MacVector 17.5

for Mac OS X

Workshop: What's New in MacVector?

MacVector, Inc.
Software for Scientists

Copyright statement

This version of "What's New in MacVector" was published in January 2020.

Contents

CONTENTS 3

INTRODUCTION 4

WORKSHOP 4

MacOS Mojave Dark Mode 4
Restriction Enzyme Picker 5
Outlining Shared Domains in Aligned Sequences 8
Gibson/Ligase Independent Cloning 9
Enhanced Help with Video Tutorials 16
Genome Comparisons by Feature 18
Scan DNA – Open Reading Frames 24
Scan DNA – Missing Features 26
Scan DNA – Primers 28
MacVector with Assembler – Job Objects 31
MacVector with Assembler – SPAdes 34
MacVector with Assembler – Flye 37
Align to Reference – Quality Values 38
Align to Reference – Problems Tab 40

Introduction

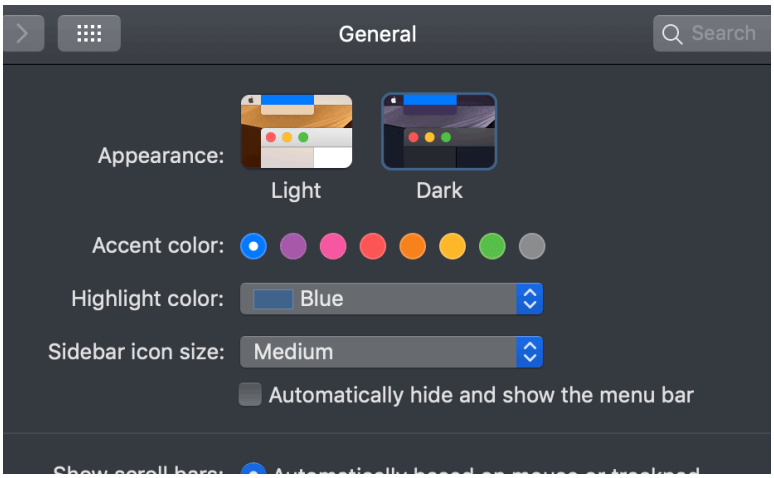
We are constantly releasing new versions of MacVector with new and improved functionality This workshop aims to bring long-term MacVector users up to speed with the latest functionality added to MacVector so you can see how it may benefit your everyday workflows.

Workshop

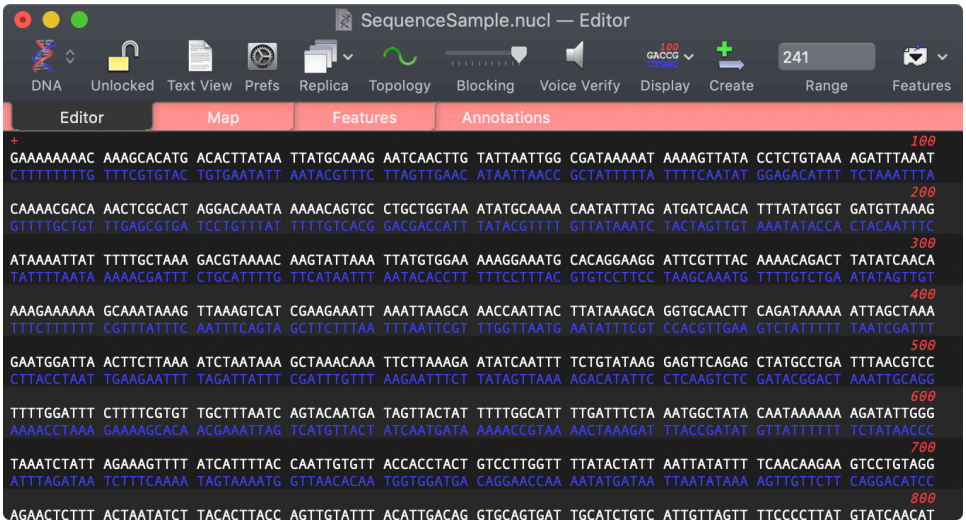
MacOS Mojave Dark Mode

MacVector 17.5 has had a large number of minor graphical enhancements made to better support the new "Dark Mode" feature introduced with macOS Mojave (macOS 10.14) and first supported in MacVector 17.0.

If you are running macOS Mojave, open **Apple | System Preferences** and click on the **General** option. Click on the **Dark** appearance icon to enable "Dark Mode"



The display updates so that all the windows have dark backgrounds with light text. MacVector not only supports the dark backgrounds, but many of the icons have been modified to that they "pop" more when running in dark mode.



Not every window in MacVector fully responds to Dark Mode. In particular, the **Map** tab always reflects the absolute colors you set in the *Symbols* editor. If you plan on using Dark Mode for most of your work, you should use the **MacVector | Preferences** -> **Map View** -> **Change Default Symbol Appearance** function to modify the colors for e.g. the **Title**, **Sequence** and **Ruler** options. MacVector 17.5 has some additional enhancements to better support switching between Light and Dark modes. For example, the defaults for the chromatogram colors automatically adjust so that the "G" traces are black in Light Mode and white in Dark Mode.

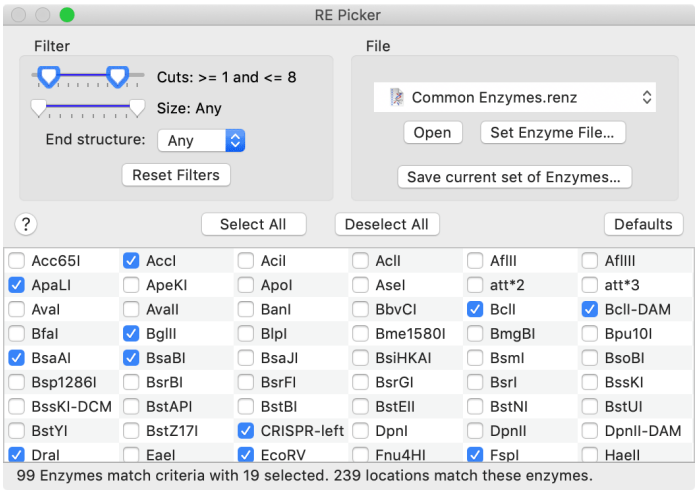
Open **Apple | System Preferences** and click on the **General** option. Click on the **Light** appearance icon to return to the normal "Light Mode"

Restriction Enzyme Picker

A new feature in MacVector is the *Restriction Enzyme Picker* (RE Picker).

Open any DNA sequence. This example uses `/Applications/MacVector/Tutorial Files/Align to Reference/Sequence Confirmation/SequenceSample.nucl` but any DNA sequence will suffice. Switch to the **Map** tab.

The *RE Picker* window opens;



By default, the window is quite large and can get in the way while you are trying to view or manipulate the **Map** tab.

Click on the **RE Picker** button on the **Map** tab toolbar. 

The *RE Picker* window closes and the icon changes to indicate it is hidden; 

Click on the **RE Picker** button again to show the *RE Picker* window.

The *RE Picker* shows an interactive list of restriction enzymes. Only those that are shown in the table and checked are displayed in the **Map** tab.

Slide the right slider of the **Cuts** control and watch the **Map** tab.

Both the *RE Picker* and the **Map** tab update to reflect the changes. The **Map** tab always shows only those enzymes that are both visible in the *RE Picker* and that are selected.

Click on the **Defaults** button. This resets the *RE Picker* to its initial default settings. Now click the checkbox next to the *Bgl*II item.

The single BglII site at 1,844 in `SequenceSample` hides and shows as you toggle the checkbox.

Slide both the left and right sliders all the way to the left.

The **Cuts** label should now indicate "0". The enzymes now visible in the *RE Picker* are all those in the default restriction enzyme file that do NOT cut the target molecule.

Click **Save current set of enzymes**... and save to your desktop with the name `Non-cutters.enz`.

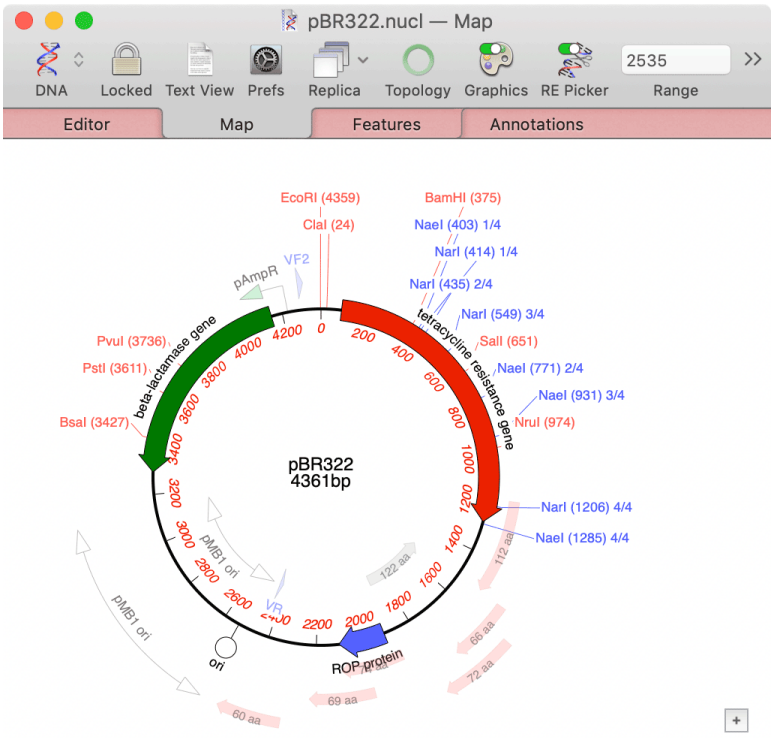
So, we now have a file that contains all of the enzymes that do not cut `SequenceSample`. Let's put this to use analyzing a different sequence.

Open the file `/Applications/MacVector/Sample Files/pBR322.nucl`. Make sure the **Map** tab is active.

Immediately we see the enzymes present in pBR322 using the default settings.

Click on the **Set enzyme file** button and navigate to select the `Non-cutters.renz` file you saved earlier.

The *RE Picker* now just shows those enzymes that did not cut `SequenceSample` and the `pBR322` **Map** tab refreshes to show those that were originally selected.

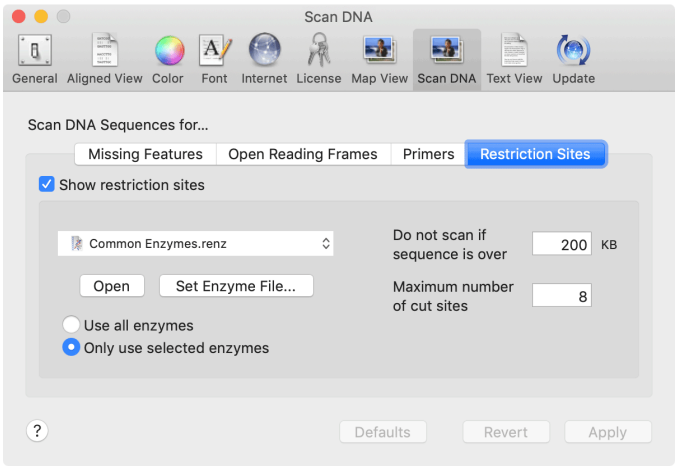


This is just one simple example of the power to be easily able to create and manipulate subsets of enzymes to help identify those that are useful for different cloning strategies.

Click on the **Defaults** button in the *RE Picker*

The `pBR322` **Map** tab once again refreshes to display many more enzymes and we see that the enzyme file is once again set to `Common Enzymes`. When you set an enzyme file as we did above it affects only the current sequence document. If you want to change the default settings used by the *RE Picker*, do this;

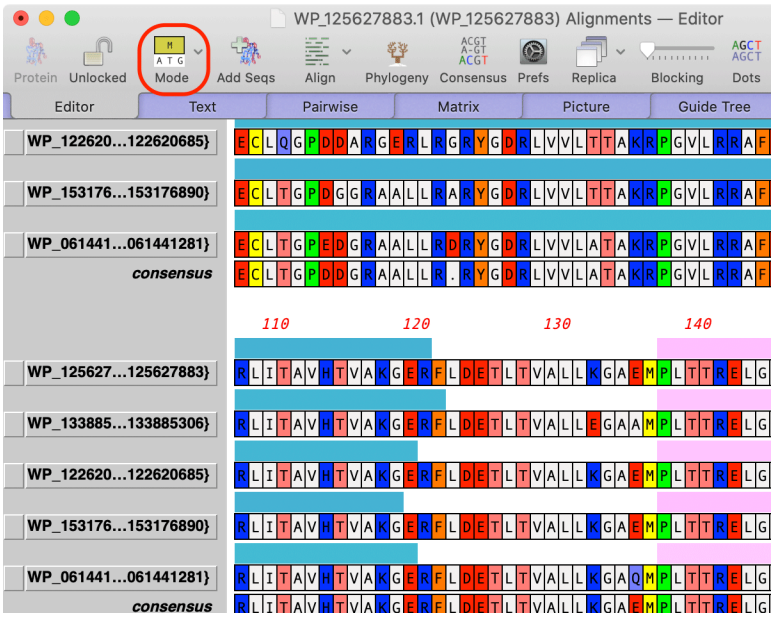
Select the **MacVector | Preferences**... menu item and click on the **Scan DNA** icon, then on the **Restriction Sites** tab.



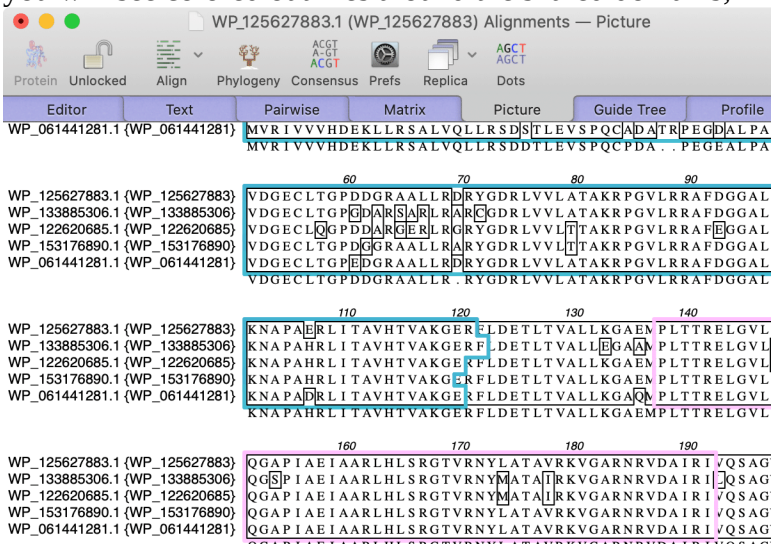
This is where you can set the default starting parameters for the *RE Picker*. If you make changes here, you will need to (a) press the **Apply** button or close the dialog, then (b) click on the *RE Picker* **Defaults** button to force a refresh of any open documents.

Outlining Shared Domains in Aligned Sequences

Multiple Sequence Alignments now retain feature information from their individual input sequences and can use this information to outline shared domains in the aligned sequences. To use this feature, first individually annotate the sequences you want to align, make sure the domains/features you are interested in are visible and set the **Fill** color to the color you would like to see in the alignment. Then add the sequences to a multiple sequence alignment document and align in the usual way (or, keep the single sequence documents open and choose **Analyze | Align Multiple Sequences Using...**). Then click on the **Mode** toolbar button (shown below) and select **Show Features**



This turns on a simple feature display mode in the **Editor** tab where you can see the extent and color of the features. When you switch to the **Picture** tab, you will see colored outlines around the shared domains;



The key to this functionality is that the individual sequence must be annotated ahead of time in a single sequence document, before being added to the alignment. The colors are taken from the **Fill** color of the graphical representation of the feature. In addition, to be considered "shared", the features must be of the same type and have the same displayed label.

Gibson/Ligase Independent Cloning

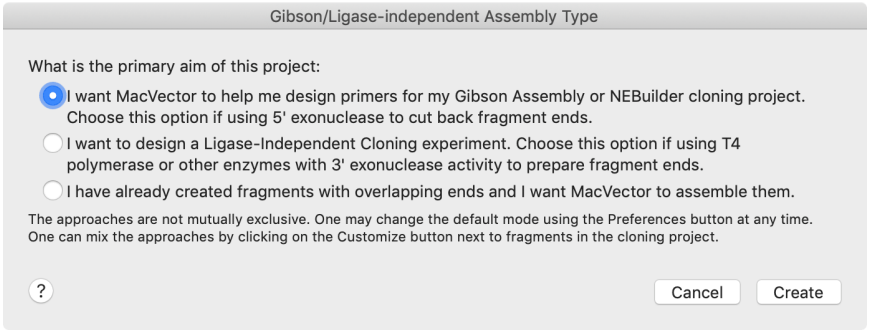
MacVector 17 introduced a new project-based interface for designing and documenting Gibson assembly and ligase-independent cloning experiments (e.g. the popular "Infusion" system).

For this example, we will ask MacVector to design a pair of primers so that we can clone a fragment into a vector. While many Gibson Assembly projects might have all of the required fragments be generated by PCR, you can also often just provide a microgram or so of linearized vector as one of the fragments, and thus you just need two appropriate primers to amplify a target fragment.

Creating a Project

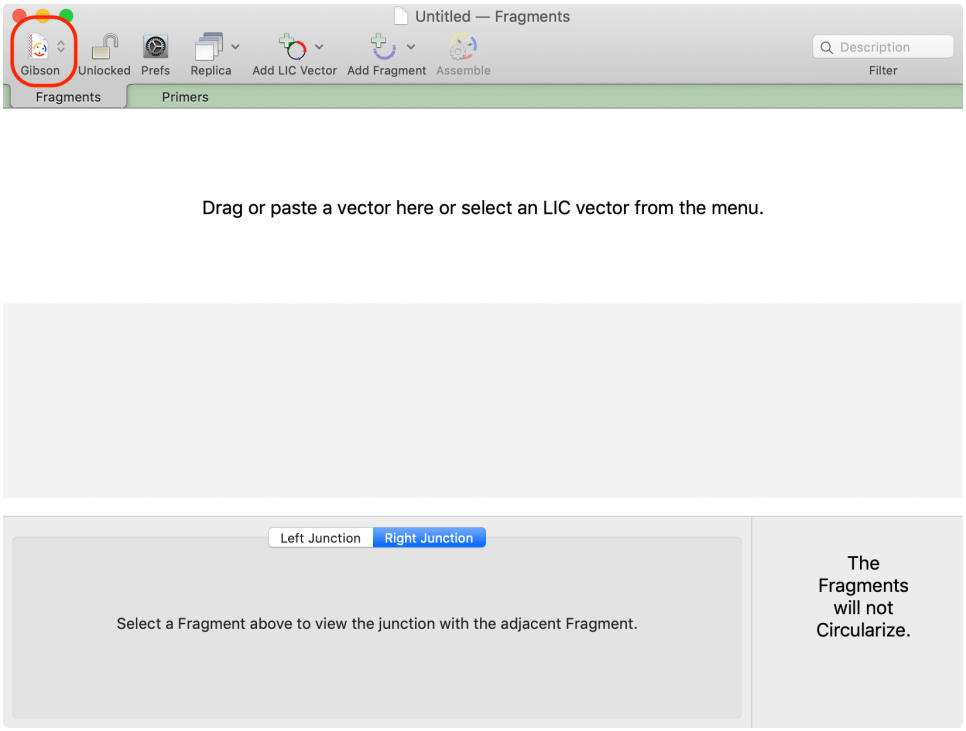
Select **File | New | Gibson/Ligase-Independent Assembly**... to create a new Gibson Assembly project.

You first need to decide what type of project you are planning on. While you can change this later, it usually easier to do this at the beginning.



Make sure you have the first "Gibson Assembly" option selected and click on the **Create** button.

A new *Gibson Assembly Project* window opens. Notice that the **Mode** button is set to **Gibson**. If you were designing primers for an Infusion experiment, you would choose the second "3' exonuclease" option.

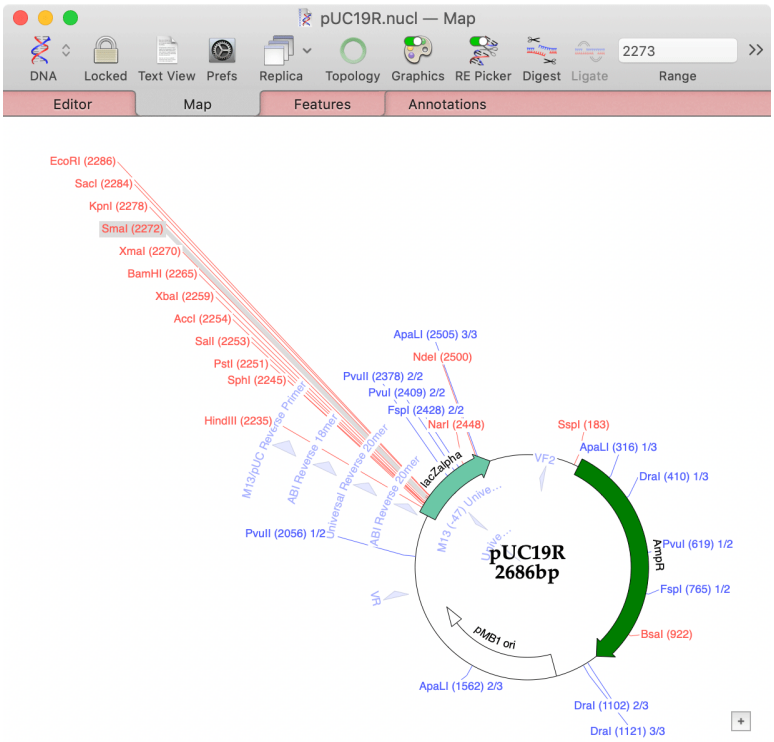


This is a primary document window, meaning it can be saved and opened later with all settings and contents preserved. It is highly interactive – there are many ways you can add fragments you would like to use in the assembly to the window.

Adding Vectors and Fragments

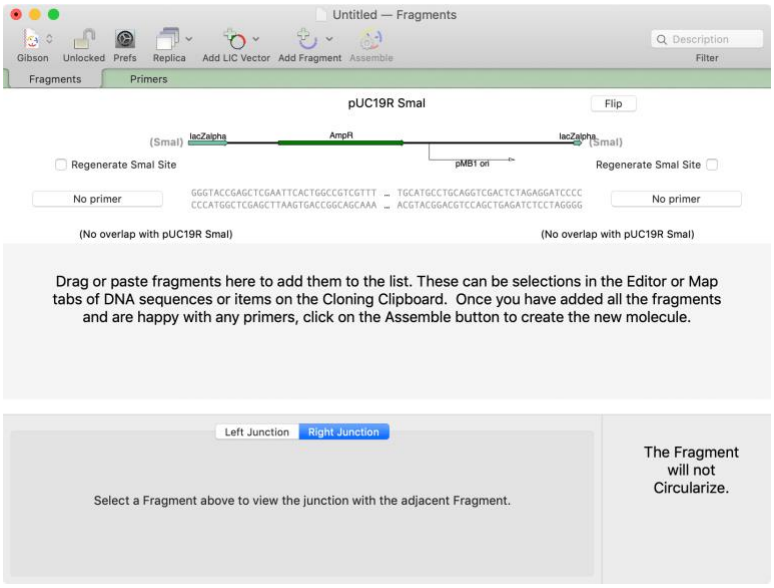
Let's do a vector first;

Open `pUC19R.nucl` (in the `/Applications/MacVector/Tutorial Files/GibsonAssembly/` folder). Switch to the **Map** tab. Select the *Sma I* site.



Click again on the *SmaI* site and carefully drag the selected site over the ***Gibson Assembly Project*** window and release the mouse.

A linearized copy of pUC19, split at the *SmaI* site, appears in the project;

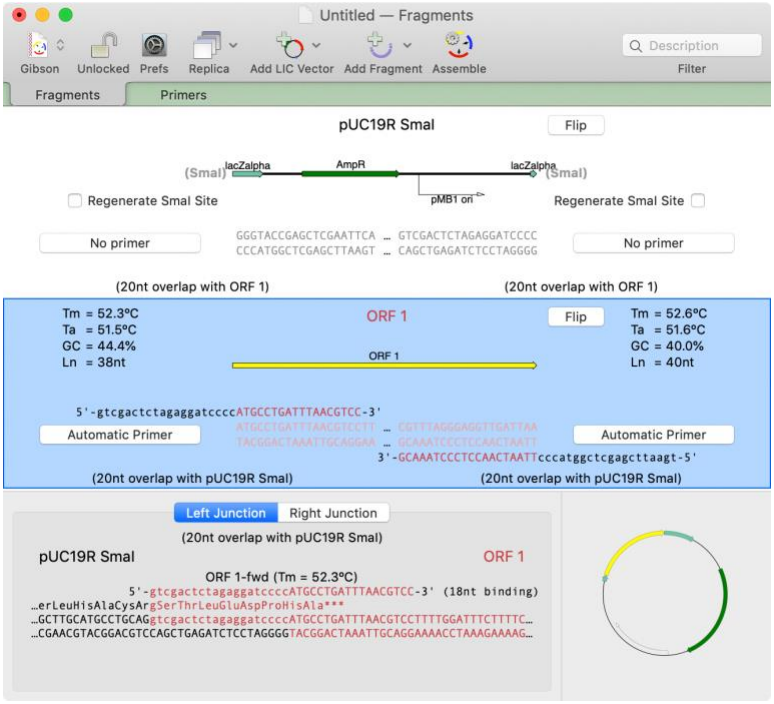


Note that when you add a vector like this (the core backbone of the vector with replication origin and selectable marker is always assumed to be the first fragment in the list), MacVector assumes you will be providing the fragment as a cut vector, rather than as the result of a PCR amplification.

Accordingly, each end is set to *No Primer*, indicating that the fragment will be accepted "as is". That means that if you add additional fragments to the project, MacVector will generate primers with extra-long tails to make sure they overlap the ends of the vector with enough residues to

Open `SequenceSample.nucl` (in the `/Applications/MacVector/Tutorial Files/GibsonAssembly/` folder). Switch to the **Map** tab. Select the yellow ORF 1 graphic, hold down the mouse button and carefully drag the item onto the second pane in the Gibson Assembly window.

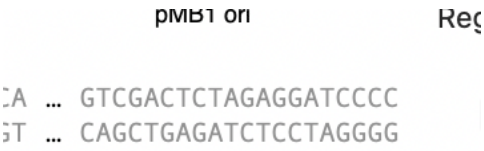
The display updates with the ORF 1 open reading frame in the second pane. However, MacVector has also automatically calculated suitable primers that could be used to amplify the ORF 1 open reading frame and provide an overlap with the *Sma*I digested pUC19R vector backbone.



Tails on Primers

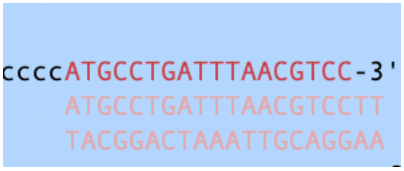
MacVector adds overhanging tails to the primers to generate suitable repeats between the ends of adjacent fragments in the final construct. Because in this case the vector backbone will be used "as is", and no primers will be used to amplify it, the primers for the insert fragment have to have much longer tails (in this case, 20nt) in order to provide sufficient repeats to enable recombination. However, we can see that the "tail" for the forward primer;

…exactly matches the sequence for the 3' end of the vector;

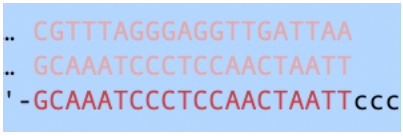


Balanced Primer Binding Tms

MacVector also tries to ensure that the forward and reverse primers have a closely matched Tm to help ensure efficient PCR amplification. In this case, the forward primer has 18nt that bind to the 5' end of the insert fragment giving it a predicted Tm of 52.3oC;



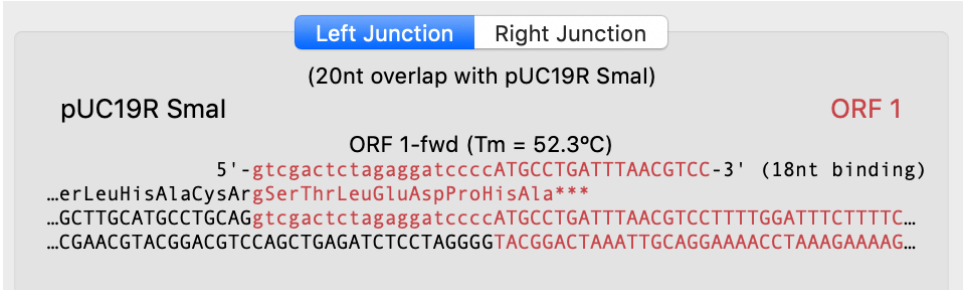
…whereas the reverse primer has been given a longer 20nt binding region, but that gives a predicted Tm of just 52.6oC, very close to that of the forward primer;



Junction Structure

The lower pane displays the details of the fragment junctions.

Click in the *ORF 1* panel to select it, then click on the **Left Junction** tab.



The junction shows the primer(s) used to generate the overlap (only one in this case) and color codes the residues so that you can see where the different sequences are derived from. By convention, for the duplicated sequence regions, MacVector shows the upper strand colored according to the fragment that provided the 5' sequence and the lower strand in the other color, so that the overlap can be viewed as the region with the complementary colors. The primer(s) used are shown above the sequence for the forward primer and below for the reverse primer (not used in this example). "Tails" are shown in lower case.

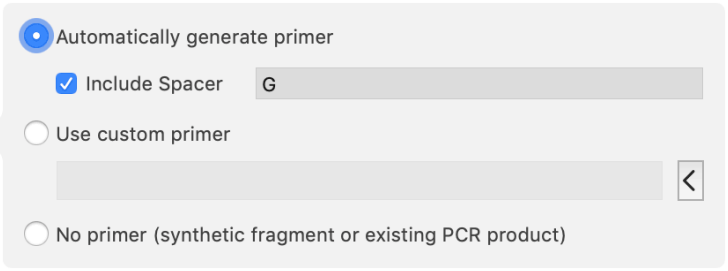
Finally, translations are shown immediately above the DNA sequence. It is important to understand that these key off existing CDS annotations in the fragment sequences, with preference given to CDS features coming into the junction from the 5' direction. So, in this case, the *lac*Z alpha CDS feature from pUC19 has precedent over the ORF 1 CDS. You can clearly see that the junction between lacZ alpha and the ORF 1 CDS is not in frame and the predicted translation terminates shortly after it passes the ATG start codon of ORF 1.

Inserting Spacer Residues

Let's add some extra residues to fix the frame.

Click on the **Automatic Primer** button on the left side of the ORF 1 pane

A popup window appears that lets you change how you want the primer to be created;



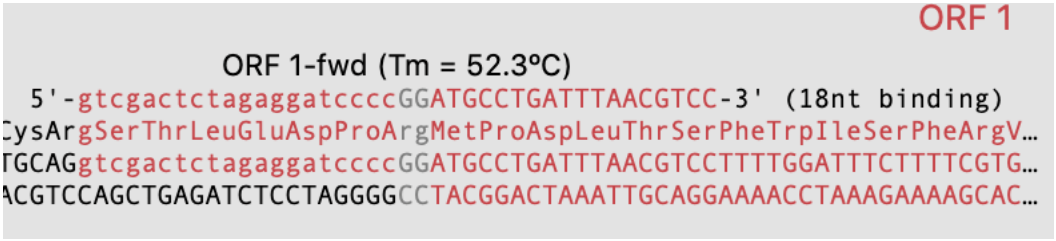
Select the **Include Spacer** checkbox, then type a "G" in the adjacent edit box.

The junction immediately updates to show the effect of the extra "G" (shown in gray).

However, we can see that is still not enough to fuse the two frames.

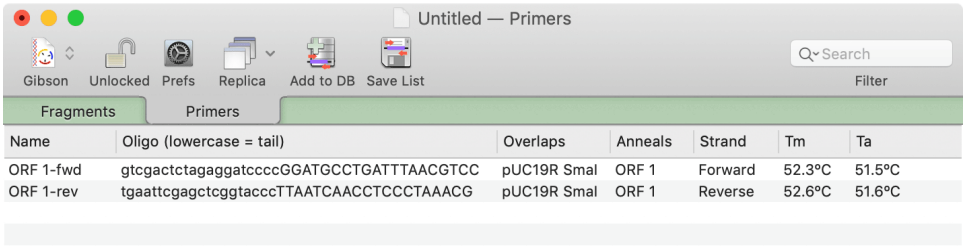
Type a second "G" in the edit box.

Now the junction shows that we have an in-frame fusion.



Now that we are happy with the primers, we can view them in a printable spreadsheet format.

Click on the **Primers** tab.

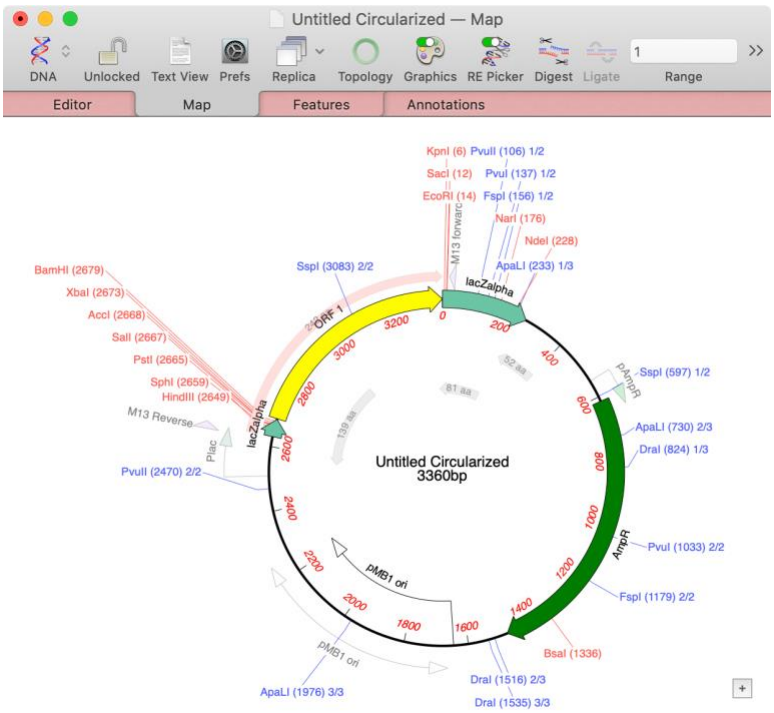


This view lists the primers, with appropriate names, along with their Tm and Ta values. The data can be printed, saved (in tab-separated or comma-separated values suitable for importing into Excel) and/or the primers added to the default MacVector Primer Database to be used in additional analyses.

Finally, the predicted construct sequence can be created;

Switch back to the **Fragments** tab. Click on the **Assemble** button.

A new window appears containing the predicted circular sequence.

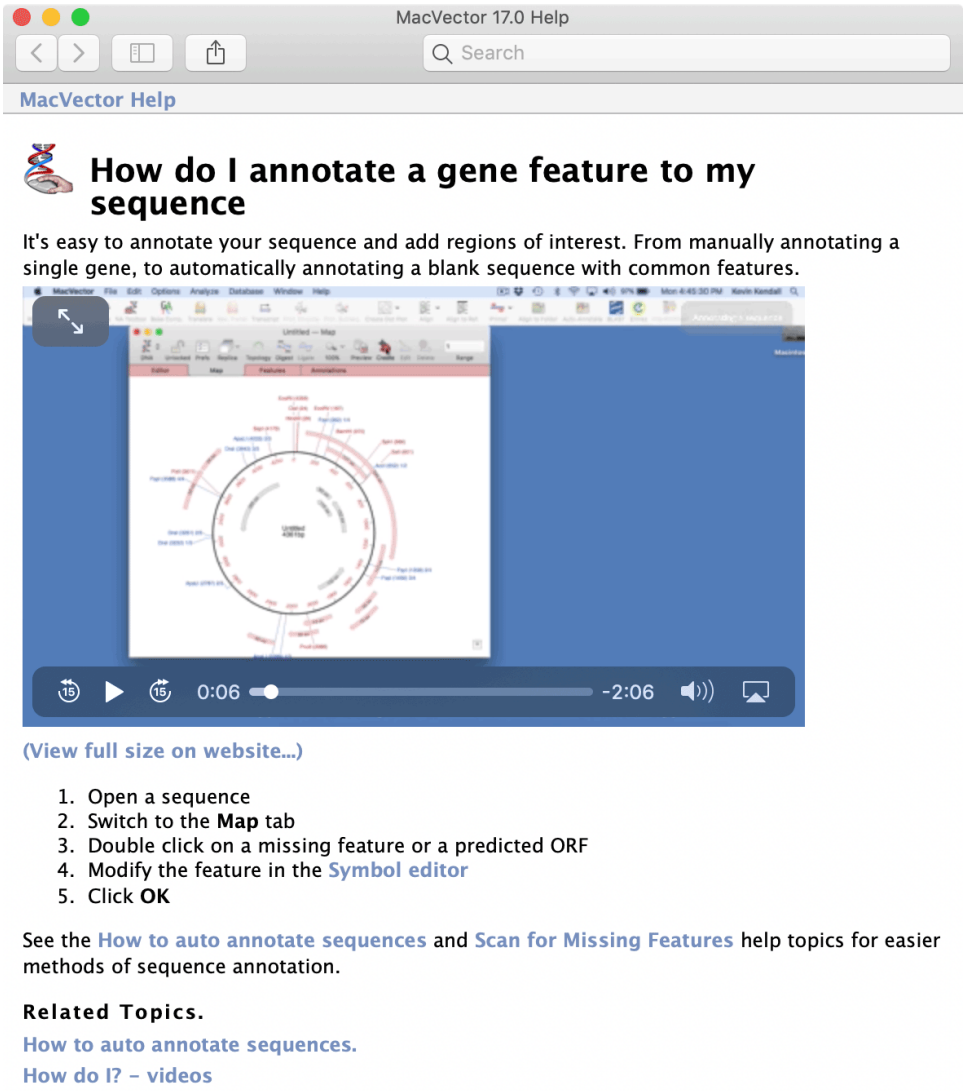


This short tutorial on Gibson Assembly only scratches the surface of what can be done in the interface. You can use your own custom primers and/or request regeneration of restriction enzyme sites and add as many fragments as you wish to the project, where MacVector will continue to try to balance the Tm's of the primers. Plus, the interface supports Ligase-independent cloning strategies, where vectors and fragments get cut back by T4 DNA polymerase, often in the presence of a single dNTP to generate long single stranded 5' overhangs. You can also simply provide your own pre-generated fragments with overlapping ends and let MacVector join them together for you.

Enhanced Help with Video Tutorials

There is a new **How Do I** menu that has links to a lot of common workflows;

Select one of the items – here we selected **Annotate a Gene to my Sequence**



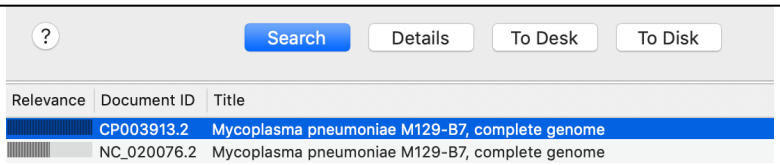
The floating help window opens at the appropriate topic. Many of these have short videos showing you how to perform the function.

Genome Comparisons by Feature

With the advent of cheap Next Generation Sequencing (NGS) technologies, it is becoming increasingly common for users to sequence an entire genome (especially with bacteria and viruses), often followed by annotation using the NCBI's Prokaryotic Genome Annotation Pipeline. The question then becomes, "what are the genetic changes in my strain that are responsible for the phenotype I observe?". MacVector 17 has an incredibly powerful tool that takes every annotated feature in a source genome and looks for that feature in a target genome to see if it exists, is annotated, and what changes are present. It is smart enough to consider translated CDS features and generates interactive lists that show identical, similar, weak and missing features. You can use the embedded interactive links to drill down to see the individual DNA and translated amino acid

changes that are potentially responsible for observed phenotypic differences. The example below uses two small Mycobacterial genomes that are not installed with MacVector, so we will need to download them from Entrez

Select **Database | Online Keyword Search for Sequences (Entrez)** and make sure the **Database** is set to *Nucleotide: Core Nucleotide db*. Then type `CP003913` into the **All Fields** edit box and press **Search**.

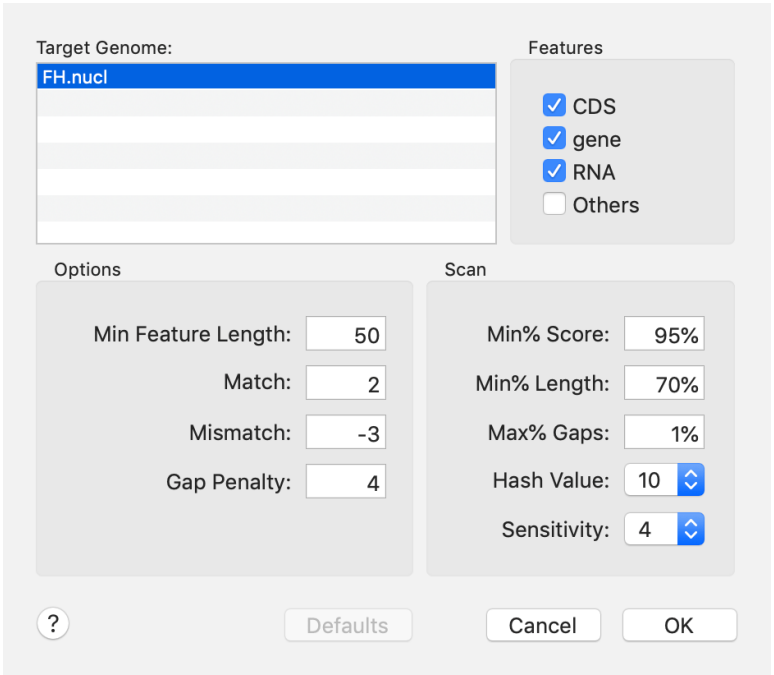


Select the first hit as shown, then click the **To Desk** button. The fully annotated sequence is downloaded and a new document window opens. Save the sequence to your desktop with the name `M129`.

Repeat the search with the accession number `CP010546`. Save this sequence to your desktop with the name `FH`.

We now have two small annotated bacterial genomes that won't take long to analyze.

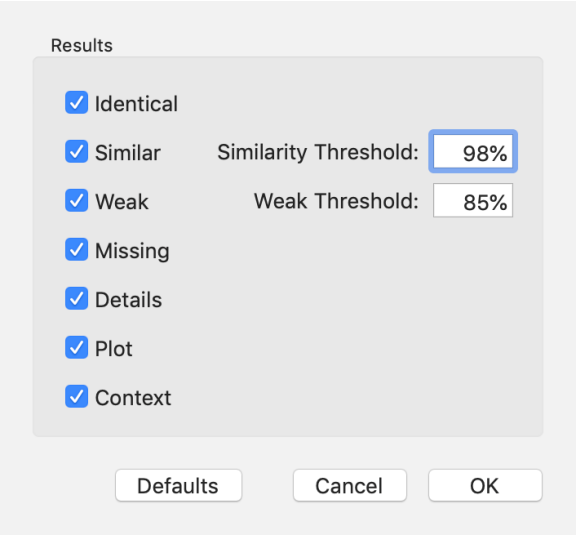
Bring `M129` to the front and select **Analyze | Compare Genomes by Feature**.



For now, let's accept the default settings.

If the **Defaults** button is active, click on it. Click **OK**.

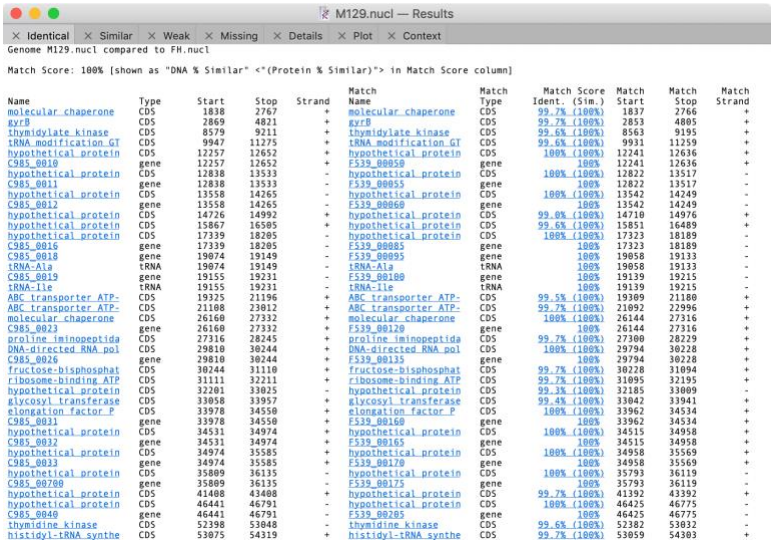
Depending on the speed of your machine, the analysis calculations may take between 5 and 20 seconds. Then a result dialog will appear;



These genomes are actually very closely related as *Mycoplasma pneumoniae* strains tend to be very genetically homogenous. So we will adjust the **Similarity Threshold** to be 98% rather than the default 95%.

Make sure all the checkboxes are selected, adjust the **Similarity Threshold** to 98% and click **OK**.

A window appears containing a tab for each of the checkboxes;



Name	Type	Start	Stop	Strand	Match Name	Match Type	Match Score Ident. (Sim.)	Match Start	Match Stop	Match Strand
molecular chaperone	CDS	1838	2767	+	molecular chaperone	CDS	99.7% (100%)	1837	2766	+
gyrB	CDS	2869	4821	+	gyrB	CDS	99.7% (100%)	2853	4805	+
thymidylate kinase	CDS	8579	9211	+	thymidylate kinase	CDS	99.6% (100%)	8563	9195	+
tRNA modification GT	CDS	9947	11275	+	tRNA modification GT	CDS	99.6% (100%)	9931	11259	+
hypothetical protein	CDS	12257	12652	+	hypothetical protein	CDS	100% (100%)	12241	12636	+
C985_0010	gene	12257	12652	+	F539_00050	gene	100%	12241	12636	+
hypothetical protein	CDS	12838	13533	-	hypothetical protein	CDS	100% (100%)	12822	13517	-
C985_0011	gene	12838	13533	-	F539_00055	gene	100%	12822	13517	-
hypothetical protein	CDS	13558	14265	-	hypothetical protein	CDS	100% (100%)	13542	14249	-
C985_0012	gene	13558	14265	-	F539_00060	gene	100%	13542	14249	-
hypothetical protein	CDS	14726	14992	+	hypothetical protein	CDS	99.0% (100%)	14710	14976	+
hypothetical protein	CDS	15867	16505	+	hypothetical protein	CDS	99.6% (100%)	15851	16489	+
hypothetical protein	CDS	17339	18205	-	hypothetical protein	CDS	100% (100%)	17323	18189	-
C985_0016	gene	17339	18205	-	F539_00085	gene	100%	17323	18189	-
C985_0018	gene	19074	19149	-	F539_00095	gene	100%	19058	19133	-
tRNA-Ala	tRNA	19074	19149	-	tRNA-Ala	tRNA	100%	19058	19133	-
C985_0019	gene	19155	19231	-	F539_00100	gene	100%	19139	19215	-
tRNA-Ile	tRNA	19155	19231	-	tRNA-Ile	tRNA	100%	19139	19215	-
ABC transporter ATP-	CDS	19325	21196	+	ABC transporter ATP-	CDS	99.5% (100%)	19309	21180	+
ABC transporter ATP-	CDS	21108	23012	+	ABC transporter ATP-	CDS	99.7% (100%)	21092	22996	+
molecular chaperone	CDS	26160	27332	+	molecular chaperone	CDS	100% (100%)	26144	27316	+
C985_0023	gene	26160	27332	+	F539_00120	gene	100%	26144	27316	+
proline iminopeptida	CDS	27316	28245	+	proline iminopeptida	CDS	99.7% (100%)	27300	28229	+
DNA-directed RNA pol	CDS	29810	30244	+	DNA-directed RNA pol	CDS	100% (100%)	29794	30228	+
C985_0026	gene	29810	30244	+	F539_00135	gene	100%	29794	30228	+
fructose-bisphosphat	CDS	30244	31110	+	fructose-bisphosphat	CDS	99.7% (100%)	30228	31094	+
ribosome-binding ATP	CDS	31111	32211	+	ribosome-binding ATP	CDS	99.7% (100%)	31095	32195	+
hypothetical protein	CDS	32201	33025	-	hypothetical protein	CDS	99.3% (100%)	32185	33009	-
glycosyl transferase	CDS	33058	33957	+	glycosyl transferase	CDS	99.4% (100%)	33042	33941	+
elongation factor P	CDS	33978	34550	+	elongation factor P	CDS	100% (100%)	33962	34534	+
C985_0031	gene	33978	34550	+	F539_00160	gene	100%	33962	34534	+
hypothetical protein	CDS	34531	34974	+	hypothetical protein	CDS	100% (100%)	34515	34958	+
C985_0032	gene	34531	34974	+	F539_00165	gene	100%	34515	34958	+
hypothetical protein	CDS	34974	35585	+	hypothetical protein	CDS	100% (100%)	34958	35569	+
C985_0033	gene	34974	35585	+	F539_00170	gene	100%	34958	35569	+
hypothetical protein	CDS	35809	36135	-	hypothetical protein	CDS	100% (100%)	35793	36119	-
C985_00700	gene	35809	36135	-	F539_00175	gene	100%	35793	36119	-
hypothetical protein	CDS	41408	43408	+	hypothetical protein	CDS	99.7% (100%)	41392	43392	+
hypothetical protein	CDS	46441	46791	-	hypothetical protein	CDS	100% (100%)	46425	46775	-
C985_0040	gene	46441	46791	-	F539_00205	gene	100%	46425	46775	-
thymidine kinase	CDS	52398	53048	-	thymidine kinase	CDS	99.6% (100%)	52382	53032	-
histidyl-tRNA synthe	CDS	53075	54319	+	histidyl-tRNA synthe	CDS	99.7% (100%)	53059	54303	+

The first tab lists all of the features that are perfectly conserved between the two genomes based on sequence identity, even if the names and qualifiers are different. CDS features are translated and the amino acid sequences compared, so there may be silent mutation differences in the encoding DNA sequences.

The first five columns are the "name", type, start, stop and strand of the feature in the parent sequence i.e. the sequence that you had frontmost when you invoked the search. The "name" is the label that appears in the **Map** tab for the feature. By default, for CDS features, this would be the */gene=* qualifier, but this can be configured on an individual feature basis or for all features of a type. See the *Creating Vector Maps* tutorial for more information on this.

The rightmost columns provide the same information for the feature(s) that matched on the target genome except that there are is an extra *Match Score* column. This displays the DNA identity score for each pair of features along with, (in brackets) the identity score for the predicted amino acid translation for CDS features given the current default genetic code.

Note that features that are duplicated in the target genome will show additional matches;

hypothetical protein	CDS	612796	613176	+	hypothetical protein	CDS	100% (100%)	607631	608011	+
C985_0510	gene	612796	613176	+	F539_02830	gene	100%	607631	608011	+
family K-like protei	CDS	614385	614639	+	family K-like protei	CDS	100% (100%)	609228	609482	+
					family K-like protei	CDS	97.0% (98.8%)	194874	195128	+
C985_0512	gene	614385	614639	+	F539_02840	gene	100%	609228	609482	+
					F539_00855	gene	97.0%	194874	195128	+
hypothetical protein	CDS	614921	617302	+	hypothetical protein	CDS	99.7% (100%)	609764	612145	+
MG032/MG096/MG288 fa	CDS	620619	622556	-	MG032/MG096/MG288 fa	CDS	99.7% (100%)	615461	617398	-

Note that when multiple matches are found, if one of them has a 100% match, all of the matching features are shown in the match list, even if they do not also have 100% identity. This approach ensures that you are always aware of duplicated/pseudogenes with significant but non-identical matches.

The display is highly interactive;

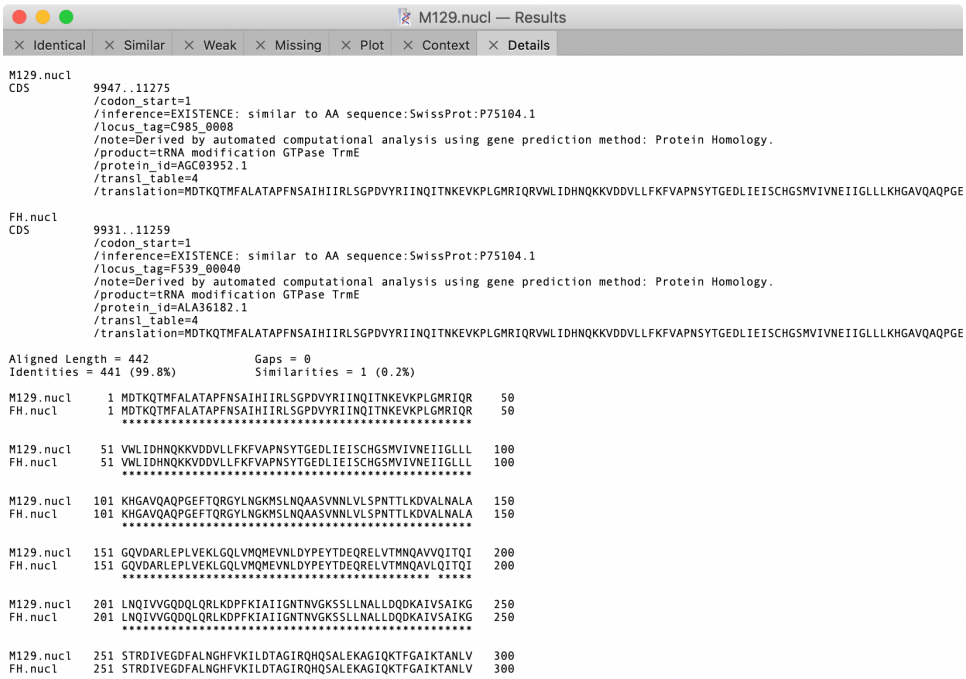
Click on any of the blue feature names in the first column.

The parent M129 sequence document is brought frontmost, switches to the Features tab and highlights and scrolls to the corresponding feature. So, you can use this shortcut to quickly jump to any feature of interest.

The same obviously applies to the target genome gene names.

Bring the result window back to the top and click on one of the links in the *Match Score* column – choose one from a CDS feature with the (extra%) column.

The window changes to select the **Details** tab;



As you scroll through the text output, you will see;

- Full GenBank definition for the parental sequence
- Full GenBank definition for the target sequence
- Aligned amino acid translations with a header containing identity and similarity information
- Aligned DNA sequences with a header containing scoring information

It can be awkward switching between tabs in this way to explore different features. MacVector has a solution!

Click and hold on the **Details** tab header, then drag the **Details** tab out of the result window to somewhere else on your desktop and let go.

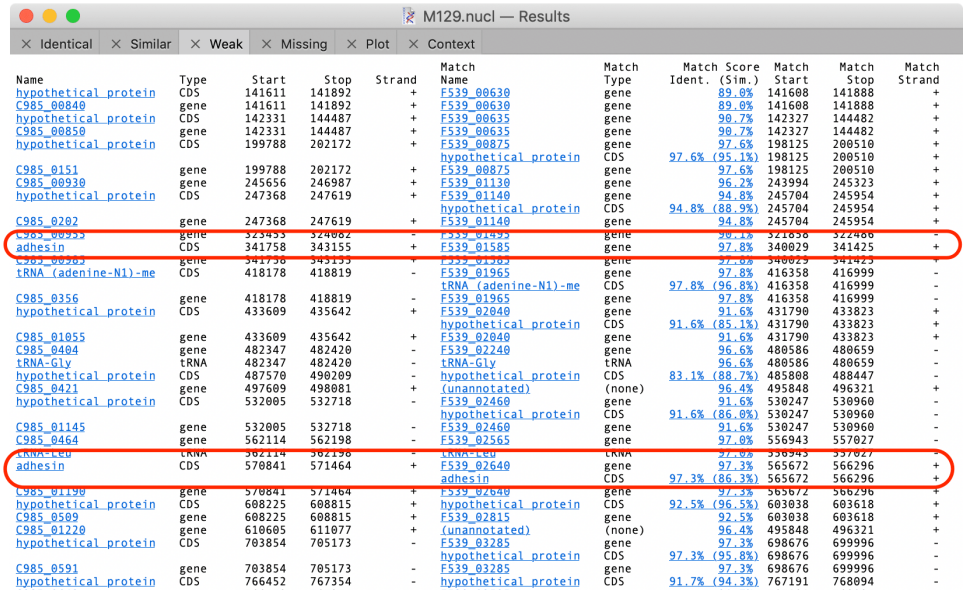
A new result window will open up containing just the **Details** tab. Now you can switch back to the Identical tab in the primary result window, click on other *Match Score* column entries and the **Details** tab window will update in real time with each click. Note that if you want to put the **Details** tab back on the main result window, you can just drag the tab back to where it came from.

Click on the **Similar** tab.

This shows "similar" features. Earlier we set the threshold to 98% so these really are almost identical, but might differ due to one or two residue changes in either DNA or translated CDS.

Click on the **Weak** tab.

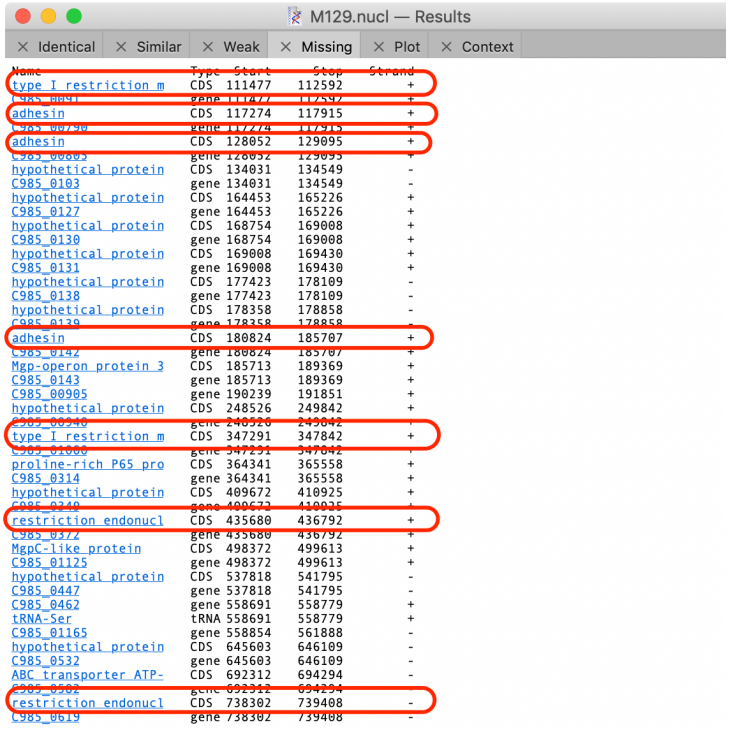
These are all the remaining matches that exceeded our initial search criteria but were not sufficiently similar to be included on the **Similar** tab.



Name	Type	Start	Stop	Strand	Match Name	Match Type	Match Score Ident. (Sim.)	Match Start	Match Stop	Match Strand
hypothetical protein	CDS	141611	141892	+	F539_00630	gene	89.0%	141608	141888	+
C985_00840	gene	141611	141892	+	F539_00630	gene	89.0%	141608	141888	+
hypothetical protein	CDS	142331	144487	+	F539_00635	gene	90.7%	142327	144482	+
C985_00850	gene	142331	144487	+	F539_00635	gene	90.7%	142327	144482	+
hypothetical protein	CDS	199788	202172	+	F539_00875	gene	97.6%	198125	200510	+
					hypothetical protein	CDS	97.6% (95.1%)	198125	200510	+
C985_0151	gene	199788	202172	+	F539_00875	gene	97.6%	198125	200510	+
C985_00930	gene	245656	246987	+	F539_01130	gene	96.2%	243994	245323	+
hypothetical protein	CDS	247368	247619	+	F539_01140	gene	94.8%	245704	245954	+
					hypothetical protein	CDS	94.8% (88.9%)	245704	245954	+
C985_0202	gene	247368	247619	+	F539_01140	gene	94.8%	245704	245954	+
C985_00955	gene	323453	324082	-	F539_01495	gene	90.1%	321858	322486	-
adhesin	CDS	341758	343155	+	F539_01585	gene	97.8%	340029	341425	+
C985_00985	gene	341758	343155	+	F539_01585	gene	97.8%	340029	341425	+
tRNA (adenine-N1)-me	CDS	418178	418819	-	F539_01965	gene	97.8%	416358	416999	-
					tRNA (adenine-N1)-me	CDS	97.8% (96.8%)	416358	416999	-
C985_0356	gene	418178	418819	-	F539_01965	gene	97.8%	416358	416999	-
hypothetical protein	CDS	433609	435642	+	F539_02040	gene	91.6%	431790	433823	+
					hypothetical protein	CDS	91.6% (85.1%)	431790	433823	+
C985_01055	gene	433609	435642	+	F539_02040	gene	91.6%	431790	433823	+
C985_0404	gene	482347	482420	-	F539_02240	gene	96.6%	480586	480659	-
tRNA-Gly	tRNA	482347	482420	-	tRNA-Gly	tRNA	96.6%	480586	480659	-
hypothetical protein	CDS	487570	490209	-	hypothetical protein	CDS	83.1% (88.7%)	485808	488447	-
C985_0421	gene	497609	498081	+	(unannotated)	(none)	96.4%	495848	496321	+
hypothetical protein	CDS	532005	532718	-	F539_02460	gene	91.6%	530247	530960	-
					hypothetical protein	CDS	91.6% (86.0%)	530247	530960	-
C985_01145	gene	532005	532718	-	F539_02460	gene	91.6%	530247	530960	-
C985_0464	gene	562114	562198	-	F539_02565	gene	97.0%	556943	557027	-
tRNA-Leu	tRNA	562114	562198	-	tRNA-Leu	tRNA	97.6%	556943	557027	-
adhesin	CDS	570841	571464	+	F539_02640	gene	97.3%	565672	566296	+
					adhesin	CDS	97.3% (86.3%)	565672	566296	+
C985_01190	gene	570841	571464	+	F539_02640	gene	97.3%	565672	566296	+
hypothetical protein	CDS	608225	608815	+	hypothetical protein	CDS	92.5% (96.5%)	603038	603618	+
C985_0509	gene	608225	608815	+	F539_02815	gene	92.5%	603038	603618	+
C985_01220	gene	610605	611077	+	(unannotated)	(none)	96.4%	495848	496321	+
hypothetical protein	CDS	703854	705173	-	F539_03285	gene	97.3%	698676	699996	-
					hypothetical protein	CDS	97.3% (95.8%)	698676	699996	-
C985_0591	gene	703854	705173	-	F539_03285	gene	97.3%	698676	699996	-
hypothetical protein	CDS	766452	767354	-	hypothetical protein	CDS	91.7% (94.3%)	767191	768094	-

They key here is that there are really not that many weak matches – the screenshot above shows almost all of them and most are uncharacterized genes. But two matches stand out, to adhesin genes. These are very important for host pathogenicity in Mycoplasma and the differences between them are responsible for typing the strains into Type 1 (M129) and Type 2 (FH).

Click on the **Missing** tab.



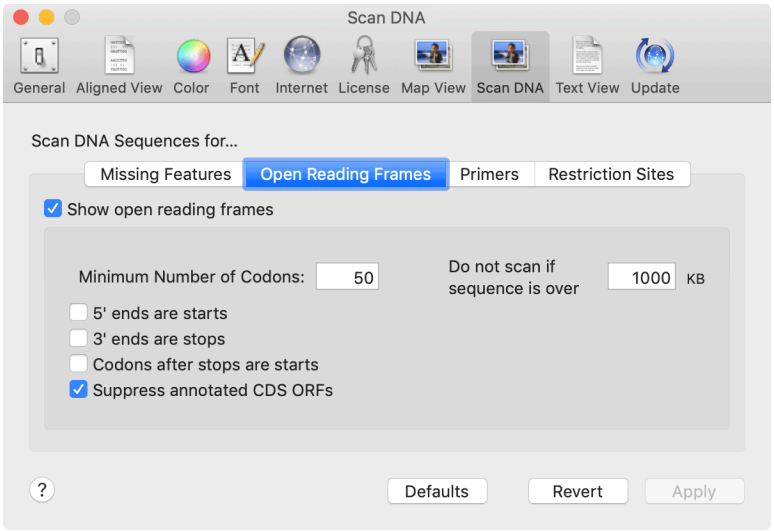
Name	Type	Start	Stop	Strand
type I restriction m	CDS	111477	112592	+
C985_00841	gene	111477	112592	+
adhesin	CDS	117274	117915	+
C985_00850	gene	117274	117915	+
adhesin	CDS	128052	129095	+
C985_00885	gene	128052	129095	+
hypothetical protein	CDS	134031	134549	-
C985_0103	gene	134031	134549	-
hypothetical protein	CDS	164453	165226	+
C985_0127	gene	164453	165226	+
hypothetical protein	CDS	168754	169008	+
C985_0130	gene	168754	169008	+
hypothetical protein	CDS	169008	169430	+
C985_0131	gene	169008	169430	+
hypothetical protein	CDS	177423	178109	-
C985_0138	gene	177423	178109	-
hypothetical protein	CDS	178358	178858	-
C985_0139	gene	178358	178858	-
adhesin	CDS	180824	185707	+
C985_0142	gene	180824	185707	+
Mgp-operon protein 3	CDS	185713	189369	+
C985_0143	gene	185713	189369	+
C985_00905	gene	190239	191851	+
hypothetical protein	CDS	248526	249842	+
C985_00940	gene	248526	249842	+
type I restriction m	CDS	347291	347842	+
C985_01000	gene	347291	347842	+
proline-rich P65 pro	CDS	364341	365558	+
C985_0314	gene	364341	365558	+
hypothetical protein	CDS	409672	410925	+
C985_0340	gene	409672	410925	+
restriction endonucl	CDS	435680	436792	+
C985_0372	gene	435680	436792	+
MgpC-like protein	CDS	498372	499613	+
C985_01125	gene	498372	499613	+
hypothetical protein	CDS	537818	541795	-
C985_0447	gene	537818	541795	-
C985_0462	gene	558691	558779	+
tRNA-Ser	tRNA	558691	558779	+
C985_01165	gene	558854	561888	-
hypothetical protein	CDS	645603	646109	-
C985_0532	gene	645603	646109	-
ABC transporter ATP-	CDS	692312	694294	-
C985_0582	gene	692312	694294	-
restriction endonucl	CDS	738302	739408	-
C985_0619	gene	738302	739408	-

These are all the features present in M129 that did not have matches in FH that exceeded our search parameters. First, note that most of the known "missing" genes are either adhesin genes or genes involved in the restriction-modification system. Again, these are well characterized variable genes in Mycobacteria, responsible for differences in pathogenicity and host specificity. Secondly, one of the limitations of the current iteration of the MacVector genome comparison tool is that the reason these genes are not matched is because of variable number of short repeats within the genes. If the repeats prevent at least 70% (the default) of the genes matching with 85% (default) identity, the match will not be reported.

For more information on the Genome Comparison tool, and for an exploration of the **Plot** and **Context** tabs and how to further analyze missing features, please take a look at the *Genome Feature Comparison* tutorial.

Scan DNA – Open Reading Frames

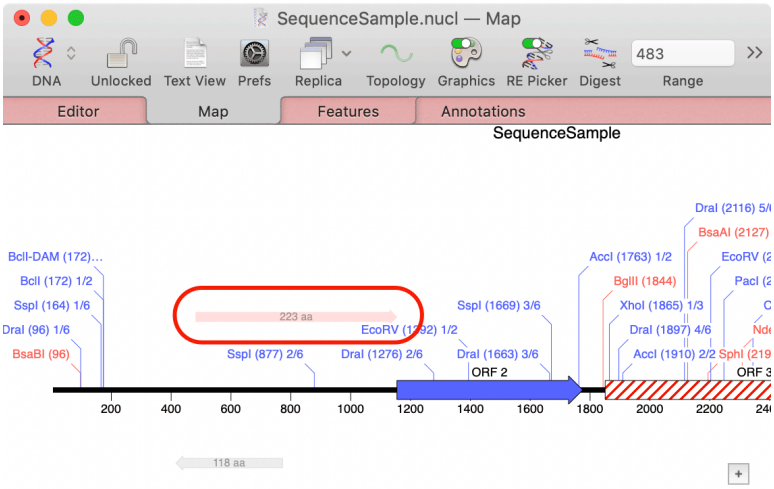
Whenever you open a DNA sequence, MacVector automatically scans it for the presence of a variety of different properties and features. We discussed the restriction sites and RE Picker above, but over the past few releases, there have been other searches added. The settings for these can all be accessed through the **MacVector | Preferences** -> **Scan DNA** tab.



Here you can control how open reading frames are displayed. You can turn them on/off and also the minimum length and how you want the ends of linear sequences to be handled.

> Make sure **Show open reading frames** is selected, bring `SequenceSample` to the front, select the yellow *ORF 1* graphic in the **Map** tab and press the `<delete>` key.

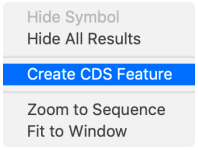
The **Map** updates to indicate there is no longer an ORF 1 feature, but there is now a pale pink arrow replacing it;



Plus strand open reading frames that exceed the default settings are shown in pale red, minus strand open reading frames in grey. Note that there is no ORF arrow shown over the ORF 2 or ORF 3 features. MacVector is intelligent enough to ignore open reading frames that have already been annotated as CDS features on the sequence.

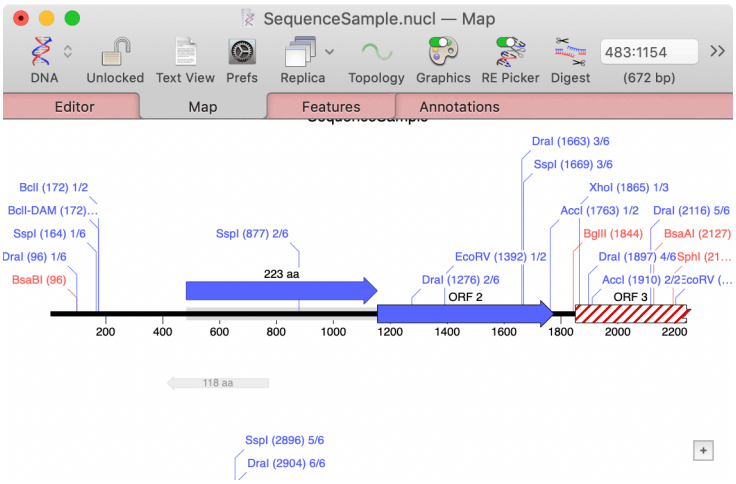
Right-click (or <ctrl>-click) on the ORF arrow where ORF 1 used to be.

A popup menu appears



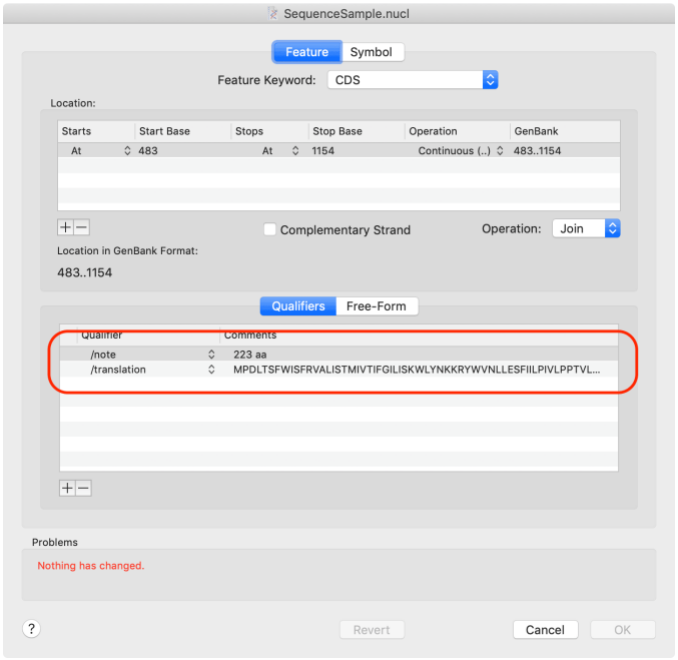
Choose **Create CDS Feature**

A new CDS feature appears, with the default appearance for CDS features.



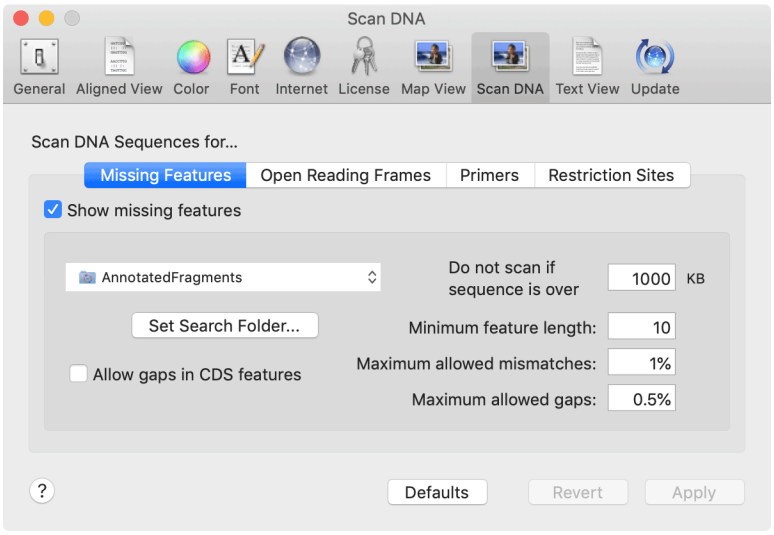
Double-click on the new CDS feature graphic

You can see that not only has a new feature been created, but the actual predicted translation has been added as a /translation qualifier;



Scan DNA – Missing Features

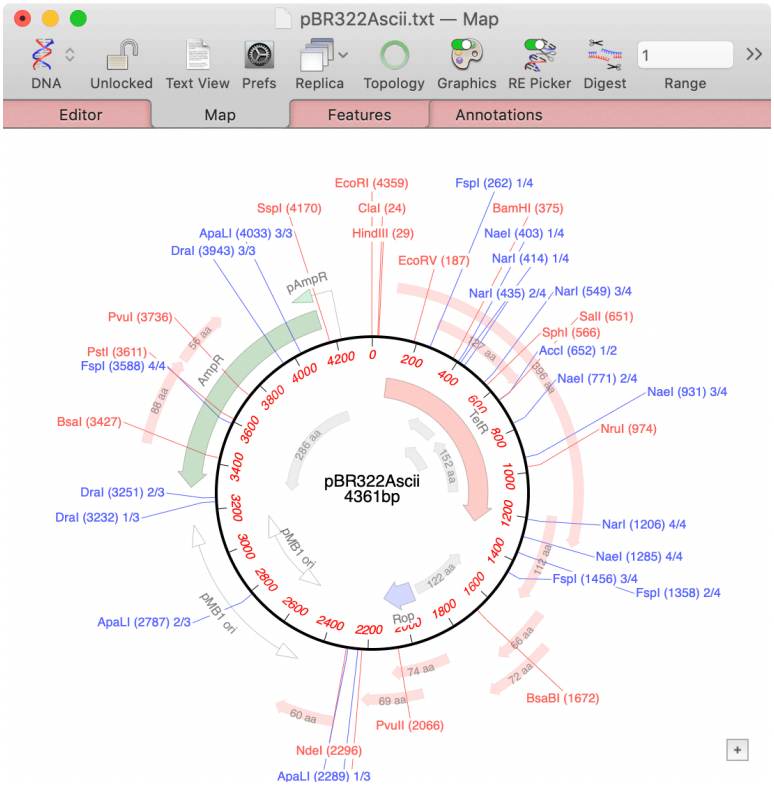
The *Scan DNA* function also scans opened DNA sequences for missing features, again controlled by the **MacVector | Preferences | Scan DNA** pane.



The key to this functionality is that it sequentially loads each of the files it can find in the **Search Folder**, takes the DNA sequence corresponding to

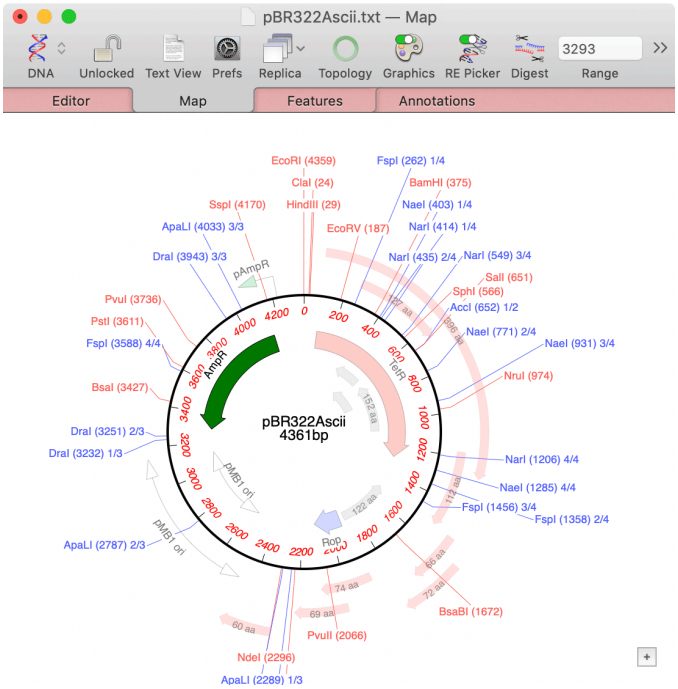
Make sure you have **Show missing features** selected, then open the sequence `/Applications/MacVector/Tutorial Files/AutoAnnotation/pBR322Ascii.txt`. It's actually a circular sequence, so click on the **Topology** button to tell MacVector.

The sequence opens, and there are a number of "greyed out" features around it;

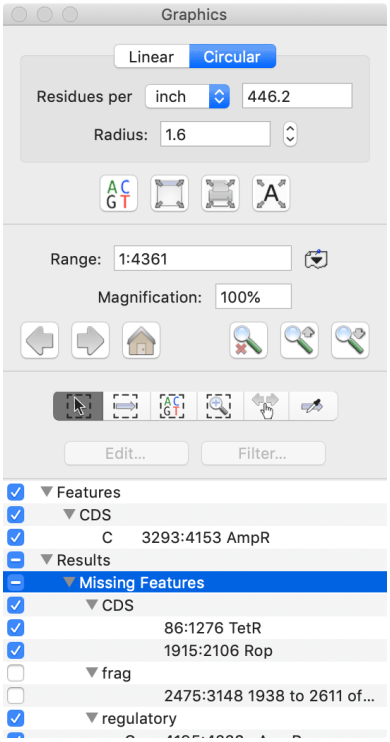


Right-click (or <cntl>-click) on the pale green *AmpR* graphic. Select **Create CDS Feature** in the resulting popup menu.

The display refreshes to show a bold AmpR gene.



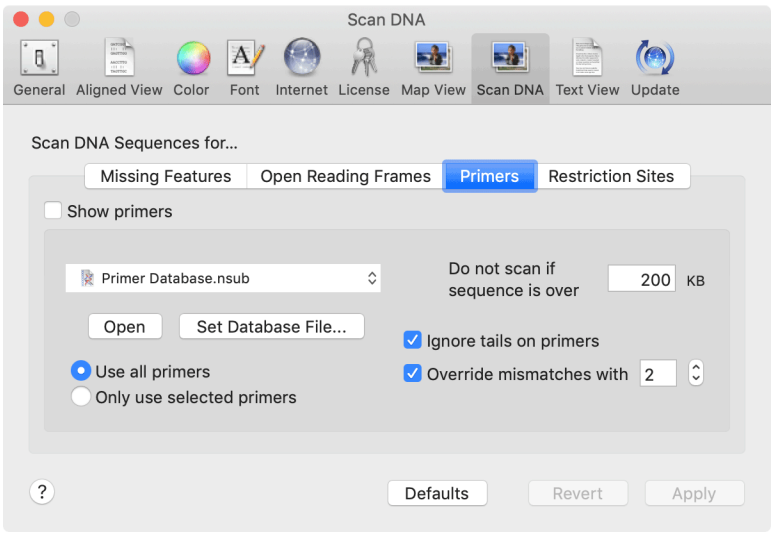
If you want to add ALL of the missing features quickly, click on the *Missing Features* list item in the floating graphics palette;



That selects them all, then you can right-click on the main **Map** tab and choose **Create Features** to add them all in one mouse click.

Scan DNA – Primers

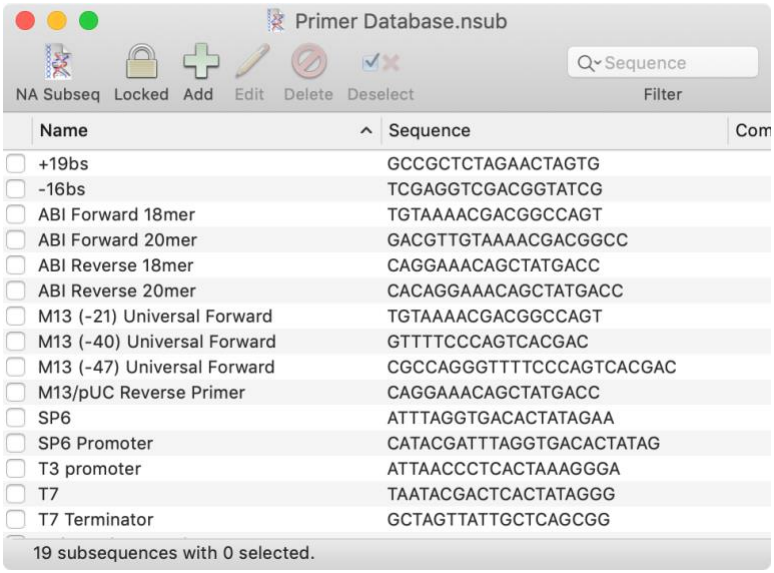
With version 17, MacVector also scans for primer binding sites. Again, this is controlled by the appropriate **Scan DNA** tab;



The default source file for the scan is `Primer Database.nsub`, a simple file containing a few common universal primers.

The file is located in `/Applications/MacVector/Subsequences/`

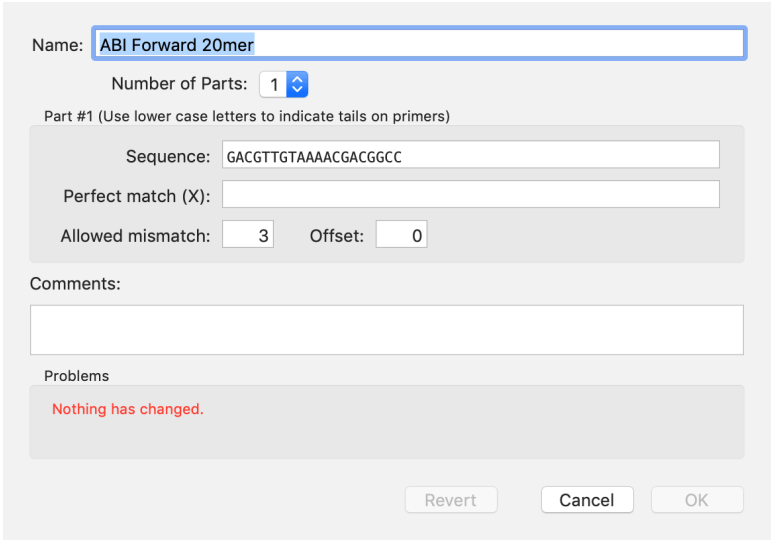
Find the file on your computer and open it with MacVector, or simply click the **Open** button under the filename.



This is a variant of the normal MacVector nucleic acid subsequence file format.

Double-click on the *ABI Forward 20mer* item.

The subsequence editor appears.



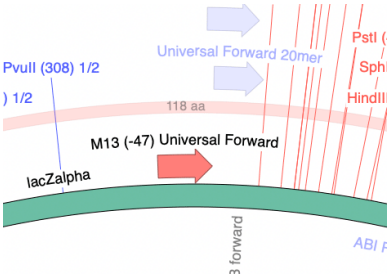
Primers only ever have a single part. However, one enhancement is that they can have 5' leading lower case "tails". You can type lower case residues at the 5'end of the primer and this sequence will be treated specially by MacVector. It will not be treated as part of the core primer binding site, but it WILL be included in any generate PCR fragments. For more details on how this works, take a look at the Primer Design Tutorial.

You can add your own primers to the list, either by hand or from other functions within MacVector, such as the **Quicktest Primer (individual)** or **Primer Design/Test (pairs)** functions. You can also create suitable Primer.nsub files from existing data stored in Excel spreadsheets. Look for the *PrimerConverter* utility on the macvector.com Downloads -> Utilities & AppleScripts page.

Open the file `/Applications/MacVector/Common Vectors/pUC/pUC19.nucl`. In the Map tab, zoom in around the lacZ alpha gene.

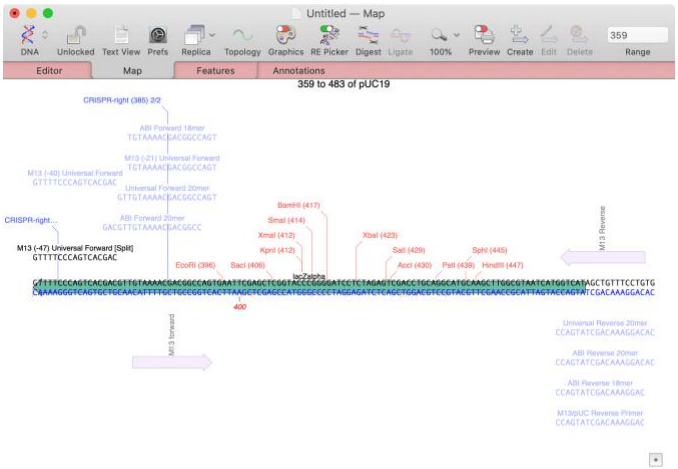
You can see many binding sites for the universal primers. Note how they are again "greyed out" compared to the permanent features to indicate that they are transient "missing" primer binding sites.

Right-click (or <ctrl>-click) on one of the primers. A popup menu appears and one option is to **Create primer_bind feature**. Select that.



A new primer_bind feature appears in normal boldness, taking on the default appearance for primer_bind features (a salmon pink hollow arrow in this case).

Click on any one of the primers above the sequence, hold down the <shift> key, then select one of the reverse primers below the sequence. Choose **Edit | Copy**. Next choose **File | New from Clipboard**.



A new window opens containing the predicted product resulting from PCR amplification using the two "facing" primers. If one or both primers had mismatched residues (e.g. from a mutagenesis experiments), they would be included in the product, as would any 5' tails added to the primers in the database.

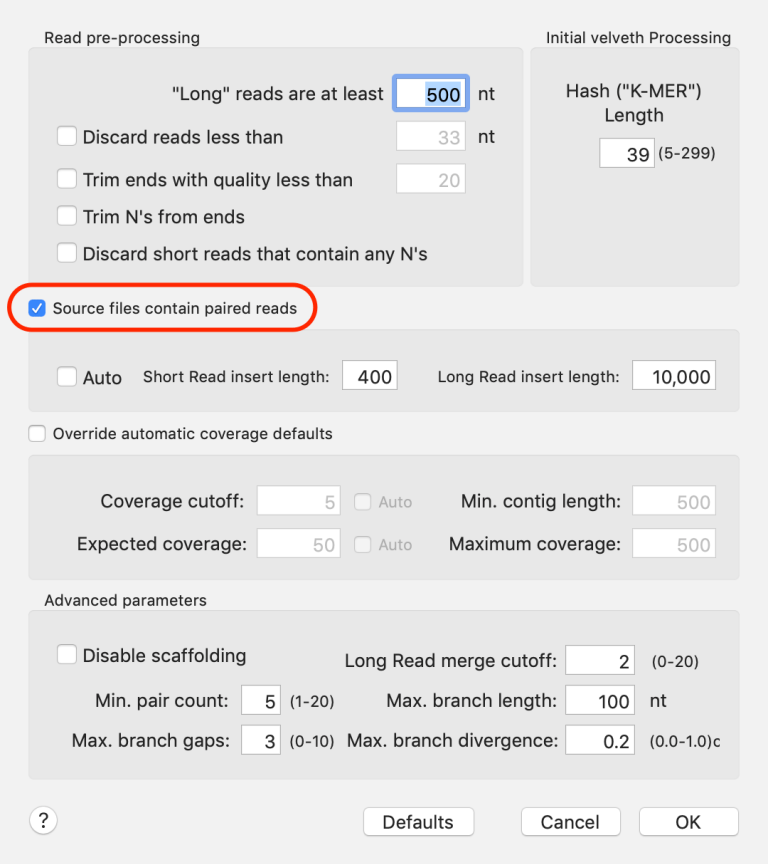
MacVector with Assembler – Job Objects

With MacVector 17, each time you run an analysis "job" in the *Assembly Project* window, the results of that job are placed in a "job object" – this is simply a "folder" in the project window containing the results of the job. Let's take a look (this requires that you have a license for *MacVector with Assembler*);

Choose **File | New | Assembly Project**, then click on the **Add Reads** toolbar button. Navigate to `/Applications/MacVector/Tutorial Files/ContigAssembly/phiX174/Fastq Data/` and select both files in that location (`phiX174-R1.fastq.gz` and `phiX174-R2.fastq.gz`).

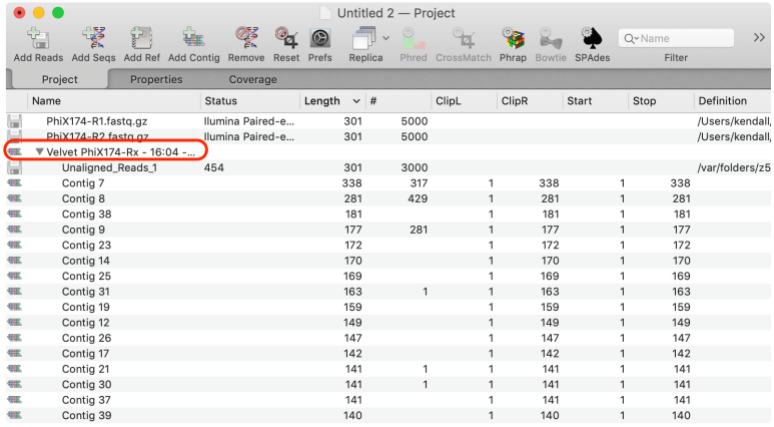
These are a pair of "gzipped" fastq files containing paired-end data from an Illumina MiSeq NGS run. Note that there is no need to unzip these types of files, which can save you enormous amounts of disk space. phiX174 is a circular 5,386bp phage molecule frequently added to Illumina sequencing runs as an internal control to confirm that the reaction proceeded as expected. The sample set included with MacVector is a small subset from such a run, chosen because of the small size on disk and speed of analysis to assemble such a short molecule.

Select the two data files in the *Assembly Project* window, then click on the **Velvet** toolbar item. Click on the **Defaults** button to use the default parameters, but then make sure you check the **Source files contain paired reads** checkbox.



Finally, click on the **OK** button.

This is a very small data set, so *Velvet* completes relatively quickly;

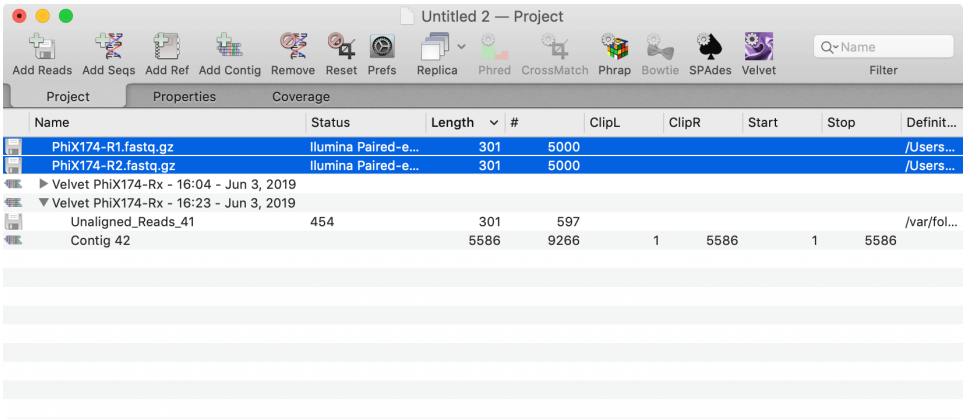


A new job object is displayed. Normally, it is automatically opened so that you can see the contents, as shown above. Note that the contigs generated are extremely small – phiX174 is 5,386bp in length and none of the contigs come anywhere near to this. This can happen with initial assembly attempts with Velvet, especially using the default settings. The most important parameter is the **Hash ("KMER") Length** value.

Note that the read files we originally imported remain at the root of the project. These are considered "read-only" copies of the data – in fact, they are not actually imported into the project at all, the project just retains "pointers" to the original data files on disk.

Close the triangle next to the *Velvet* job object. Select the two data files again. Repeat the analysis but set the **Hash ("KMER") Length** value to 201.

In general, a good place to start with *Velvet* assemblies is at 2/3rds of the average length of the input reads. In this case, the reads are around 300nt each, so 201 (values should be odd, though internally they will be rounded up if you choose an even value) should be a fair starting point.

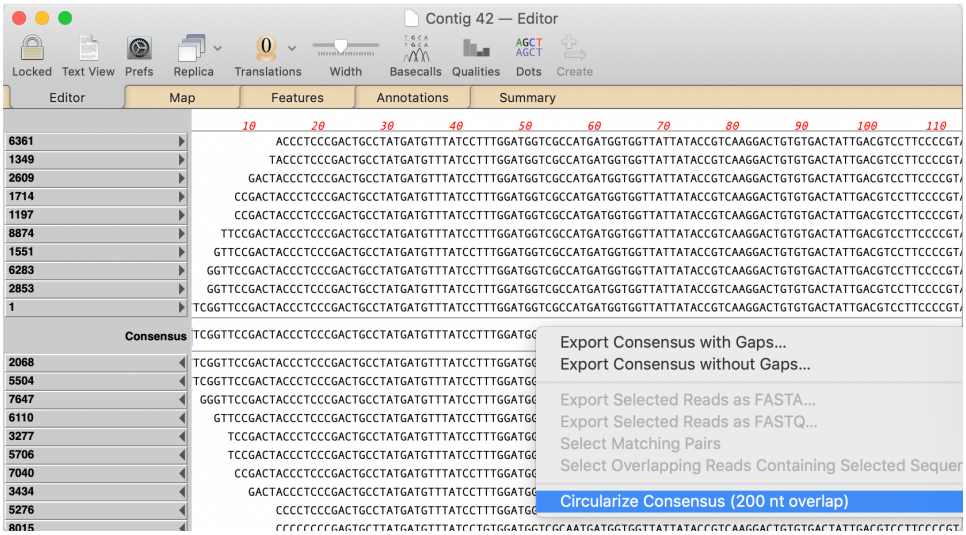


Now we see a second job object. In this case, the job has just two sub-items: a single *Contig* and an "Unaligned Reads" item.

The contig (5,586nt) is longer than the known length of phiX174 (5,386nt). In common with most assemblers, *Velvet* does not automatically identify circular molecules. However, MacVector has a solution!

Double-click on the single contig (*Contig 42* in the above image).

A *Contig Editor* window opens;



This shows the alignments of the input reads to the consensus. There is a fair amount of functionality in this editor than can be accessed using a right-click context-sensitive menu item.

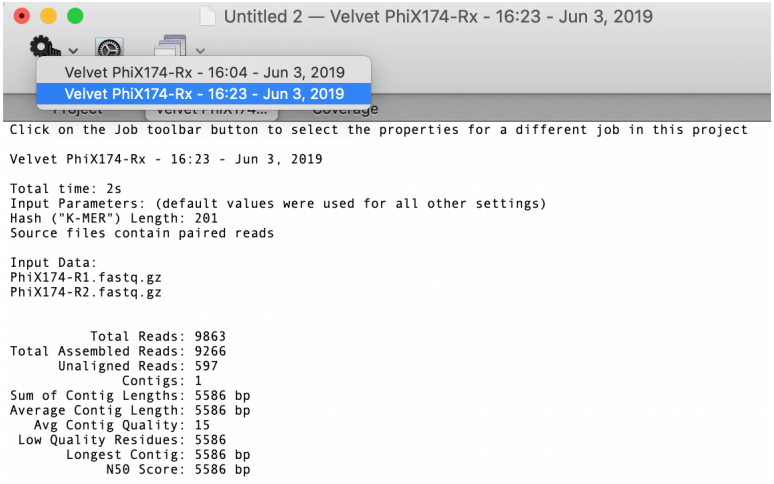
Right-click (or <ctrl>-click) to bring up the context sensitive menu.

In the absence of any selection in the window, the active menu items are to either export the consensus (with or without gaps) or, in this case, to **Circularize Consensus**. This item is only active if direct repeats have been detected at the ends of the consensus – if that is the case, the length of the overlap is reported, otherwise the menu item will be disabled and read *Cannot Circularize Consensus*.

Select the **Circularize Consensus (200 nt overlap)** menu item.

A new window appears with the circularized consensus.

Close all of the windows except for the *Assembly Project* window. Click on the **Properties** tab. This is the middle tab of the three and gets renamed to reflect the name of the currently select job object.



This tab displays the properties of the selected job, including the parameters used to generate the results, where they differ from the defaults. You can click on the upper left **Job** button to select a different job. You can also click on the **Replica** button to open a second window set to the **Properties** tab and then, each time you click on a job, the second window will update with the appropriate details.

MacVector with Assembler – SPAdes

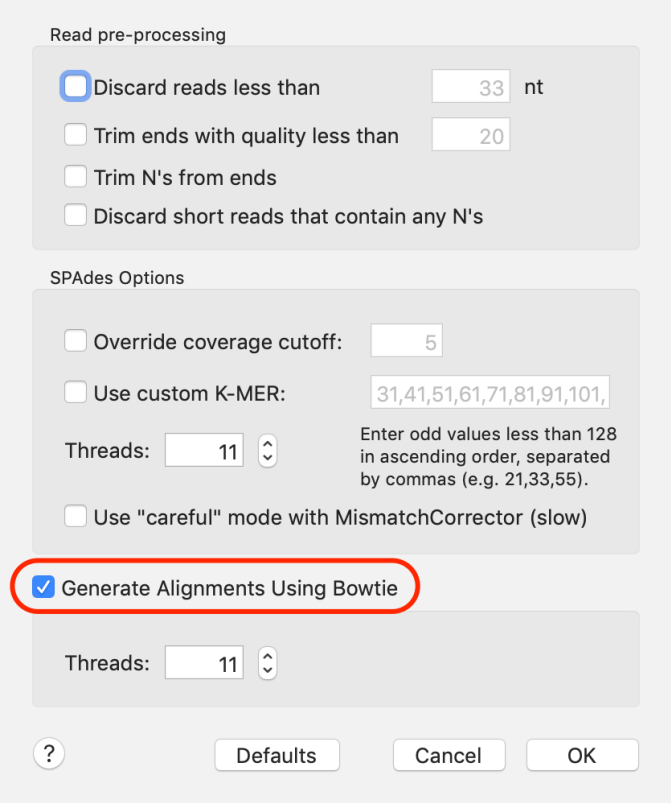
MacVector has used the popular *Velvet* short read assembly algorithm for a number of years. MacVector 16 introduced a new algorithm, *SPAdes* that has a number of advantages over *Velvet*;

- It generally requires less tweaking of parameters to get an optimum assembly
- It often generates longer contigs as it is a little better at resolving repeat sequences.
- It generally uses less memory (RAM) than *Velvet*, though that does depend on the input data.
- It can handle mixed input of short and long (e.g. Oxford Nanopore or PacBio along with Illumina/IonTorrent) reads.

On the other hand;

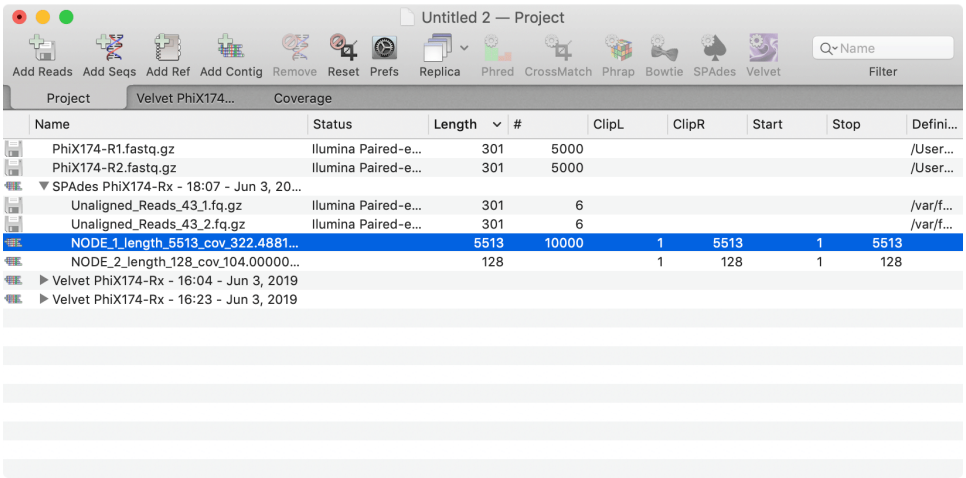
- It is significantly slower than *Velvet*. Typically, assemblies take 5-10 times longer to complete.
- It requires a greater depth of coverage
- It has a slightly greater frequency of mis-assemblies.

Hopefully, you've still got the phiX174 project open from the last section. Select the two `.fastq.gz` data files and click on the **SPAdes** toolbar item. Click on **Defaults** (if it is active indicating the settings have been modified), then make sure the **Generate Alignments Using Bowtie** checkbox is select and click on **OK**.



Unlike *Velvet*, the *SPAdes* algorithm does not generate alignments, but just consensus sequences. There are many times where seeing the actual alignments can be extremely helpful. So MacVector gives the option of running a post-assembly alignment using *Bowtie*. This will take each consensus sequence resulting from the *SPAdes* alignment and align it to the input reads. If you really don't care about viewing the alignments, leave this unchecked as it will increase the processing time by 25-50%. But for short assemblies like this, we should definitely turn it on.

Once the job completes, we get a *SPAdes xxx* job object;

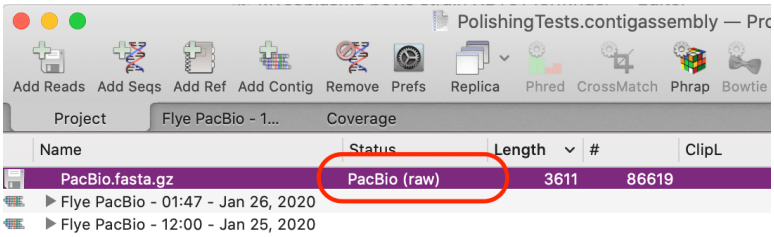


In this case, even using the defaults we get a full-length assembly. If you double-click on the longest "NODE_xxx" (that's a *SPAdes*-specific nomenclature) you can open a *Contig Editor* window where you can circularize the consensus via a right-click exactly as we saw with the *Velvet* assembly.

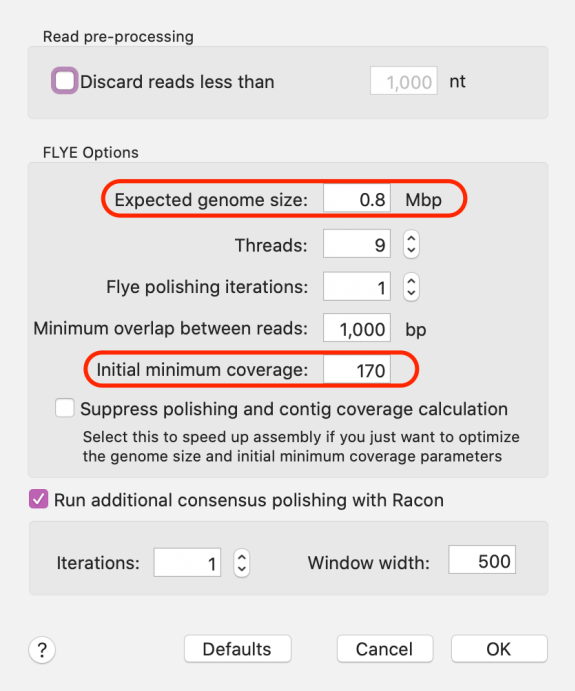
MacVector with Assembler – Flye

Pacific Biosciences and Oxford Nanopore Technologies are two companies that have pioneered single molecules sequencing techniques that can generate much longer reads than the Illumina and IonTorrent technologies. However, they also have significantly higher error rates (typically 10-15% or more) which causes significant assembly problems for typical short read assemblers. *Flye* is one of a new breed of assemblers that can assemble these high error rate long reads relatively quickly. It additionally has the ability to "polish" the consensus sequences of contigs – a procedure where the reads are re-aligned with the consensus to generate a more optimal consensus. This can be repeated for several iterations. In addition, MacVector includes a stand-alone polisher called Racon that can also improve the consensus sequences generated.

As with *SPAdes*, *Bowtie* and *Velvet*, *Flye* can directly use gzipped fasta or fastq files, saving disk space. One important difference with *Flye* compared to short read assemblers is that you MUST tell it what type of data is present in the input file(s) by double-clicking on the **Status** column entry and setting the **Source of data** appropriately.



The most important *Flye* parameters are **Expected genome size** and **Initial minimum coverage**.

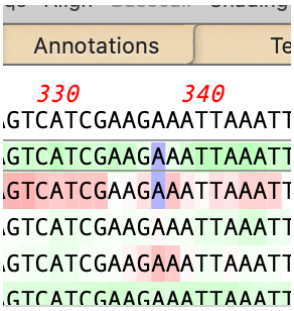


Sometimes it can take some perseverance to find the optimal **Initial minimum coverage**. To help speed up assemblies, you can temporarily select **Suppress polishing and contig coverage calculation**. This lets you assemble small bacterial genomes in just a few minutes. Once you find the best **Initial minimum coverage**, you can turn everything back on again for more accurate consensus calculations.

Align to Reference – Quality Values

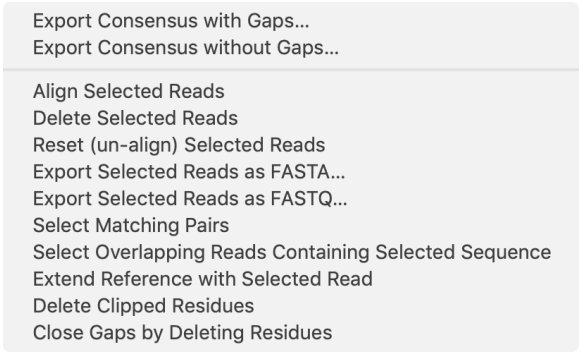
A new **Shading** button in the *Align to Reference* **Editor** window. When selected it turns on background shading for the residues in the upper pane.

As ever, you can mouse-over residues to see a tooltip displaying the details of each residue. This functionality has also been added to the *Contig* **Editor**. Edited residues are shown with a blue background;



In addition, if you have the MacVector plus Assembler module, you can now directly run the popular basecaller *phred* by clicking on the **Basecalls** toolbar button.

There have been a number of editing enhancements. In particular the right-click (`<ctrl>`-click) context-sensitive menu has additional functions;



While most of these are self-explanatory, some benefit from additional discussion. Many of these are also present in the *Contig* **Editor** window.

Select Matching Pairs – if you have aligned paired-end fastq NGS data, this will also select the opposite read of any reads you have selected. Combined with **Export Selected Reads as FASTA/FASTQ**, this lets you easily pull out related reads representing specific SNPs or repeats.

Select Overlapping Reads Containing Selected Sequence – if you select a few residues containing a SNP or other sequencing difference, this will select all the other reads containing that same variation(s).

Extend Reference with Selected Read – if a read overhangs either end of the reference, use this to extend the reference with the read. Great for extending contigs to generate overlaps and close sequencing gaps.

Delete Clipped Residues – this permanently removes all the greyed out "clipped" residues in the alignment. While those residues do not get included in consensus calculations, many users prefer the cleaned up display.

Close Gaps by Deleting Residues – it is very common for reads to have additional insertions of one or two residues due to sequencing or base-calling errors. Again, these do not typically affect the consensus calculation, but you can use this menu item to clean up alignments.

There have been a few other editing enhancements included in the last few releases;

- Hold down the `<option>` key and type a character or a gap to insert a residue or gap immediately before the currently selected base.
- You can "nudge" entire reads left or right my selecting the sequence in the left hand name panel and using the left/right arrow keys.

Align to Reference – Problems Tab

MacVector 17 added a new tab called **Problems** to the *Align to Reference* window. The idea behind this window output is to alert you to potential sequencing problems where the consensus you have generated (from MacVector or from external assemblers) might not completely match with the NGS read data you have. While there are limits to the number of reads and length of reference that MacVector can handle (mostly memory related), you can use *Align to Reference* to align 10+ million reads to a typical 5 Mbp+ bacterial genome if you are patient.

After running an *Align to Reference* alignment, the **Problems** tab will list the top 2,000 locations that exhibit differences versus the reference. The

algorithm checks every individual read against the reference and counts up the mismatches, gaps and masked/clipped regions where the reads disagree with the reference. This can help you focus on areas where the reads might indicate that the original assembly consensus is incorrect.

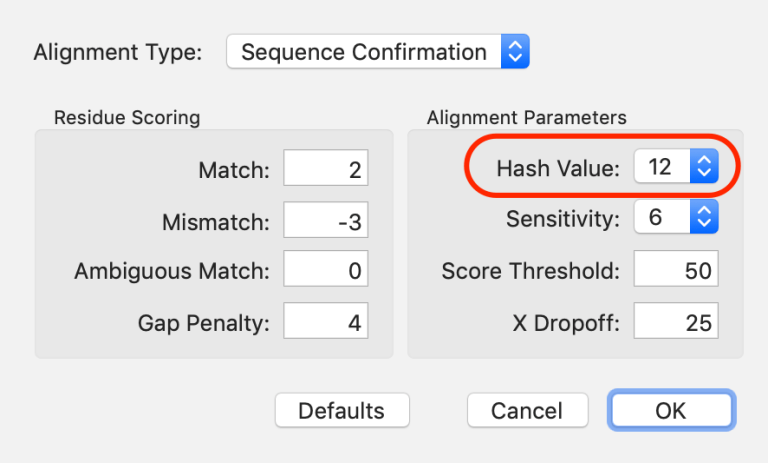
Open the file `/Applications/MacVector/Tutorial Files/Contig Assembly/phiX174/phiX174(a).nucl`.

There are a number of variants of phiX174 available. This is the one that matches our data set.

Choose **Analyze | Align to Reference**, then click on the **Add Seqs** toolbar item. Locate the folder `/Applications/MacVector/Tutorial Files/Contig Assembly/phiX174/Fastq Data/` and select the two `.fasta.gz` files in the folder and click **OK**.

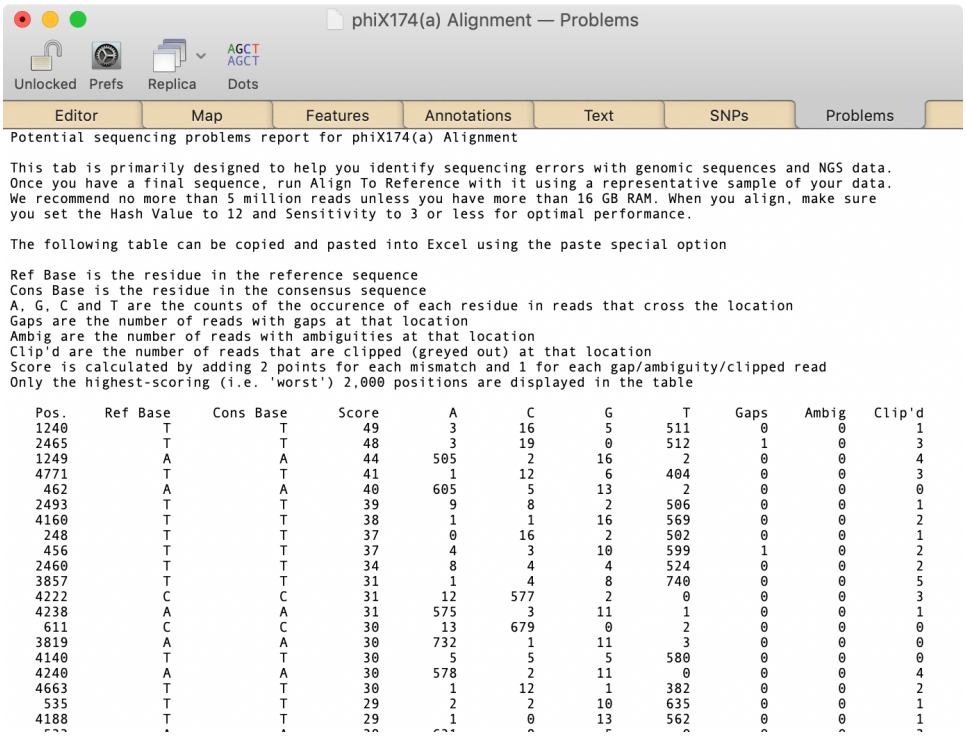
After a short pause, the display updates to show all of the imported sequences. They are shown in italics to indicate that they have not yet been aligned.

No need to select them all. Just click on the **Align** toolbar item and set up the parameters as below. A **Hash Value** of 12 really helps to speed up alignments. Click **OK**.



The alignment takes very little time with this small number of reads.

Switch to the **Problems** tab.



In this case, the problems are minimal. The "worst" position at 1240 has a score of 49, but there are 511 T's at that position compared to just 3 A's, 16 C's, 5 G's and just one clipped residue. Let's see what happens when we have real mismatches;

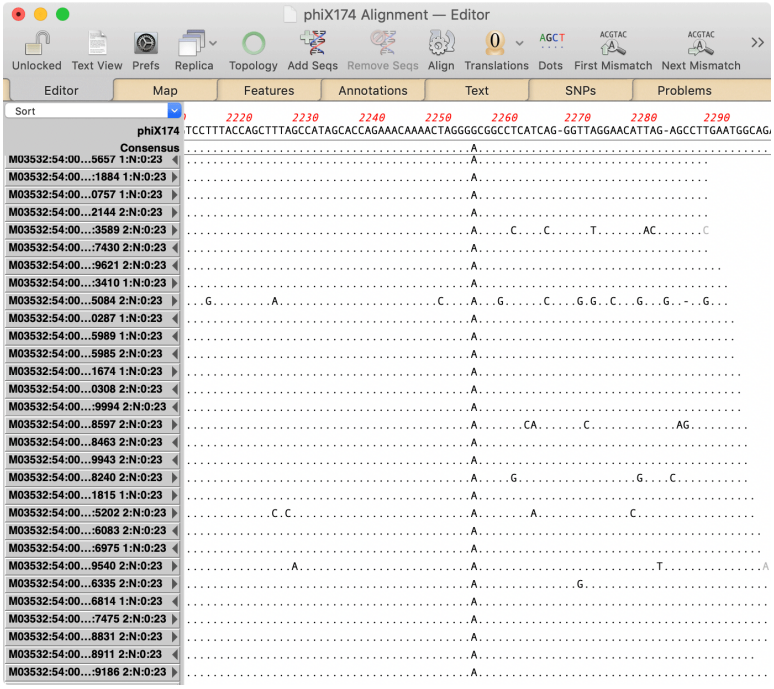
Repeat the analysis using `/Applications/MacVector/Tutorial Files/Contig Assembly/phiX174/phiX174.nucl`

This classic reference version of phiX174 has a few base changes relative to the version sequenced in this data set.

Pos.	Ref Base	Cons Base	Score	A	C	G	T	Gaps	Ambig	Clip'd
2254	G	A	1270	631	0	1	0	0	0	8
2656	T	C	1089	2	539	0	1	0	0	7
2576	G	A	1006	499	1	3	0	0	0	6
4800	C	T	881	0	1	1	436	1	0	6
4554	C	T	880	0	3	1	435	0	0	8
1240	T	T	49	3	16	5	511	0	0	1
2465	T	T	48	3	19	0	512	1	0	3
1249	A	A	44	505	2	16	2	0	0	4
4771	T	T	41	1	12	6	404	0	0	3

Here we can clearly see that the first 5 lines have very high scores, the reference and consensus bases differ and the counts of the A/C/G/T residues indicate an issue.

Scroll the **Editor** pane horizontally to around 2254, then scroll vertically until you see sequences. You can also just click on the reference sequence to have the display automatically scroll to show the reads at that location. Click on the **Dots** toolbar item to make the differences more noticeable.



Here you can immediately see the difference between the sequenced molecule and the reference. This is a real variation between the reference `phiX174` sequence and the version used to spike the sequencing reaction. You can edit the reference to match the reads and the **Problems** tab will update in real time. Note that with genome-size sequences, this can take a few seconds.