MacVector 17.5

for Mac OS X

Workshop: What's New in MacVector?

Maclector Inc.

Software for Scientists

Copyright statement

Copyright MacVector, Inc, 2020. All rights reserved.

This document contains proprietary information of **MacVector**, **Inc** and its licensors. It is their exclusive property. It may not be reproduced or transmitted, in whole or in part, without written agreement from **MacVector**, **Inc**.

The software described in this document is furnished under a license agreement, a copy of which is packaged with the software. The software may not be used or copied except as provided in the license agreement.

MacVector, Inc reserves the right to make changes, without notice, both to this publication and to the product it describes. Information concerning products not manufactured or distributed by **MacVector, Inc** is provided without warranty or representation of any kind, and **MacVector, Inc** will not be liable for any damages.

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"

	General	Q Search
Appearance:	Light Dark	
Accent color:		
Highlight color:	Blue 🗘	3
Sidebar icon size:	Medium	
	Automatically hide and sho	w the menu bar

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.

•••			🖹 Se	equenceS	ample.nucl	— Editor			
			-	\sim -				241	- 😴 - 1
DNA L	Inlocked Tex	t View Prefs	Replica T	opology (Blocking Vo	bice Verify Di	splay Create	e Range	Features
Edito	r 🦵	Мар	Featu	res	Annotations				
+ GAAAAAAAAAC CTTTTTTTG	AAAGCACATG TTTCGTGTAC	ACACTTATAA TGTGAATATT	TTATGCAAAG AATACGTTTC	AATCAACTT TTAGTTGAA	G TATTAATTG C ATAATTAAC	G CGATAAAAAT C GCTATTTTA	AAAAGTTATA TTTTCAATAT	CCTCTGTAAA GGAGACATTT	100 AGATTTAAAT TCTAAATTTA
CAAAACGACA GTTTTGCTGT	AACTCGCACT TTGAGCGTGA	AGGACAAATA TCCTGTTTAT	AAAACAGTGC TTTTGTCACG	CTGCTGGTA GACGACCAT	A ATATGCAAA T TATACGTTT	A CAATATTTAG T GTTATAAATC	ATGATCAACA TACTAGTTGT	TTTATATGGT	200 GATGTTAAAG CTACAATTTC
ATAAAATTAT TATTTTAATA	TTTTGCTAAA AAAACGATTT	GACGTAAAAC CTGCATTTTG	AAGTATTAAA TTCATAATTT	TTATGTGGA	A AAAGGAAAT T TTTCCTTTA	G CACAGGAAGG C GTGTCCTTCC	ATTCGTTTAC TAAGCAAATG	AAAACAGACT TTTTGTCTGA	300 TATATCAACA ATATAGTTGT
AAAGAAAAAA TTTCTTTTT	GCAAATAAAG CGTTTATTTC	TTAAAGTCAT AATTTCAGTA	CGAAGAAATT GCTTCTTTAA	AAATTAAGC	A AACCAATTA T TTGGTTAAT	C TTATAAAGCA G AATATTTCGT	GGTGCAACTT CCACGTTGAA	CAGATAAAAA GTCTATTTTT	400 ATTAGCTAAA TAATCGATTT
GAATGGATTA CTTACCTAAT	ACTTCTTAAA TGAAGAATTT	ATCTAATAAA TAGATTATTT	GCTAAACAAA CGATTTGTTT	TTCTTAAAG AAGAATTTC	A ATATCAATT T TATAGTTAA	T TCTGTATAAG A AGACATATTC	GAGTTCAGAG CTCAAGTCTC	CTATGCCTGA GATACGGACT	500 TTTAACGTCC AAATTGCAGG
TTTTGGATTT AAAACCTAAA	CTTTTCGTGT GAAAAGCACA	TGCTTTAATC ACGAAATTAG	AGTACAATGA TCATGTTACT	TAGTTACTA ATCAATGAT	T TTTTGGCAT A AAAACCGTA	T TTGATTTCTA A AACTAAAGAT	AATGGCTATA TTACCGATAT	CAATAAAAAA GTTATTTTTT	600 AGATATTGGG TCTATAACCC
TAAATCTATT ATTTAGATAA	AGAAAGTTTT TCTTTCAAAA	ATCATTTTAC TAGTAAAATG	CAATTGTGTT GTTAACACAA	ACCACCTAC TGGTGGATG	T GTCCTTGGT A CAGGAACCA	T TTATACTATT A AATATGATAA	ΑΑΤΤΑΤΑΤΤΤ ΤΤΑΑΤΑΤΑΑΑ	TCAACAAGAA AGTTGTTCTT	700 GTCCTGTAGG CAGGACATCC
	ΔΟΤΔΑΤΑΤΟΤ	ТАСАСТТАСС	ΔΩΤΤΩΤΑΤΤΤ				ΔΤΤΩΤΤΔΩΤΤ	ттессеттат	600 CTATCAACAT

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;

00		RI	E Picker			
Filter			File			
Cuts: >= 1 and <= 8			Comm	on Enzymes.re	enz ≎	
End struc	Reset Filter	s	Save c	urrent set of Er	nzymes	
?		Select All	Deselect All		Defaults	
Acc65I	🗸 Accl	Acil	Acli	AfIII	AfIII	
🗹 ApaLl	ApeKI	Apol	Asel	att*2	att*3	
Aval	Avall	Banl	BbvCl	🗹 Bcll	🗹 Bcll-DAM	
🗌 Bfal	🗹 Bglll	Blpl	Bme1580I	BmgBl	Bpu10I	
🗹 BsaAl	🗹 BsaBl	BsaJI	BsiHKAI	Bsml	BsoBl	
Bsp1286I	BsrBI	BsrFl	BsrGl	Bsrl	BssKI	
BssKI-DCM	BstAPI	BstBl	BstEll	BstNI	BstUI	
BstYI	BstZ17I	CRISPR-I	eft 🗌 Dpnl	Dpnll	Dpnll-DAM	
🔽 Dral	Eael	EcoRV	🗌 Fnu4HI	🗸 Fspl	Haell	
99 Enzymes match criteria with 19 selected, 239 locations match these enzymes.						

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

Sample.nucl
Click on the RE Picker button on the Map tab toolbar.
The <i>RE Picker</i> window closes and the icon changes to indicate it is hidden;
Click on the RE Picker button again to show the <i>RE Picker</i> window.
The <i>RE Picker</i> 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 <i>RE Picker</i> and the Map tab update to reflect the changes. The Map tab always shows only those enzymes that are both visible in the <i>RE Picker</i> and that are selected.
Click on the Defaults button. This resets the <i>RE Picker</i> to its initial default settings. Now click the checkbox next to the <i>Bgl</i> 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 <i>RE Picker</i> 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 Noncutters.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

•••	D M	/P_125627883.1	(WP_12562788	3) Alignments	— Editor
Protein Unlocked	Add Seqs	Align Phylogeny	ACGT A-GT ACGT Consensus Pref	s Replica	AGCT AGCT Blocking Dots
Editor	Text	Pairwise	Matrix	Picture	Guide Tree
WP_12262012262068	85) ECLQG	P D D A R G E R L		VVLTTAK	PGVL RRAF
WP_15317615317689	90} ECLTG	<mark>P D</mark> G G <mark>R</mark> A A L L	R A R Y G D R L		PGVLRRAF
WD 061441 0614410					
WP_06144106144128					
consens		PDDGRAALLL			
	110	120		130	140
WP_12562712562788	33} RLITA	V H T V A K G E R	F L D E T L T V	ALLKGAEM	1 P L T T R E L G
WP_13388513388530	06} RLITA	V H T V A K G E R	FLDETLTV	ALLEGAAM	1 <mark>P</mark> L T T <mark>R E</mark> L G
WP_12262012262068	85} RLITA	V H T V A K G E R	FLDETLTV	ALLKGAEM	1 <mark>P</mark> LTTRELG
WP_15317615317689	0) RLITA	V H T V A K G E R	FLDETLTV	ALLKGAEM	1 P L T T R E L G
WP_06144106144128	81) RLITA	VHTVAKGER	FLDETLTV	A L L <mark>K</mark> G A Q M	IPLTTRELG
consens			FLDETLTV		

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.

Gibson/Ligase-independent Assembly Type
What is the primary aim of this project:
I want MacVector to help me design primers for my Gibson Assembly or NEBuilder cloning project. Choose this option if using 5' exonuclease to cut back fragment ends.
I want to design a Ligase-Independent Cloning experiment. Choose this option if using T4 polymerase or other enzymes with 3' exonuclease activity to prepare fragment ends.
\bigcirc I have already created fragments with overlapping ends and I want MacVector to assemble them.
The approaches are not mutually exclusive. One may change the default mode using the Preferences button at any time. One can mix the approaches by clicking on the Customize button next to fragments in the cloning project.
? Cancel Create

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 *Smal* 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;



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;

CCCCATGCCTGATTTAACGTCC-3' ATGCCTGATTTAACGTCCTT TACGGACTAAATTGCAGGAA

...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;

.. CGTTTAGGGAGGTTGATTAA .. GCAAATCCCTCCAACTAATT '-GCAAATCCCTCCAACTAATTccc

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.

	Left Junction	Right Junction					
	(20nt overlap w	ith pUC19R Smal)					
pUC19R Smal			ORF 1				
ORF 1-fwd (Tm = 52.3°C)							
5'-gtcgactctagaggatccccATGCCTGATTTAACGTCC-3' (18nt binding)							
<pre>erLeuHisAlaCysArgSerThrLeuGluAspProHisAla***</pre>							
GCTTGCATGCCTGCAGgtcgactctagaggatccccATGCCTGATTTAACGTCCTTTTGGATTTCTTTTC							
CGAACGTACGGACGTCCAG	CTGAGATCTCCTAGG	GGTACGGACTAAATTGCAGGAAA	ACCTAAAGAAAAG				

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 *lacZ* 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;

ORF 1-fwd (Tm = 52.3°C) 5'-gtcgactctagaggatccccGATGCCTGATTTAACGTCC JysArgSerThrLeuGluAspProArgCysLeuIle*** IGCAGgtcgactctagaggatccccGATGCCTGATTTAACGTCC ACGTCCAGCTGAGATCTCCTAGGGGCTACGGACTAAATTGCAGG

• Automatically genera	te primer	
🗸 Include Spacer	G	
O Use custom primer		<
O No primer (synthetic	fragment or existing PCR product)	

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).

```
ORF 1-fwd (Tm = 52.3°C)
5'-gtcgactctagaggatccccGATGCCTGATTTAACGTCC
JysArgSerThrLeuGluAspProArgCysLeuIle***
FGCAGgtcgactctagaggatccccGATGCCTGATTTAACGTCC
ACGTCCAGCTGAGATCTCCTAGGGGCTACGGACTAAATTGCAGG
```

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.

```
ORF 1
```

```
ORF 1-fwd (Tm = 52.3°C)
```

```
5'-gtcgactctagaggatccccGGATGCCTGATTTAACGTCC-3' (18nt binding)

CysArgSerThrLeuGluAspProArgMetProAspLeuThrSerPheTrpIleSerPheArgV...

FGCAGgtcgactctagaggatccccGGATGCCTGATTTAACGTCCTTTTGGATTTCTTTCGTG...

ACGTCCAGCTGAGATCTCCTAGGGGCCTACGGACTAAATTGCAGGAAAACCTAAAGAAAAGCAC...
```

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

Click on	the Prir	ners ta	ab.						
				Untitled	— Primers				
🕤 🗘 🕤		- - -		;				Q~ Sea	arch
Gibson Un	locked Prefs	Replica	Add to DB	Save List					Filter
Fragment	s Pri	mers							
Name	Oligo (lower	case = tail)		Overlaps	Anneals	Strand	Tm	Та
ORF 1-fwd	gtcgactctag	aggatcccc	GGATGCCT	GATTTAACGTCC	pUC19R Smal	ORF 1	Forward	52.3°C	51.5°C
ORF 1-rev	tgaattcgagc	tcggtaccc	ГТААТСААС	CTCCCTAAACG	pUC19R Smal	ORF 1	Reverse	52.6°C	51.6°C

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 Ligaseindependent 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;

Check the Orientation of a Ligated Fragment Determine RE Sites for Cloning Hide the Graphics and RE Picker Floating Windows Design Gibson Cloning/LIC Strategies Subclone Digested Fragments into a Vector Bring Sequences into MacVector Optimize Codon Usage Generate a Transcript of a DNA Sequence Do an Online Keyword Search for Sequences Annotate a Gene to my Sequence Automatically Annotate Blank Sequences Find Functional Domains in a Protein Import Features from a Genome Browser Change the Default Appearance of a Feature Display Missing Features, Predicted ORFs and Restriction Sites Design Primers with Tails and Mismatches Add a Primer to the Primer Database Design a Primer to Match an Existing Primer Design Primers to Amplify a Gene Test a Pair of Primers Annotate Where a Primer Binds on a Sequence Align Reads Against a Reference Sequence Map NGS Data Against a Reference Genome Extract Reads from a FASTQ Dataset Finish a Genome Assembly Create a De Novo Assembly (Velvet) Create a De Novo Assembly (SPAdes) Create an Assembly Project Choose the Right Alignment Tool Identify Important Differences Between Two Genomes Visually Align a Pair of Sequences Do an Online BLAST Search

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

MacVector 17.0 Help

 MacVector 17.0 Help

 Q Search

 MacVector Help

Key How do I annotate a gene feature to my sequence

It's easy to annotate your sequence and add regions of interest. From manually annotating a single gene, to automatically annotating a blank sequence with common features.



(View full size on website...)

- 1. Open a sequence
- 2. Switch to the Map tab
- 3. Double click on a missing feature or a predicted ORF
- 4. Modify the feature in the Symbol editor
- 5. Click OK

See the How to auto annotate sequences and Scan for Missing Features help topics for easier methods of sequence annotation.

Related Topics.

How to auto annotate sequences. How do I? - videos

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.

?		Search Details To Desk To Disk
Relevance	Document ID	Title
	CP003913.2	Mycoplasma pneumoniae M129-B7, complete genome
	NC 020076.2	Mycoplasma pneumoniae M129-B7, complete genome

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.

Target Genome:		Features
FH.nucl		 ✓ CDS ✓ gene ✓ RNA Others
Options		Scan
Min Feature Length:	50	Min% Score: 95%
Match:	2	Min% Length: 70%
Mismatch:	-3	Max% Gaps: 1%
Gap Penalty:	4	Hash Value: 10 ᅌ
		Sensitivity: 4 ᅌ
?	Defaults	Cancel OK

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;

Results		
🗸 Identical		
🗸 Similar	Similarity Threshold:	98%
🗸 Weak	Weak Threshold:	85%
🗸 Missing		
🗸 Details		
🗸 Plot		
🗸 Context		
Defau	Its Cancel	ОК

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;



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	ČDS	100% (100%)	609228	609482	+
					family K-like protei	CDS	<u>97.0% (98.8%)</u>	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	ČDS	99.7% (100%)	609764	612145	+
MG032/MG096/MG288 fa	CDS	620619	622556	-	MG032/MG096/MG288 fa	CDS	<u>99.7% (100%)</u>	615461	617398	-
mambeana neatain	CDC	634031	C 3 E 7 A 3		mambrana aratain	CDC	1000 (1000)	C107CA	C 205 4C	

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;

	🕅 M129.nucl — Results
imes Identical	imes Similar $ imes$ Weak $ imes$ Missing $ imes$ Plot $ imes$ Context $ imes$ Details
M129.nucl CDS	994711275 /codon_start=1 /inference=EXISTENCE: similar to AA sequence:SwissProt:P75104.1 /locus_tagre985_0008 /note=Derived by automated computational analysis using gene prediction method: Protein Homology. /produc=TkNA modification GTPase TrmE /protein_id=ACG03952.1 /transl_table=4 /transl_table=4
FH.nucl CDS	993111259 /codon_start=1 /inference=EXISTENCE: similar to AA sequence:SwissProt:P75104.1 /locus_tag=F539_00040 /note=Derived by automated computational analysis using gene prediction method: Protein Homology. /product=tRNA modification GTPase TrmE /protein_id=ALA36182.1 /transl_table=4 /transl_table=4
Aligned Len	gth = 442 Gaps = 0
Identities	= 441 (99.8%) Similarities = 1 (0.2%)
M129.nucl	1 MDTKQTMFALATAPFNSAIHIIRLSGPDVYRIINQITNKEVKPLGMRIQR 50
FH.nucl	1 MDTKQTMFALATAPFNSAIHIIRLSGPDVYRIINQITNKEVKPLGMRIQR 50
M129.nucl	51 VWLIDHNQKKVDDVLLFKFVAPMSYTGEDLIEISCHGSMVIVNEIIGLLL 100
FH.nucl	51 VWLIDHNQKKVDDVLLFKFVAPMSYTGEDLIEISCHGSMVIVNEIIGLLL 100
M129.nucl	101 KHGAVQAQPGEFTQRGYLNGKMSLNQAASVNNLVLSPNTTLKDVALNALA 150
FH.nucl	101 KHGAVQAQPGEFTQRGYLNGKMSLNQAASVNNLVLSPNTTLKDVALNALA 150
M129.nucl FH.nucl	151 GQVDARLEPLVEKLGQLVMQMEVNLDYPEYTDEQRELVTMNQAVVQITQI 200 151 GQVDARLEPLVEKLGQLVMQMEVNLDYPEYTDEQRELVTMNQAVLQITQI 200
M129.nucl	201 LNQIVVGQDQLQRLKDPFKIAIIGNTNVGKSSLLNALLDQDKAIVSAIKG 250
FH.nucl	201 LNQIVVGQDQLQRLKDPFKIAIIGNTNVGKSSLLNALLDQDKAIVSAIKG 250
M129.nucl	251 STRDIVEGDFALNGHFVKILDTAGIRQHQSALEKAGIQKTFGAIKTANLV 300
FH.nucl	251 STRDIVEGDFALNGHFVKILDTAGIRQHQSALEKAGIQKTFGAIKTANLV 300

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.

				ž	M129.nucl — Results				
imes Identical $ imes$ Simila	r × We	eak × Mis	sing ×	Plot ×	Context				
					Match	Match	Match Score Ma	tch Match	Match
Name	Type	Start	Stop	Strand	Name	Type	Ident. (Sim.) Sta	art Stop	Strand
hypothetical protein	CDS	141611	141892	+	F539 00630	gene	89.0% 141	508 141888	+
C985 00840	gene	141611	141892	+	F539 00630	gene	89.0% 141	508 141888	+
hypothetical protein	ČDS	142331	144487	+	F539 00635	gene	90.7% 142	327 144482	+
C985 00850	gene	142331	144487	+	F539 00635	gene	90.7% 142	327 144482	+
hypothetical protein	CDS	199788	202172	+	E539 00875	gene	97.6% 198	200510	+
house and the second					hypothetical protein	CDS	97.6% (95.1%) 198	200510	+
C985 0151	gene	199788	202172	+	E539 00875	gene	97 6% 198	200510	+
C985 00930	gene	245656	246987	+	E539 01130	gene	96.2% 243	94 245323	+
hypothetical protein	CDS	247368	247619	+	F539 01140	gene	94.8% 245	245954	+
					hypothetical protein	CDS	94 8% (88 9%) 245	704 245954	+
C985_0202	gene	247368	247619	+	F539_01140	gene	94.8% 245	245954	+
C302 00322	gene	323433	524002	-	<u>L223_01432</u>	gene	90.1% 321	000 022400	-
adhesin	CDS	341758	343155	+	F539_01585	gene	<u>97.8%</u> 340	929 341425	+
<u>C903_00903</u>	gene	341730	343133		<u>1333_01303</u>	gene	<u>97.0%</u> 340	341423	
<u>tRNA (adenine-N1)-me</u>	CDS	418178	418819	-	F539_01965	gene	<u>97.8%</u> 416	358 416999	-
					tRNA (adenine-N1)-me	CDS	<u>97.8% (96.8%)</u> 416	358 416999	-
C985 0356	gene	418178	418819	-	F539 01965	gene	<u>97.8%</u> 416	358 416999	-
hypothetical protein	ČDS	433609	435642	+	F539 02040	gene	91.6% 431	790 433823	+
					hypothetical protein	ČDS	91.6% (85.1%) 431	790 433823	+
C985 01055	gene	433609	435642	+	F539 02040	gene	91.6% 431	790 433823	+
C985 0404	gene	482347	482420	-	F539 02240	gene	96.6% 480	480659	-
tRNA-G1v	tRNA	482347	482420	-	tRNA-G1v	tRNA	96.6% 480	586 480659	-
hypothetical protein	CDS	487570	490209	-	hypothetical protein	CDS	83.1% (88.7%) 485	308 488447	-
C985 0421	gene	497609	498081	+	(unannotated)	(none)	96.4% 495	348 496321	+
hypothetical protein	CDS	532005	532718	-	E539 02460	gene	91.6% 530	47 530960	-
in the second second second					hypothetical protein	CDS	91.6% (86.0%) 530	47 530960	-
C985 01145	gene	532005	532718		E539 02460	gene	91 6% 530	247 530960	-
C985_0464	gene	562114	562198	-	F539 02565	gene	97 0% 556	43 557027	-
	Bene	30/114	30/130		INNO-LEU	Bene	2/ 86 330	140 00/02/	_
adhesin	CDS	570841	571464	+	E539 02640	gene	97.3% 565	572 566296	+
Maria De la	000	570012	572101		adhesin	CDS	97.3% (86.3%) 565	572 566296	÷,
C382 01130	gene	570841	571464	+	1539 02640	gene	97.3% 565	0/2 566296	+
hypothetical protein	ČDS	608225	608815	+	hypothetical protein	CDS	92,5% (96,5%) 603	603618	+
C985 0509	gene	608225	608815	+	F539 02815	gene	92.5% 603	038 603618	+
C985 01220	gene	610605	611077	+	(unannotated)	(none)	96 4% 495	348 496321	+
hypothetical protein	CDS	703854	705173	-	E539 03285	gene	97 3% 698	576 699996	-
in procession processi		, , , , , , , , , , , , , , , , , , , ,			hypothetical protein	CDS	97 3% (95 8%) 698	576 699996	
C985 0591	gene	703854	705173		F539 A3285	gene	97 3% 698	76 699996	
hypothetical protein	CDS	766452	767354		hypothetical protein	CDS	91 7% (94 3%) 767	191 768094	
insportierical proteill	CD 3	/00452	, 0, 5, 54	-	nypothetical protein	CD3	JA. (M (JH. JM) /0/.		-

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.

		N MADO	nuel Des	l.k	
		₹ M129.I	nuci — Res	Suits	
V Identical V Cimilar	V Week	V Missing	V Diet	V Contout	
 Identical < Similar 	~ Weak	< wissing	~ PI01	< Context	
Name	Type Start	. Stop	Strand		
type I restriction m	CDS 111477	112592	+		
985 0091	gene 1114//	117597			
adhesin	CDS 117274	117915			
1985 00/90	gene 11/2/4	120005			
adnesth	CDS 128052	129095			
hypothetical protein	CDS 134031	134549	+		
C985 0103	gene 134031	134549	_		
hypothetical protein	CDS 164453	165226	+		
C985 0127	gene 164453	165226	+		
hypothetical protein	ČDS 168754	169008	+		
<u>C985_0130</u>	gene 168754	169008	+		
hypothetical protein	CDS 169008	169430	+		
<u>C985_0131</u>	gene 169008	169430	+		
hypothetical protein	CDS 177423	178109	-		
<u>C985_0138</u>	gene 177423	178109	-		
COSE 0130	CDS 1/8358	170050	-		
adhesin	CDS 18082/	185707	+		
L985 0147	gene 1808/4	185/0/	÷		
Mgp-operon protein 3	CDS 185713	189369	+		
C985 0143	gene 185713	189369	+		
C985_00905	gene 190239	191851	+		
hypothetical protein	CDS 248526	249842	+		
2905_00940	gene 240520	243042			
type I restriction m	CDS 347291	. 347842	+		
proling rich R65 pro	CDS 264241	265559	+		
C985 0314	Gene 364341	365558	+		
hypothetical protein	CDS 409677	410925	+		
CORE 0240	gono 400672	410025			
restriction endonucl	CDS 435680	436792	+		
<u>C985_0372</u>	gene 435686	436/92	+		
MgpC-like protein	CDS 498372	499613	+		
<u>C985_01125</u>	gene 498372	499613	+		
hypothetical protein	CDS 537818	541795	-		
<u>C985_0447</u>	gene 537818	541795	-		
1985_0462	gene 558691	558779	+		
C085 01165	TKNA 558691	558//9	+		
hypothetical protein	CDS 6456034	646100			
C985 0532	gene 645603	646109	-		
ABC transporter ATP-	CDS 692312	694294	-		
2005 0502	ECHC 692312	604204			
restriction endonucl	CDS 738302	739408	-)		
1985 0619	gene /3830/	/39408			

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;

				× 5	Seque	nceSample.i	nucl			
					Featur	e Svmbo				
						000				
			Fea	iture Keyi	vora:	CDS		~		
Locatio	on:									
Star	ts	Start Base		Stops	St	op Base	Operation		GenBank	
At	0	483		At	\$ 11	54	Continuo	us () 🗘	4831154	
								0		
		Deals Frances		C	omple	mentary Str	and	Oper	ration:	Join 🔽
Locat	tion in Gei	Bank Format	t:							
483.	.1154									
				Qu	alifiers	Free-Fo	orm			
				Qu	alifiers	Free-Fo	orm			
	quaimer			Qu	alifiers	Free-Fo	orm			
C	Qualifier /note /translat	tion	0	Qu omments 223 aa MPDI TSEW		Free-Fo		WWW	ESEIII PIVI E	PTVI
C	Quaimer /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-Fo		YWVNLL	ESFIILPIVLF	PTVL
C	Qualifier /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-Fo	ILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PTVL
C	Quaimer /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-Fo	orm IILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
C	Quaimer /note /translat	tion	0 :	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-Fo	PILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
C	Qualifier /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	ALISTMIVTIFG	DILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
	Quaimer /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	ALISTMIVTIFG	DY MILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
+-	Quaimer /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	ALISTMIVTIFG	DYM HLISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
+-	Quaimer /note /translat	tion	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	ALISTMIVTIFG	DYM HILISK WLYNKKR	YWVNLL	ESFIILPIVLF	PPTVL
+	Quaimer /note /translat	ilon	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-FC	97m	YWVNLL	ESFIILPIVLF	PPTVL
+	Quaimer /note /translat	tion ed.	0	Qu omments 223 aa MPDLTSFW	alifiers	Free-Fc	97m	YWVNLL	ESFIILPIVLF	νΡΤΥL
+	Quaimer /note /translat	ed.	0	Qu omments 223 aa MPDLTSFW	alifiers /ISFRV/	Free-Fc	n iliskwi yn kkr	YWVNLL	ESFIILPIVLF	PPTVL
+	Qualiner /note /translat	ilon	0	Qu mments 223 aa MPDLTSFW	alifier:	ALISTMIVTIFG	ILISKWLYNKKR	YWVNLL	ESFIILPIVLF	PPTVI

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.

		Scan DNA				
	🥥 A/ 🙆	1 🕅 🎿		Reported States Barrier Construction Constru		
General Aligned Vie	w Color Font Interne	et License Map Vie	w Scan DNA	Text View	Update	
Scan DNA Sequ	ences for					
Missing	g Features Open R	eading Frames	Primers	Restrictio	n Sites	
🗹 Show missi	ing features					
ia Annotat	tedFragments	≎ ⊑ s)o not scan equence is	if over	1000	КВ
	Set Search Folder	Minim	um feature	length:	10	
Allow ga	aps in CDS features	Maximum all	owed misma	atches:	1%	
		Maxin	num allowed	d gaps:	0.5%	
?		Defaul	ts	Revert	A	oply

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;

Graphics
Linear Circular
Residues per inch 0 446.2
Radius: 1.6
AF I I X
Range: 1:4361
Magnification: 100%
Edit Filter
✓ ▼Features
✓ ▼CDS
C 3293:4153 AmpR
Kesults Missing Features
86:1276 TetR
1915:2106 Rop
▼ frag
2475:3148 1938 to 2611 of
✓ regulatory

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;

•••			Scan DNA	4				
8		A/ 💿	R	-de		Normania Mariana Maria		
General Align	ed View Color	Font Internet	License Ma	p View	Scan DNA	Text View	Update	
Scan DNA	Sequences fo	r						
N	lissing Feature	es Open Re	ading Frame	s P	rimers	Restriction	n Sites	
Show	primers							
P	rimer Database.ı	nsub	٥	Do sec	not scan quence is	if over	200	КВ
C	Open Set	t Database Fil	e 🗸	Ignore	e tails on p	orimers		
O Us	se all primers	ad primers		Overr	ide misma	atches with	1 2	0
	ily use selecte							
?			De	efaults		Revert	Ар	ply

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.

🖹 🖹 🚔 🖉 🖉	1		Q~Sequence			
IA Subseq Locked Add Edit Delete	e Desel	ect	Filter			
Name	~ :	Sequence		Com		
+19bs	(GCCGCTCTAGAACTAG	TG			
-16bs	٦	CGAGGTCGACGGTAT	CG			
ABI Forward 18mer	1	IGTAAAACGACGGCCA	GT			
ABI Forward 20mer	(GACGTTGTAAAACGACGGCC				
ABI Reverse 18mer	(CAGGAAACAGCTATGACC				
ABI Reverse 20mer	(CACAGGAAACAGCTATGACC				
M13 (-21) Universal Forward	1	IGTAAAACGACGGCCA	GT			
M13 (-40) Universal Forward	(GTTTTCCCAGTCACGAC				
M13 (-47) Universal Forward	(CGCCAGGGTTTTCCCAGTCACGAC				
M13/pUC Reverse Primer	(CAGGAAACAGCTATGACC				
SP6	/	ATTTAGGTGACACTATAGAA				
SP6 Promoter	(CATACGATTTAGGTGACACTATAG				
T3 promoter	/	ATTAACCCTCACTAAAGGGA				
Т7	Ţ	TAATACGACTCACTATAGGG				
T7 Terminator	(GCTAGTTATTGCTCAG	CGG			

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.

Name: ABI Forward 20mer
Number of Parts: 1 ᅌ
Part #1 (Use lower case letters to indicate tails on primers)
Sequence: GACGTTGTAAAACGACGGCC
Perfect match (X):
Allowed mismatch: 3 Offset: 0
Comments:
Problems
Nothing has changed.
Revert Cancel OK

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.

Read pre-processing	Initial velveth Processing
"Long" reads are at least500ntDiscard reads less than33ntTrim ends with quality less than20Trim N's from ends	Hash ("K-MER") Length 39 (5-299)
✓ Source files contain paired reads	
Auto Short Read insert length: 400 Long Read	d insert length: 10,000
Override automatic coverage defaults	
Coverage cutoff: 5 Auto Min. c Expected coverage: 50 Auto Maximum	contig length: 500 um coverage: 500
Advanced parameters	
Disable scaffoldingLong Read merge ofMin. pair count:5 (1-20)Max. branch leMax. branch gaps:3 (0-10)Max. branch diverge	eutoff: 2 (0-20) ength: 100 nt gence: 0.2 (0.0-1.0)c
? Defaults	Cancel OK

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

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

	•				- L L	Jntit	led 2 -	- Pro	oject						
	ta 😵	2	÷ 🦉	a	0	P] ~		्य		Re	٩	Q~Nan	10	>>
Ad	d Reads Add Seqs	Add Ref Add	d Contig Remove	Reset	Prefs	Rep	olica	Phred	CrossMatch	Phrap	Bowtie	SPAdes		Filter	
1	Project	Properties	Covera	ge											
	Name		Status		Length	~	#		ClipL	ClipR		Start	Stop		Definition
	PhiX174-R1.fas	stq.gz	Ilumina Paire	ed-e		301		5000							/Users/kendal
	PhiX174-R2 fa	sta az	Ilumina Paire	ed-e		301		5000							/Users/kendal
THE.	Velvet PhiX174	-Rx - 16:04)												
	Unaligned_	Reads_1	454			301		3000							/var/folders/z8
48.	Contig 7					338		317	1		338		1	338	
-98.	Contig 8					281		429	1		281		1	281	
98.	Contig 38					181			1		181		1	181	
-98	Contig 9					177		281	1		177		1	177	
48.	Contig 23					172			1		172		1	172	
-98.	Contig 14					170			1		170		1	170	
480.	Contig 25					169			1		169		1	169	
E.	Contig 31					163		1	1		163		1	163	
-980	Contig 19					159			1		159		1	159	
TE	Contig 12					149			1		149		1	149	
-98.	Contig 26					147			1		147		1	147	
TE .	Contig 17					142			1		142		1	142	
·CEE.	Contig 21					141		1	1		141		1	141	
۹Ľ.	Contig 30					141		1	1		141		1	141	
-98.	Contig 37					141			1		141		1	141	
err.	Contig 29					140			1		140		1	140	

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.

•	. 😐 🖷			Untitled	2 —	Project						
	험 생 김 속	QZ (¥ 🙆	- -		ेम	8	2.	35	Q~Na	me	
Ac	d Reads Add Seqs Add Ref Add Cor	ntig Remove R	teset Prefs F	Replica	Phred	CrossMatch	Phrap B	lowtie SPAdes	Velvet		Filter	
	Project Properties	Coverage										
	Name	Sta	tus	Length	~	#	ClipL	ClipR	Start	Stop	C	Definit
	PhiX174-R1.fastq.gz	llum	nina Paired-e		301	5000					/\	Jsers
	PhiX174-R2.fastq.gz	llum	nina Paired-e		301	5000					/\	Jsers
-986.	Velvet PhiX174-Rx - 16:04 - Jur	1 3, 2019										
¶E.	Velvet PhiX174-Rx - 16:23 - Jur	n 3, 2019										
	Unaligned_Reads_41	454	ļ.		301	597					/\	/ar/fol
-9 1 E.	Contig 42			55	586	9266		1 5580	5	1	5586	

Now we see a second job object. In this case, the job has just two subitems: 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;

•••	Contig 42 — Editor
	🗇 🗸 👱 🗸 🛲 👯 📭 AGCT 🖄
Locked Text View Prefs Re	plica Translations Width Basecalls Qualities Dots Create
Editor	n Features Appetations Summary
Editor	
	<u>10 20 30 40 50 60 70 80 90 100 110</u>
6361	ACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
1349	TACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
2609	GACTACCCCCCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
1714	CCGACTACCCCCCCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
1197	CCGACTACCCCCCCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
8874	TTCCGACTACCCCCCCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
1551	GTTCCGACTACCCCCCCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
6283	GGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATCCCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
2853	GGTTCCGACTACCGTCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
1	TCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGGTCGCCATGATGGTGGTTATTATCCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGT
Consensus	TCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGG Export Consensus with Gaps
2068	TCGGTTCCGACTACCCCCCGACTGCCTATGATGTTTATCCTTTGGATGC Export Consensus without Gaps
5504	TCGGTTCCGACTACCCCCCGACTGCCTATGATGTTTATCCTTTGGATGC
7647	GGGTTCCGACTACCCCCGACTGCCTATGATGTTTATCCTTTGGATGC Export Selected Reads as FASTA
6110	GTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGC Export Selected Reads as FASTO
3277	TCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGC
5706	TCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGC
7040	ccgactaccctcccgactgcctatgatgtttatcctttggatgg Select Overlapping Reads Containing Selected Sequer
3434	GACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGGATGC
5276	CICCULARIZE CONSENSUS (200 ht overlap)
8015	CCCCCCCCGAGTGCTTATGATGTTTATCCTGTGGATGGTCGCAATGATGGTGGTTATTATCGTGACGGCGATGATGACTGTGGGCTATTGACGTCCTTCCCCGT

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.

Read pre-processing							
Discard reads less than 33 nt							
Trim ends with quality less than							
Trim N's from ends							
Discard short reads that contain any N's							
SPAdes Options							
Override coverage cutoff: 5							
Use custom K-MER: 31,41,51,61,71,81,91,101,							
Threads: 11 C Enter odd values less than 128 in ascending order, separated by commas (e.g. 21,33,55).							
Use "careful" mode with MismatchCorrector (slow)							
Generate Alignments Using Bowtie							
Threads: 11 🗘							
? Defaults Cancel 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;

•	• • •		Jntitled 2	— F	Project							
	1 생 김 4 일	š 💁 🙆	7 ~ (°		्य	-	2		35	Q~Na	me	
A	dd Reads Add Seqs Add Ref Add Contig Remo	ove Reset Prefs F	teplica Ph	red	CrossMatch	h Phrap	Bowtie	SPAdes	Velvet		Filter	
	Project Velvet PhiX174 Cove	erage										
	Name	Status	Length	~ #	#	ClipL	CI	ipR	Start	Stop)	Defini
	PhiX174-R1.fastq.gz	Ilumina Paired-e	30)1	5000							/User
	PhiX174-R2.fastq.gz	Ilumina Paired-e	30	01	5000							/User
48	SPAdes PhiX174-Rx - 18:07 - Jun 3, 20											
	Unaligned_Reads_43_1.fq.gz	Ilumina Paired-e	30)1	6							/var/f
	Unaligned_Reads_43_2.fq.gz	Ilumina Paired-e	30)1	6							/var/f
-010	NODE_1_length_5513_cov_322.4881		551	3	10000		1	5513		1	5513	
€#	NODE_2_length_128_cov_104.00000		12	8			1	128		1	128	
4	Velvet PhiX174-Rx - 16:04 - Jun 3, 2019											
48	Velvet PhiX174-Rx - 16:23 - Jun 3, 2019											

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.

Read pre-processing Discard reads less than 1,000 nt
FLYE Options
Expected genome size: 0.8 Mbp
Threads: 9
Flye polishing iterations:
Minimum overlap between reads: 1,000 bp
Initial minimum coverage: 170
Suppress polishing and contig coverage calculation Select this to speed up assembly if you just want to optimize the genome size and initial minimum coverage parameters
🗹 Run additional consensus polishing with Racon
Iterations: 1 C Window width: 500

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 rightclick (<ctrl>-click) context-sensitive menu has additional functions;

```
Export Consensus with Gaps...
Export Consensus without Gaps...
Align Selected Reads
Delete Selected Reads
Reset (un-align) Selected Reads
Export Selected Reads as FASTA...
Export Selected Reads as FASTQ...
Select Matching Pairs
Select Overlapping Reads Containing Selected Sequence
Extend Reference with Selected Read
Delete Clipped Residues
Close Gaps by Deleting Residues
```

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**.

Alignment Type: Sec	uence Con	firmation ᅌ	
Residue Scoring		Alignment Parameters	
Match:	2	Hash Value: 12	
Mismatch:	-3	Sensitivity: 6	\Diamond
Ambiguous Match:	0	Score Threshold:	50
Gap Penalty:	4	X Dropoff:	25
	Defaults	Cancel	к

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

Switch to the **Problems** tab.

• • •			phiX1	74(a) Alignment	. — Problems				
	0 1.	AGCT							
Unlocked P	refs Replica	Dots							
Editor	. N	lap	Features	Annotations	Text	SNPs	Prot	lems	
Potential :	otential sequencing problems report for phiX174(a) Alignment								
This tab i Once you h We recommen you set the	s primarily d ave a final s nd no more th e Hash Value	esigned to he equence, run an 5 million to 12 and Ser	elp you iden Align To Re reads unles nsitivity to	tify sequencing ference with it s you have more 3 or less for	errors with ge using a repres than 16 GB RAM optimal perform	nomic sequen entative sam . When you a ance.	ces and NGS o ple of your o lign, make su	ata. ata. re	
The follow	ing table can	be copied ar	nd pasted in	to Excel using	the paste speci	al option			
Ref Base i: Cons Base A, G, C and Gaps are th Ambig are Clip'd are Score is ca Only the h	s the residue is the residu d T are the c ne number of the number of the number o alculated by ighest-scorin	in the refer e in the cons ounts of the reads with ga reads with a f reads that adding 2 poir g (i.e. 'wors	ence sequen sensus seque occurence o aps at that ambiguities are clipped nts for each st') 2,000 p	ce nce f each residue location at that locatio (greyed out) a mismatch and 1 ositions are di	in reads that o n t that locatior for each gap/a splayed in the	ross the loc mbiguity/cli table	ation pped read		
Pos.	Ref Base	Cons Base	Score	A	C G	T Ga	ps Ambig	Clip'd	
1240	Т	Т	49	3 1	6 5	511	0 0	1	
2465	Т	Т	48	3 1	90	512	1 0	3	
1249	A	A	44	505	2 16	2	0 0	4	
4771	Т	Т	41	1 1	2 6	404	0 0	3	
462	A	A	40	605	5 13	2	0 0	Θ	
2493	Т	Т	39	9	B 2	506	0 0	1	
4160	Т	Т	38	1	1 16	569	0 0	2	
248	Т	Т	37	0 1	6 2	502	0 0	1	
456	Т	Т	37	4	3 10	599	1 0	2	
2460	Т	Т	34	8	4 4	524	0 0	2	
3857	Т	Т	31	1 .	4 8	740	0 0	5	
4222	С	С	31	12 57	7 2	Θ	0 0	3	
4238	Α	Α	31	575	3 11	1	0 0	1	
611	C	C	30	13 67	9 0	2	0 0	Ō	
3819	Ā	Ā	3.0	732	1 11	3	0 0	ē	
4140	т	т	30	5	5 5	580	0 0	õ	
4740	Å	Å	30	578	2 11	0		Ă	
4663	Ť	Ť	30	1 1	2 1	382	6 6	2	
535	Ť	+	20	2 1	2 10	635	6 6	1	
1100	Ť	1 T	29	1	2 10 a 12	555	0 0	1	
4100	1		29	C . 1	0 15	502	0 0	1	

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	Α	С	G	т	Gaps	Ambig	Clip'd
2254	G	A	1270	631	Θ	1	Θ	Θ	ō	. 8
2656	Т	С	1089	2	539	Θ	1	Θ	Θ	7
2576	G	A	1006	499	1	3	Θ	Θ	Θ	6
4800	С	Т	881	Θ	1	1	436	1	Θ	6
4554	С	Т	880	Θ	3	1	435	Θ	Θ	8
1240	Т	Т	49	3	16	5	511	Θ	Θ	1
2465	Т	Т	48	3	19	Θ	512	1	Θ	3
1249	A	A	44	505	2	16	2	Θ	Θ	4
4771	Т	Т	41	1	12	6	404	Θ	Θ	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.

			phiX1	174 Alignmen	t — Ed	litor				
_n =	@		-33	OB	as	0 ~	AGCT	ACGTAC	ACGTAC	>>
Unlooked Toxt View	Drofo Dopli		Add Soco	Es Domovio Como	Alian '	Translations	Dete	Linet Miama	toh Novt Mismotoh	
Onlocked Text view	Preis Repli	ica iopology	Auu Seys	Remove Seqs	Aligh	Translations	DOIS	FIIST WISHIG	ten next mismaten	
Editor	Map	Feature	s	Annotations		Text	S	NPs	Problems	
Sort	∠,	2220	2230	2240	2250	2260	2.	270 2	280 2290	
	phiX174 TC	CTTTACCAGCTTT	AGCCATAG	CACCAGAAACAAA	ACTAGG	GGCGGCCTCA	TCAG-GO	GTTAGGAACA	TTAG-AGCCTTGAATG	GCAG
M03532-54-00 5657.1	Consensus					.A				
M03532:54:00:1884	1:N:0:23					A				
M03532:54:000757 1	:N:0:23					A				
M03532:54:002144 2	:N:0:23					. A				
M03532:54:00:3589	2:N:0:23)					.AC	c	т	.ACC	
M03532:54:00:7430	2:N:0:23					.A				
M03532:54:00:9621	2:N:0:23 ┥					. A				
M03532:54:00:3410	1:N:0:23 🕨					. A				
M03532:54:005084 2	::N:0:23 🕨	.GA.			C	. A G	c	G.GC	GGG	
M03532:54:000287 1	:N:0:23 ┥					. A				
M03532:54:005989 1	:N:0:23 🔳					.A				
M03532:54:005985 2	::N:0:23 ┥					. A				
M03532:54:001674 1	:N:0:23 🕨					. A				
M03532:54:000308 2	::N:0:23 ┥					. A				
M03532:54:00:9994	2:N:0:23 ┥					.A				
M03532:54:008597 2	::N:0:23 ▶					. A C	A	c	AG	
M03532:54:008463 2	::N:0:23 🖣					. A				
M03532:54:009943 2	::N:0:23 🔳					. A				
M03532:54:008240 2	::N:0:23 ▶					.AG		• • • • • • • • • • •	GC	
M03532:54:001815 1	:N:0:23 ▶					. A				• •
M03532:54:00:5202	2:N:0:23 ▶	C.(. A	A	C		• •
M03532:54:00:6083	2:N:0:23	• • • • • • • • • • • • • • • •		• • • • • • • • • • • • • • •	• • • • • •	. A			• • • • • • • • • • • • • • • • • • • •	
M03532:54:00:6975	1:N:0:23	•••••		• • • • • • • • • • • • • • •	• • • • • •	. A	• • • • • • •		••••••	
M03532:54:009540 2	::N:0:23 ▶	•••••	. A	• • • • • • • • • • • • • • •		. A	• • • • • • •		T	A
M03532:54:006335 2	::N:0:23 ▶	•••••	• • • • • • • • •		• • • • • • •	. A		.G	••••••	• • • •
M03532:54:006814 1	:N:0:23 4	• • • • • • • • • • • • • • • • • • • •	• • • • • • • • •	••••••	• • • • • • •	. A	• • • • • • •	• • • • • • • • • • • •	••••••	
M03532:54:00:7475	2:N:0:23)	•••••				.A	• • • • • • •			
M03532:54:008831 2	:IN:U:23	• • • • • • • • • • • • • • • • • • •	• • • • • • • • •	••••••	• • • • • • •	.A	• • • • • • •	• • • • • • • • • • • •	•••••	• • • •
M03532:54:008911 2	:N:0:23 4	• • • • • • • • • • • • • • • • • •	• • • • • • • • •	••••••	• • • • • • •	. A	• • • • • • •	• • • • • • • • • • • •	••••••	• •
M03532:54:00:9186	2:N:0:23 🕨					. A				

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.