIGINAL_SEQ".

- Each component sequence has an "interval-loc" entry that defines its location and strand on the alignment. This is a standard BSML model, but is usually associated with features rather than sequences.

- Each component sequence has additional attributes;

o "IsAssembled" – "0" for no, "1" for yes o "ClipLeft" – the extent of the masking at the 5' end of the sequence.
o "ClipRight" – the start of the masking at the 3' end of the sequence.

- There are additional attributes stored at the "definitions" level describing the parameters used in the assembly. Most correspond directly to parameters in the assembly dialog;

o "GapPenalty"
o "MismatchScore"
o "MatchScore"
o "AmbiguousScore"
o "HashValue"

o “ScoreThreshold"
o “XDropOff"
o “Recursion" – this is shown in the dialog as “sensitivity”
o “Threshold"

• There are additional attributes stored at the “definitions” level describing the current settings of the assembly window;

o “ShowDots” – “0” for no, “1” for yes – determines if matches to the reference are shown as dots or standard residue characters.

Saving Alignments

The user can save the alignment at any time. The usual rules apply;

• Save As... is always enabled. The user can choose to save the alignment under a different name or in a different format.

• Save is only enabled if the alignment has been edited since the last save. It is also enabled for new alignments, but will bring up the “Save As...” dialog as a valid file name is not initially specified.

Saving the Modified Reference Sequence

The user can choose “Save As...” at any time and choose “MacVector NA Sequence” format from the “Format” popup menu. This saves the reference sequence (complete with any annotations and features) in standard MacVector single sequence format.

Loading Alignments

The user can open an existing alignment at any time using the “File:Open” command. Currently, alignments files can only be viewed/selected if “All Documents” or “All Readable Documents” is selected in the “Enable” popup menu. After selection, the alignment will load and be positioned at the start of the alignment. The status of the “show dots” toolbar button is remembered, but all other display information is reset to the default.

Printing Alignments

You print from the **Editor** tab, though only the currently displayed segment of the alignment will be printed. However, a new **Text** tab was added in MacVector 10.5 that displays the complete alignment as plain text. This view can be printed, and can also be copied to the Clipboard and pasted into other text editing applications. You can adjust the line length by clicking on the **Prefs** toolbar icon. All other settings are controlled by the appearance of the Editor pane (e.g. **Show Dots**, sequence ordering, position of consensus etc).