MacVector 17.5

for Mac OS X

Virtual Gene Cloning from NGS RNA-Seq Data

MacVector, Inc.

Software for Scientists

Copyright statement

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 the Virtual Gene Cloning tutorial was published in June 2020.

Contents

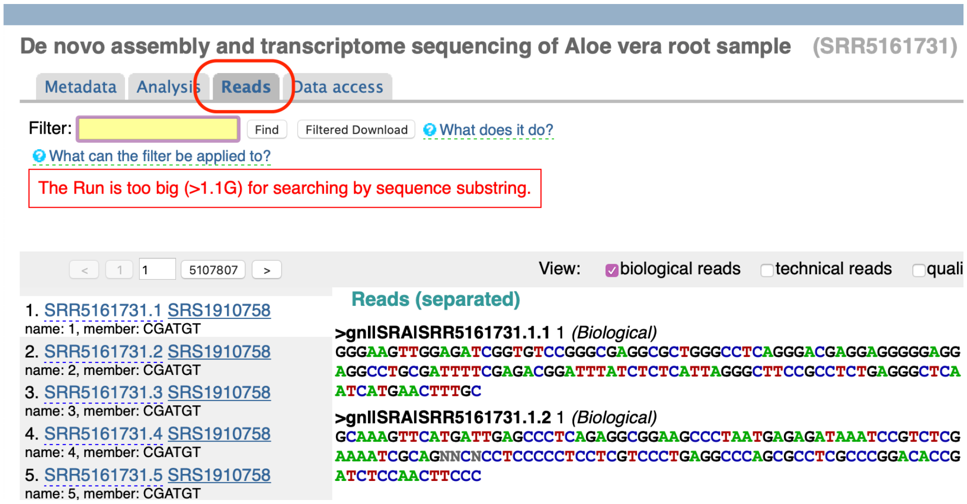
Introduction

The NCBI Sequence Read Archive (SRA) database is a huge resource of Next Generation Sequencing experimental data. Many groups and laboratories deposit data here that they have generated for their own specific projects that can be datamined for other unrelated projects with a minimum of effort. In this tutorial, we will demonstrate how you can use MacVector to identify and assemble genes encoding zinc finger proteins from RNA-Seq data from a plant species.

For this tutorial we will "clone" zinc finger proteins of the C2H2 type from a particular plant, *Aloe vera*, that is commonly used in the cosmetic, pharmaceutical and food industries. C2H2 zinc finger proteins are important transcriptional regulators that, in plants, have a highly conserved sequence, QALGGH, located within a putative DNA-contacting surface of each finger.

Sample Files

The data for this tutorial is too large to be included with the MacVector software distribution so needs to be downloaded from the NCBI SRA website. The tutorial uses RNA-Seq data isolated from an *Aloe vera* root sample - https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR5161731. To download the data, first click on the **Reads** tab;



De novo assembly and transcriptome sequencing of Aloe vera root sample (SRR5161731)

Filter:

The Run is too big (>1.1G) for searching by sequence substring.

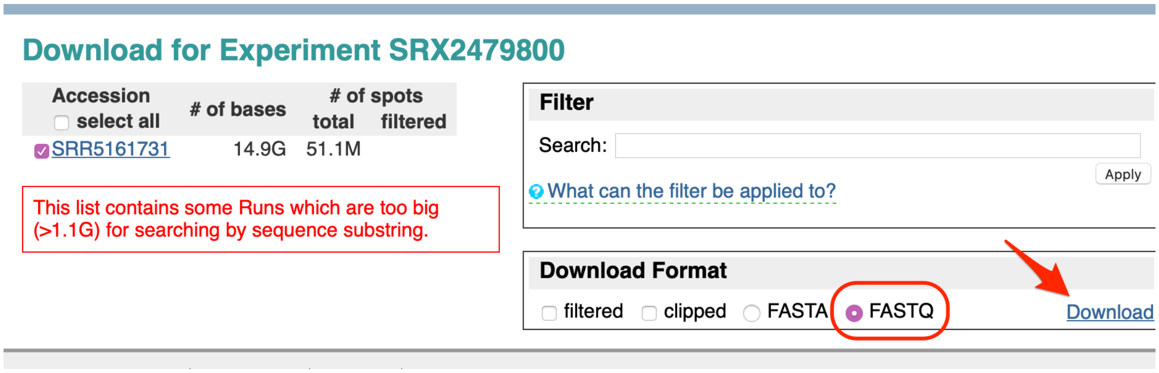
View: ☑ biological reads □ technical reads □ quali

1. SRR5161731.1 SRS1910758
name: 1, member: CGATGT
2. SRR5161731.2 SRS1910758
name: 2, member: CGATGT
3. SRR5161731.3 SRS1910758
name: 3, member: CGATGT
4. SRR5161731.4 SRS1910758
name: 4, member: CGATGT
5. SRR5161731.5 SRS1910758
name: 5, member: CGATGT

Reads (separated)

```
>gnl|SRA|SRR5161731.1.1 1 (Biological)
GGGAAGTTGGAGATCGGTGTCCGGGCGAGGCGCTGGGCCTCAGGGACGAGGAGGGGGAGG
AGGCCTGCGATTTTCGAGACGGATTTATCTCTCATTAGGGCTTCCGCCTCTGAGGGCTCA
ATCATGAACTTTGC
>gnl|SRA|SRR5161731.1.2 1 (Biological)
GCAAAGTTCATGATTGAGCCCTCAGAGGCGGAAGCCCTAATGAGAGATAAATCCGTCTCG
AAAATCGCAGNNCNCCTCCCCCTCCTCGTCCCTGAGGCCCAGCGCCTCGCCCGGACACCG
ATCTCCAACTTCCC
```

Then click on the **Filtered Download** button to get to the download page;



Download for Experiment SRX2479800

Accession □ select all	# of bases	# of spots total	 filtered
☑ SRR5161731	14.9G	51.1M	

This list contains some Runs which are too big (>1.1G) for searching by sequence substring.

Filter

Search:

Download Format

□ filtered □ clipped ○ FASTA ● FASTQ

Select **FASTQ** format and click on the **Download** button. By default, the data will download as a compressed file called `sra_data.fastq.gz`. You may find that your computer automatically decompresses the file. This is not a problem; it just means that the file takes up more space on your hard drive. The file contains paired-end reads in "interleaved" format where the reads are organized as Read1-left, Read1-right, Read2-left, Read2-right etc. This allows all of the data to be contained within a single file. More typically, you may receive data in a pair of files. MacVector can handle either type, with or without gzip compression.

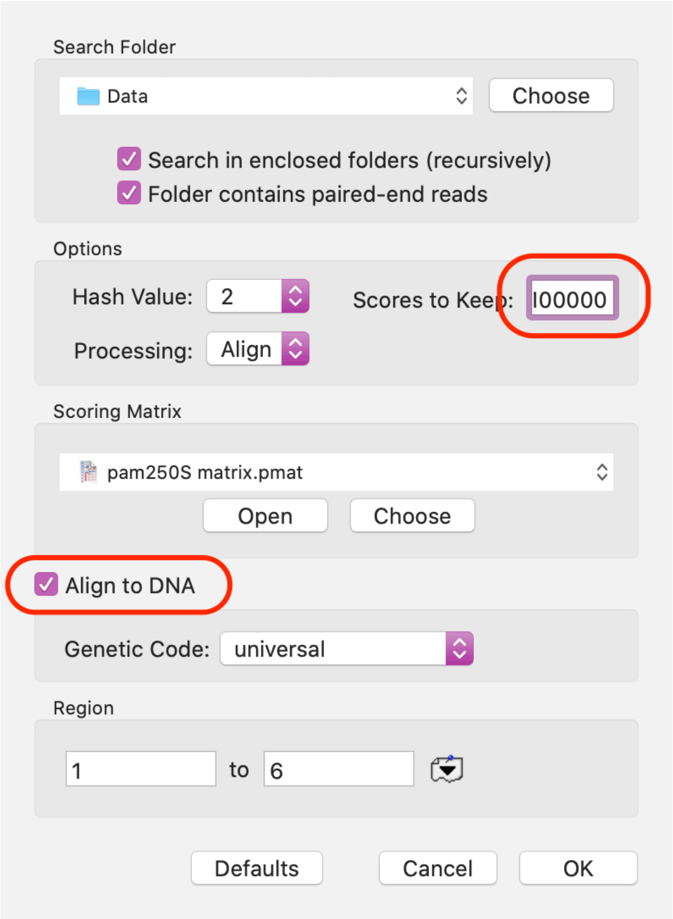
Finally, move the file into a unique folder e.g. "`Data`"

Note that there is a companion "leaf" dataset (SRR5167034) that is incomplete. The "reverse" reads in that set are actually a duplication of the "forward" reads. The data is still usable, but it is now effectively unpaired data and the analysis is more tedious, though you can follow the steps in the tutorial and still get successful assemblies. This tutorial was first written using SRR5167034 data and so some of the example screenshots may still show that data set.

Tutorial

Find Reads that Encode QALGGH

The first step is to identify all RNA-Seq reads in the data set that potentially encode the conserved QALGGH domain. Choose **File | New | Protein** to create a new protein sequence window, then type "QALGGH" into the **Editor**. Then choose **Database | Align to Folder**;



Search Folder

Data

☑ Search in enclosed folders (recursively)

☑ Folder contains paired-end reads

Options

Hash Value: 2

Scores to Keep: 100000

Processing: Align

Scoring Matrix

pam250S matrix.pmat

☑ Align to DNA

Genetic Code: universal

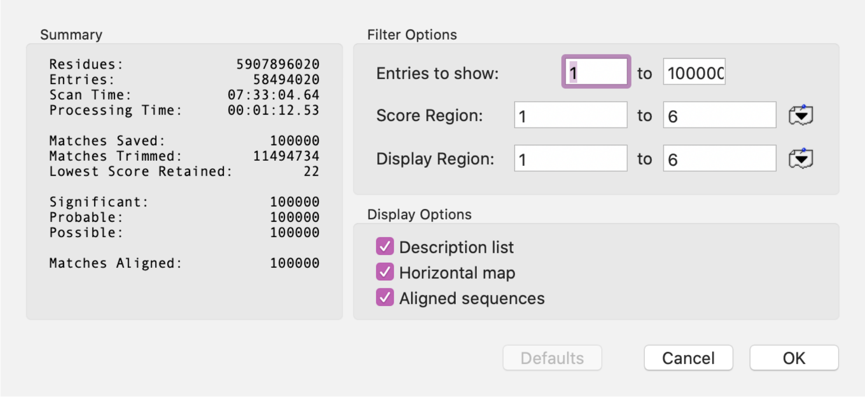
Region

1 to 6

There are two important changes we need to make. (a) make sure you set **Scores to Keep** to a large number (e.g. 100,000) as we need to save as many reads as we can. (b) Make sure you have checked the **Align to DNA** checkbox. We are starting with an amino acid sequence, but we want to find all the *DNA sequences* that potentially could encode that protein.

When you click **OK**, the search will start. There is a lot of data to process in the fasta file, and each read needs to be translated in all 6 reading frames before being aligned to our input sequence, so this will take some time. On an i7 MacBook Pro the search took 7 hours and 45 minutes.

When complete, select all three output options and click **OK**.



Summary

```
Residues:                 5907896020
Entries:                    58494020
Scan Time:               07:33:04.64
Processing Time:         00:01:12.53

Matches Saved:                100000
Matches Trimmed:            11494734
Lowest Score Retained:            22

Significant:                  100000
Probable:                     100000
Possible:                     100000

Matches Aligned:              100000
```

Filter Options

Entries to show: 1 to 10000C

Score Region: 1 to 6

Display Region: 1 to 6

Display Options

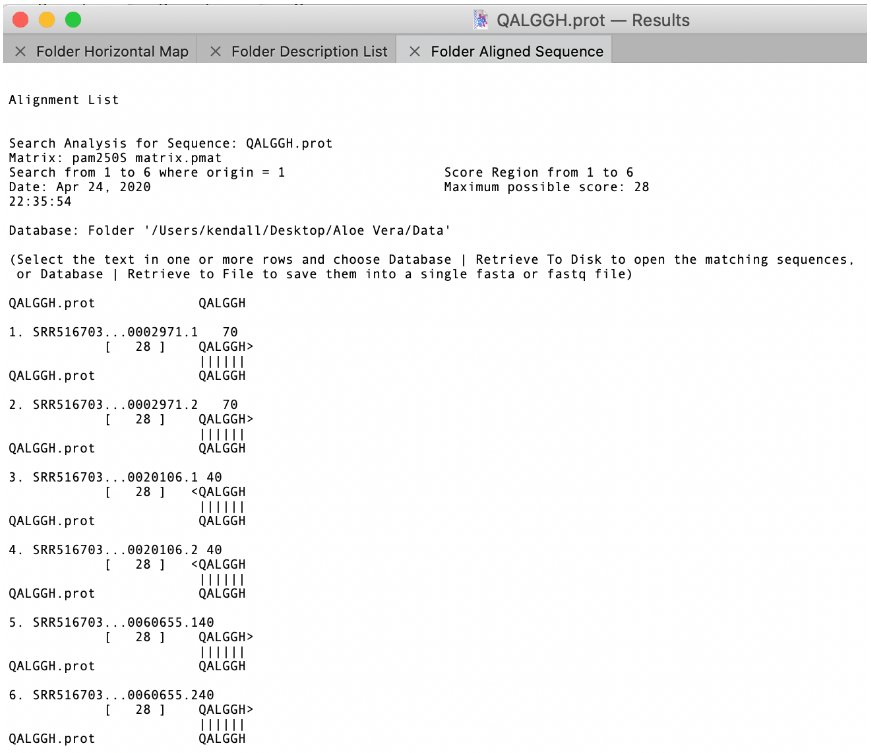
☑ Description list

☑ Horizontal map

☑ Aligned sequences

With 100,000 hits to display, the result windows can take some time to generate their content. On a MacBook Pro it took 3 minutes for the results to appear.

The **Folder Aligned Sequence** tab displays the translation of each hit aligned against the query sequence. You can see that the first few entries are perfect matches;



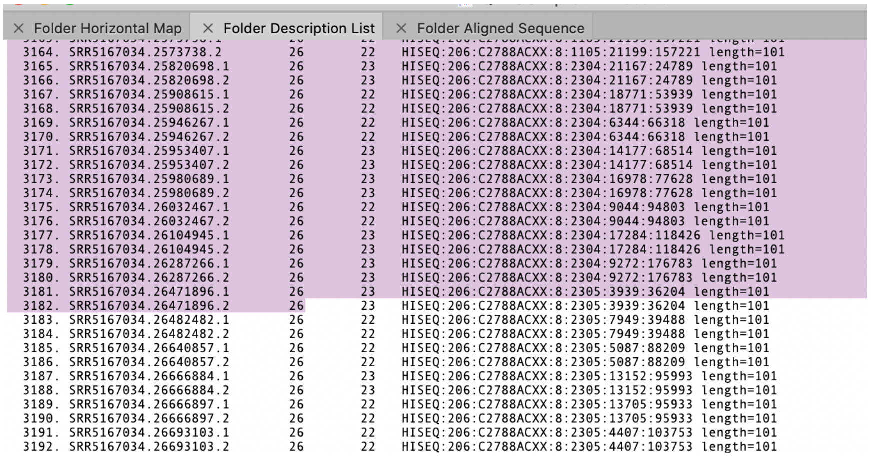
```
Alignment List


Search Analysis for Sequence: QALGGH.prot
Matrix: pam250S matrix.pmat
Search from 1 to 6 where origin = 1                     Score Region from 1 to 6
Date: Apr 24, 2020                                      Maximum possible score: 28
22:35:54

Database: Folder '/Users/kendall/Desktop/Aloe Vera/Data'

(Select the text in one or more rows and choose Database | Retrieve To Disk to open the matching sequences,
 or Database | Retrieve to File to save them into a single fasta or fastq file)

QALGGH.prot              QALGGH

1. SRR516703...0002971.1   70
            [   28 ]     QALGGH>
                         ||||||
QALGGH.prot              QALGGH

2. SRR516703...0002971.2   70
            [   28 ]     QALGGH>
                         ||||||
QALGGH.prot              QALGGH

3. SRR516703...0020106.1 40
            [   28 ]    <QALGGH
                         ||||||
QALGGH.prot              QALGGH

4. SRR516703...0020106.2 40
            [   28 ]    <QALGGH
                         ||||||
QALGGH.prot              QALGGH

5. SRR516703...0060655.140
            [   28 ]     QALGGH>
                         ||||||
QALGGH.prot              QALGGH

6. SRR516703...0060655.240
            [   28 ]     QALGGH>
                         ||||||
QALGGH.prot              QALGGH
```

As you scroll down the list, eventually you start to see imperfect matches;

```
3181. SRR51670...71896.1   10
            [   26 ]    <HPLGGH
                         ||||||
QALGGH.prot              QALGGH

3182. SRR51670...71896.2   10
            [   26 ]    <HPLGGH
                         ||||||
QALGGH.prot              QALGGH

3183. SRR51670...82482.1
            [   26 ]     QQLGGH>
                         | ||||
QALGGH.prot              QALGGH

3184. SRR51670...82482.2
            [   26 ]     QQLGGH>
                         | ||||
QALGGH.prot              QALGGH
```

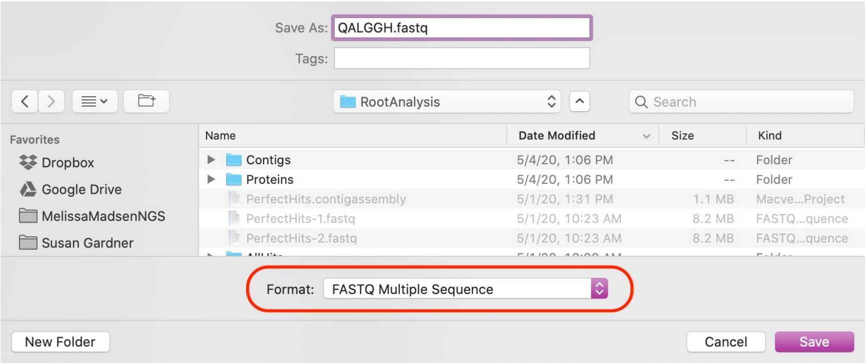
For this tutorial, we want to retrieve just those reads with perfect matches. If you are looking for different domains, you may prefer to include reads with just partial matches to the query.

To retrieve the reads, switch to the **Folder Description List** tab, click at the beginning of the row corresponding to the first hit, hold down the `<shift>` key, scroll down to the row corresponding to the last perfect match and click at the end of that row. You don't have to actually select an entire row – selecting any part of a row is treated as selecting the entire row. Note that if you want to retrieve all of the hits in a **Folder Description List** tab, you can simply choose **Edit | Select All** (`<command>-A`).



```
3164. SRR5167034.2573738.2      26      22    HISEQ:206:C2788ACXX:8:1105:21199:157221 length=101
3165. SRR5167034.25820698.1     26      23    HISEQ:206:C2788ACXX:8:2304:21167:24789 length=101
3166. SRR5167034.25820698.2     26      23    HISEQ:206:C2788ACXX:8:2304:21167:24789 length=101
3167. SRR5167034.25908615.1     26      22    HISEQ:206:C2788ACXX:8:2304:18771:53939 length=101
3168. SRR5167034.25908615.2     26      22    HISEQ:206:C2788ACXX:8:2304:18771:53939 length=101
3169. SRR5167034.25946267.1     26      22    HISEQ:206:C2788ACXX:8:2304:6344:66318 length=101
3170. SRR5167034.25946267.2     26      22    HISEQ:206:C2788ACXX:8:2304:6344:66318 length=101
3171. SRR5167034.25953407.1     26      23    HISEQ:206:C2788ACXX:8:2304:14177:68514 length=101
3172. SRR5167034.25953407.2     26      23    HISEQ:206:C2788ACXX:8:2304:14177:68514 length=101
3173. SRR5167034.25980689.1     26      23    HISEQ:206:C2788ACXX:8:2304:16978:77628 length=101
3174. SRR5167034.25980689.2     26      23    HISEQ:206:C2788ACXX:8:2304:16978:77628 length=101
3175. SRR5167034.26032467.1     26      22    HISEQ:206:C2788ACXX:8:2304:9044:94803 length=101
3176. SRR5167034.26032467.2     26      22    HISEQ:206:C2788ACXX:8:2304:9044:94803 length=101
3177. SRR5167034.26104945.1     26      23    HISEQ:206:C2788ACXX:8:2304:17284:118426 length=101
3178. SRR5167034.26104945.2     26      23    HISEQ:206:C2788ACXX:8:2304:17284:118426 length=101
3179. SRR5167034.26287266.1     26      23    HISEQ:206:C2788ACXX:8:2304:9272:176783 length=101
3180. SRR5167034.26287266.2     26      23    HISEQ:206:C2788ACXX:8:2304:9272:176783 length=101
3181. SRR5167034.26471896.1     26      23    HISEQ:206:C2788ACXX:8:2305:3939:36204 length=101
3182. SRR5167034.26471896.2     26      23    HISEQ:206:C2788ACXX:8:2305:3939:36204 length=101
3183. SRR5167034.26482482.1     26      22    HISEQ:206:C2788ACXX:8:2305:7949:39488 length=101
3184. SRR5167034.26482482.2     26      22    HISEQ:206:C2788ACXX:8:2305:7949:39488 length=101
3185. SRR5167034.26640857.1     26      22    HISEQ:206:C2788ACXX:8:2305:5087:88209 length=101
3186. SRR5167034.26640857.2     26      22    HISEQ:206:C2788ACXX:8:2305:5087:88209 length=101
3187. SRR5167034.26666884.1     26      23    HISEQ:206:C2788ACXX:8:2305:13152:95993 length=101
3188. SRR5167034.26666884.2     26      23    HISEQ:206:C2788ACXX:8:2305:13152:95993 length=101
3189. SRR5167034.26666897.1     26      22    HISEQ:206:C2788ACXX:8:2305:13705:95933 length=101
3190. SRR5167034.26666897.2     26      22    HISEQ:206:C2788ACXX:8:2305:13705:95933 length=101
3191. SRR5167034.26693103.1     26      22    HISEQ:206:C2788ACXX:8:2305:4407:103753 length=101
3192. SRR5167034.26693103.2     26      22    HISEQ:206:C2788ACXX:8:2305:4407:103753 length=101
```

Finally choose **Database | Retrieve to File...** and choose a suitable file name, making sure to select **FASTQ Multiple Sequence** as the output file type;



Save As: QALGGH.fastq

Tags:

Name	Date Modified	Size	Kind
▶ Contigs	5/4/20, 1:06 PM	--	Folder
▶ Proteins	5/4/20, 1:06 PM	--	Folder
PerfectHits.contigassembly	5/1/20, 1:31 PM	1.1 MB	Macve...Project
PerfectHits-1.fastq	5/1/20, 10:23 AM	8.2 MB	FASTQ...quence
PerfectHits-2.fastq	5/1/20, 10:23 AM	8.2 MB	FASTQ...quence

Format: FASTQ Multiple Sequence

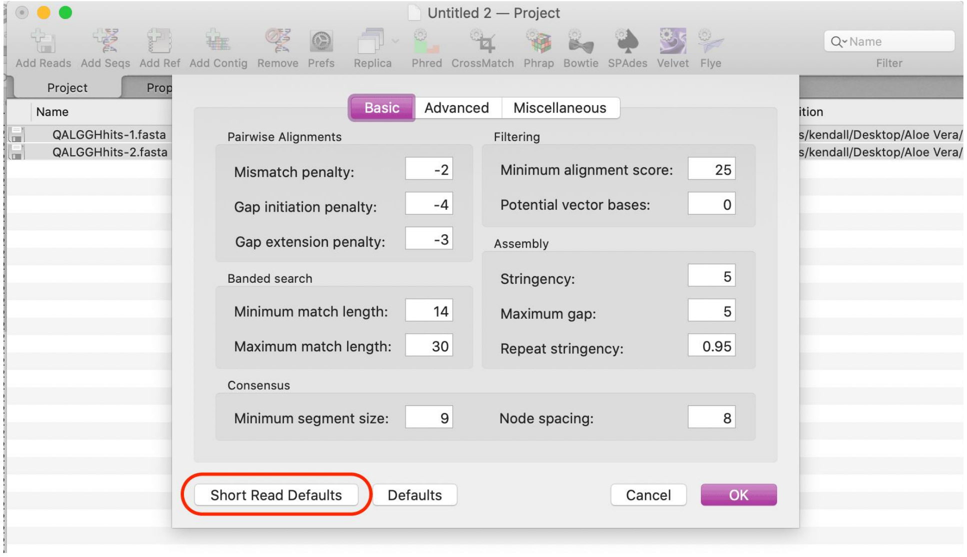
The reads are saved into a pair of files, with "-1" and "-2" appended to the filename. E.g. `QALGGH-1.fastq` and `QALGGH-2.fastq`.

Assembling Matching Reads

The next step is to assemble the reads into contigs. With luck, each contig will represent a different Zn finger gene, where they all contain one or more conserved QALGGH domains.

First create a **File | New | Assembly Project**. Then click on the **Add Reads** button and select the pair of fastq files you saved in the last section. MacVector has four separate *de novo* assembly algorithms, each of which has different characteristics and uses. *Flye* is used only for long read assemblies (e.g. Oxford Nanopore and PacBio reads). *Velvet* and *SPAdes* are tuned primarily for genomic assemblies of many millions of short reads. In this case, we are expecting the reads to assemble into a number of short cDNA-length contig sequences, rather than a single long genome. *Velvet* and *SPAdes* are not really optimized for this, type of assembly though in many cases they will generate reasonable results. However, for this data, the optimal algorithm to use is *phrap*. This is an older algorithm, originally optimized for Sanger sequencing data, but it does an extremely good job of assembling reads into a series of short contigs, even if there are few overlapping reads. The downside is that it can be slow with a large number of reads.

Select the two reads, then click on the **phrap** toolbar button.



Name	
QALGGHhits-1.fasta	s/kendall/Desktop/Aloe Vera/
QALGGHhits-2.fasta	s/kendall/Desktop/Aloe Vera/

Pairwise Alignments
- Mismatch penalty: -2
- Gap initiation penalty: -4
- Gap extension penalty: -3

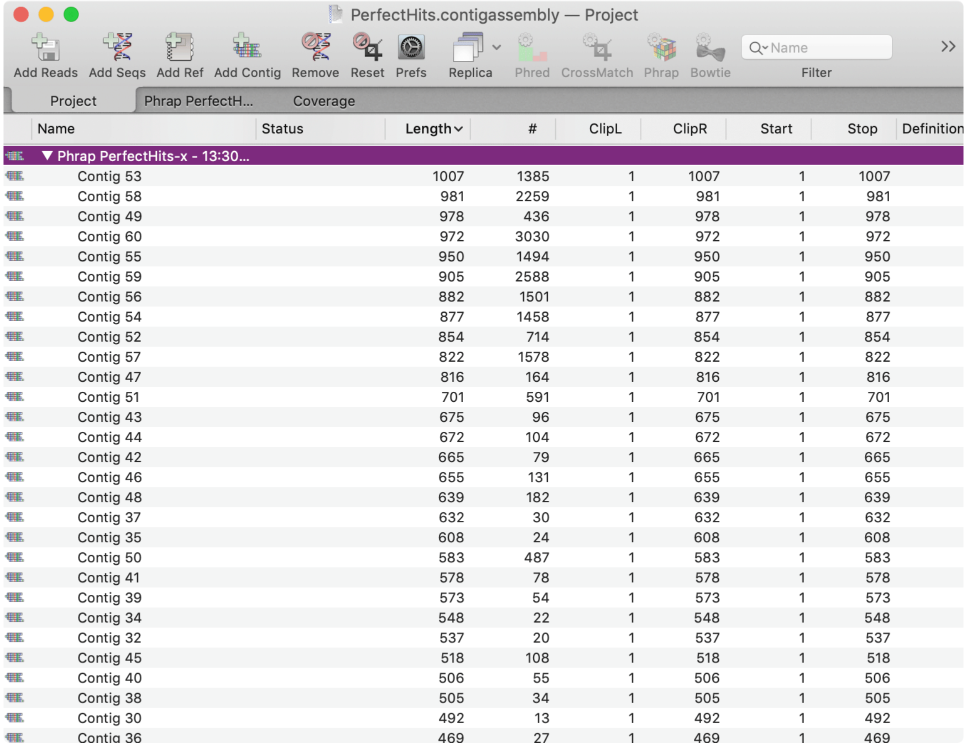
Banded search
- Minimum match length: 14
- Maximum match length: 30

Filtering
- Minimum alignment score: 25
- Potential vector bases: 0

Assembly
- Stringency: 5
- Maximum gap: 5
- Repeat stringency: 0.95

Consensus
- Minimum segment size: 9
- Node spacing: 8

Click on the **Short Read Defaults** button to set up the optimal parameters for this data, then click **OK**. Because we are only using ~20,000 pairs of reads, and the reads are short (~100nt each) the job completes in a moderate length of time (3 hours), resulting in a ***Phrap*** job appearing in the project window with ~60 contigs;



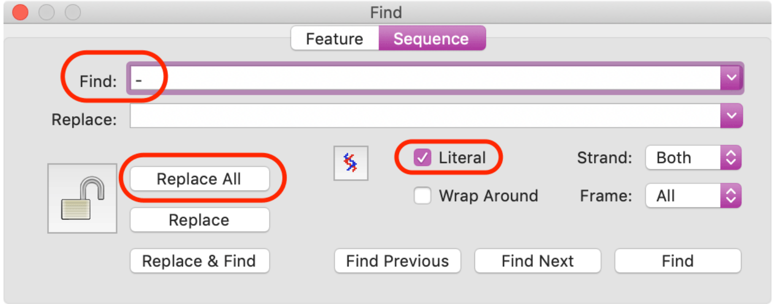
Name	Status	Length	#	ClipL	ClipR	Start	Stop	Definition
▼ Phrap PerfectHits-x - 13:30...								
Contig 53		1007	1385	1	1007	1	1007	
Contig 58		981	2259	1	981	1	981	
Contig 49		978	436	1	978	1	978	
Contig 60		972	3030	1	972	1	972	
Contig 55		950	1494	1	950	1	950	
Contig 59		905	2588	1	905	1	905	
Contig 56		882	1501	1	882	1	882	
Contig 54		877	1458	1	877	1	877	
Contig 52		854	714	1	854	1	854	
Contig 57		822	1578	1	822	1	822	
Contig 47		816	164	1	816	1	816	
Contig 51		701	591	1	701	1	701	
Contig 43		675	96	1	675	1	675	
Contig 44		672	104	1	672	1	672	
Contig 42		665	79	1	665	1	665	
Contig 46		655	131	1	655	1	655	
Contig 48		639	182	1	639	1	639	
Contig 37		632	30	1	632	1	632	
Contig 35		608	24	1	608	1	608	
Contig 50		583	487	1	583	1	583	
Contig 41		578	78	1	578	1	578	
Contig 39		573	54	1	573	1	573	
Contig 34		548	22	1	548	1	548	
Contig 32		537	20	1	537	1	537	
Contig 45		518	108	1	518	1	518	
Contig 40		506	55	1	506	1	506	
Contig 38		505	34	1	505	1	505	
Contig 30		492	13	1	492	1	492	
Contig 36		469	27	1	469	1	469	

You can double-click on a contig to open the *Contig Editor* so that you can see the actual alignments. As we are hoping that each contig will represent a different Zn finger coding gene, lets save them all as individual MacVector `.nucl` sequences. Select the first contig, hold down <`shift`>, select the last contig, then choose **File | Export Selected Contigs To...** and choose a suitable destination directory.

Identification of Zn Finger Encoding Contigs

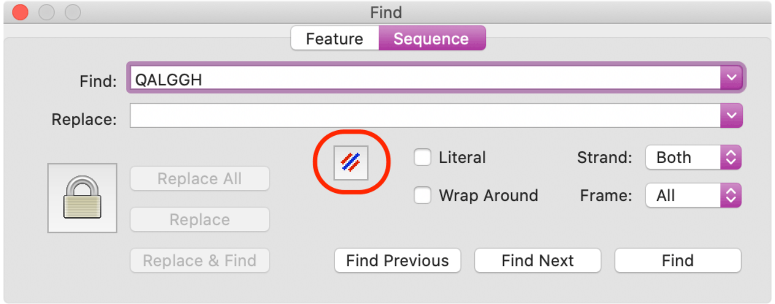
We can assume that every contig *should* have a DNA sequence that can potentially encode QALGGH. However, we cannot be sure that the corresponding DNA sequence actually lies in-frame within a potential protein-coding region. So, let's, (a) confirm that each contig does indeed code for QALGGH and then (b) look to see if there is a potential ORF in the correct frame to include QALGGH in the predicted amino acid sequence.

Open the longest contig (`Contig 53.nucl` in our example, but yours may vary). You can actually perform the analysis that follows directly on the contigs from the project if you wish. Note: versions of MacVector prior to 18.0 retain gaps in the exported contigs. You should remove any gaps before performing analyses. The easiest way to do this is using the **Edit | Find | Find** function. Set up the **Find** dialog like this with the **Literal** box checked, a dash in the **Find** box, and empty **Replace** box and click **Replace All** (make sure to unlock the sequence);



Find: -
Replace:
☑ Literal
☐ Wrap Around
Strand: Both
Frame: All

Now we can search for QALGGH domains in the `Contig 53.nucl` window. Choose **Edit | Find | Find..** once again but this time set up the dialog like this;



Find: QALGGH

Replace:

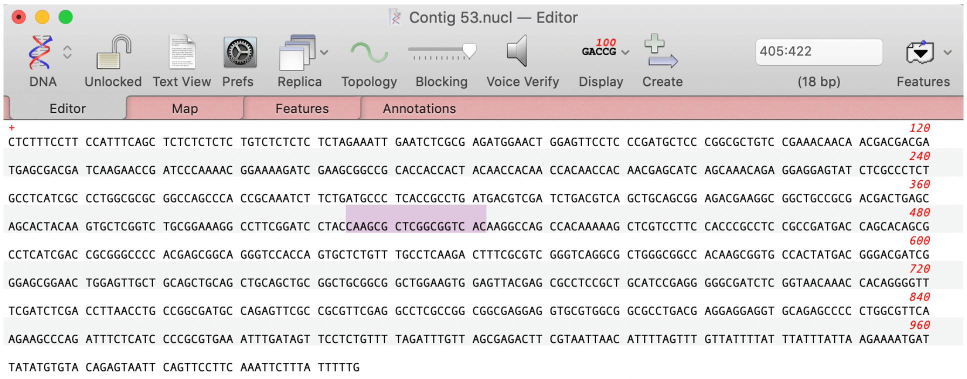
□ Literal

□ Wrap Around

Strand: Both

Frame: All

Note in particular to toggle the circled "mode" button to the protein symbol. This lets you search a DNA sequence with an amino acid sequence. If you type "QALGGH" into the **Find:** box and click on the **Find** button, you should see that the **Editor** tab updates to show this;

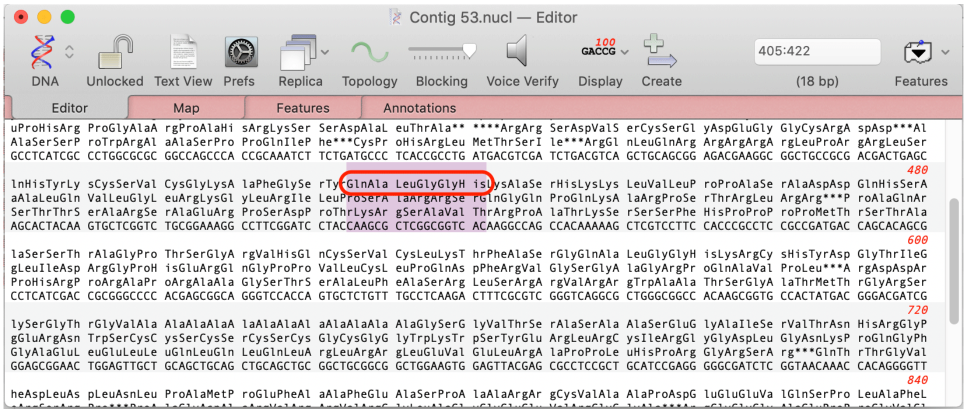


405:422

(18 bp)

```
+                                                                                                              120
CTCTTTCCTT CCATTTCAGC TCTCTCTCTC TGTCTCTCTC TCTAGAAATT GAATCTCGCG AGATGGAACT GGAGTTCCTC CCGATGCTCC CGGCGCTGTC CGAAACAACA ACGACGACGA
                                                                                                              240
TGAGCGACGA TCAAGAACCG ATCCCAAAAC GGAAAAGATC GAAGCGGCCG CACCACCACT ACAACCACAA CCACAACCAC AACGAGCATC AGCAAACAGA GGAGGAGTAT CTCGCCCTCT
                                                                                                              360
GCCTCATCGC CCTGGCGCGC GGCCAGCCCA CCGCAAATCT TCTGATGCCC TCACCGCCTG ATGACGTCGA TCTGACGTCA GCTGCAGCGG AGACGAAGGC GGCTGCCGCG ACGACTGAGC
                                                                                                              480
AGCACTACAA GTGCTCGGTC TGCGGAAAGG CCTTCGGATC CTACCAAGCG CTCGGCGGTC ACAAGGCCAG CCACAAAAAG CTCGTCCTTC CACCCGCCTC CGCCGATGAC CAGCACAGCG
                                                                                                              600
CCTCATCGAC CGCGGGCCCC ACGAGCGGCA GGGTCCACCA GTGCTCTGTT TGCCTCAAGA CTTTCGCGTC GGGTCAGGCG CTGGGCGGCC ACAAGCGGTG CCACTATGAC GGGACGATCG
                                                                                                              720
GGAGCGGAAC TGGAGTTGCT GCAGCTGCAG CTGCAGCTGC GGCTGCGGCG GCTGGAAGTG GAGTTACGAG CGCCTCCGCT GCATCCGAGG GGGCGATCTC GGTAACAAAC CACAGGGGTT
                                                                                                              840
TCGATCTCGA CCTTAACCTG CCGGCGATGC CAGAGTTCGC CGCGTTCGAG GCCTCGCCGG CGGCGAGGAG GTGCGTGGCG GCGCCTGACG AGGAGGAGGT GCAGAGCCCC CTGGCGTTCA
                                                                                                              960
AGAAGCCCAG ATTTCTCATC CCCGCGTGAA ATTTGATAGT TCCTCTGTTT TAGATTTGTT AGCGAGACTT CGTAATTAAC ATTTTAGTTT GTTATTTTAT TTATTTATTA AGAAAATGAT

TATATGTGTA CAGAGTAATT CAGTTCCTTC AAATTCTTTA TTTTTG
```

Initially, this may appear confusing as the highlighted sequence bears little resemblance to the search sequence. However, if you click on the **Display** toolbar button and select the **Show Plus Strand Translations** options all should become much clearer;

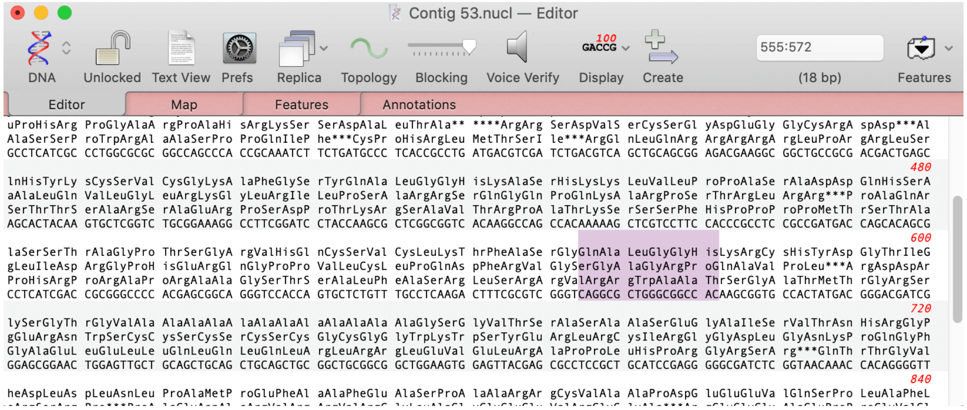


405:422

(18 bp)

```
uProHisArg ProGlyAlaA rgProAlaHi sArgLysSer SerAspAlaL euThrAla** ****ArgArg SerAspValS erCysSerGl yAspGluGly GlyCysArgA spAsp***Al
AlaSerSerP roTrpArgAl aAlaSerPro ProGlnIleP he***CysPr oHisArgLeu MetThrSerI le***ArgGl nLeuGlnArg ArgArgArgA rgLeuProAr gArgLeuSer
GCCTCATCGC CCTGGCGCGC GGCCAGCCCA CCGCAAATCT TCTGATGCCC TCACCGCCTG ATGACGTCGA TCTGACGTCA GCTGCAGCGG AGACGAAGGC GGCTGCCGCG ACGACTGAGC
                                                                                                              480
lnHisTyrLy sCysSerVal CysGlyLysA laPheGlySe rTyrGlnAla LeuGlyGlyH isLysAlaSe rHisLysLys LeuValLeuP roProAlaSe rAlaAspAsp GlnHisSerA
aAlaLeuGln ValLeuGlyL euArgLysGl yLeuArgIle LeuProSerA laArgArgSe rGlnGlyGln ProGlnLysA laArgProSe rThrArgLeu ArgArg***P roAlaGlnAr
SerThrThrS erAlaArgSe rAlaGluArg ProSerAspP roThrLysAr gSerAlaVal ThrArgProA laThrLysSe rSerSerPhe HisProProP roProMetTh rSerThrAla
AGCACTACAA GTGCTCGGTC TGCGGAAAGG CCTTCGGATC CTACCAAGCG CTCGGCGGTC ACAAGGCCAG CCACAAAAAG CTCGTCCTTC CACCCGCCTC CGCCGATGAC CAGCACAGCG
                                                                                                              600
laSerSerTh rAlaGlyPro ThrSerGlyA rgValHisGl nCysSerVal CysLeuLysT hrPheAlaSe rGlyGlnAla LeuGlyGlyH isLysArgCy sHisTyrAsp GlyThrIleG
gLeuIleAsp ArgGlyProH isGluArgGl nGlyProPro ValLeuCysL euProGlnAs pPheArgVal GlySerGlyA laGlyArgPr oGlnAlaVal ProLeu***A rgAspAspAr
ProHisArgP roArgAlaPr oArgAlaAla GlySerThrS erAlaLeuPh eAlaSerArg LeuSerArgA rgValArgAr gTrpAlaAla ThrSerGlyA laThrMetTh rGlyArgSer
CCTCATCGAC CGCGGGCCCC ACGAGCGGCA GGGTCCACCA GTGCTCTGTT TGCCTCAAGA CTTTCGCGTC GGGTCAGGCG CTGGGCGGCC ACAAGCGGTG CCACTATGAC GGGACGATCG
                                                                                                              720
lySerGlyTh rGlyValAla AlaAlaAlaA laAlaAlaAl aAlaAlaAla AlaGlySerG lyValThrSe rAlaSerAla AlaSerGluG lyAlaIleSe rValThrAsn HisArgGlyP
gGluArgAsn TrpSerCysC ysSerCysSe rCysSerCys GlyCysGlyG lyTrpLysTr pSerTyrGlu ArgLeuArgC ysIleArgGl yGlyAspLeu GlyAsnLysP roGlnGlyPh
GlyAlaGluL euGluLeuLe uGlnLeuGln LeuGlnLeuA rgLeuArgAr gLeuGluVal GluLeuArgA laProProLe uHisProArg GlyArgSerA rg***GlnTh rThrGlyVal
GGAGCGGAAC TGGAGTTGCT GCAGCTGCAG CTGCAGCTGC GGCTGCGGCG GCTGGAAGTG GAGTTACGAG CGCCTCCGCT GCATCCGAGG GGGCGATCTC GGTAACAAAC CACAGGGGTT
                                                                                                              840
heAspLeuAs pLeuAsnLeu ProAlaMetP roGluPheAl aAlaPheGlu AlaSerProA laAlaArgAr gCysValAla AlaProAspG luGluGluVa lGlnSerPro LeuAlaPheL
```

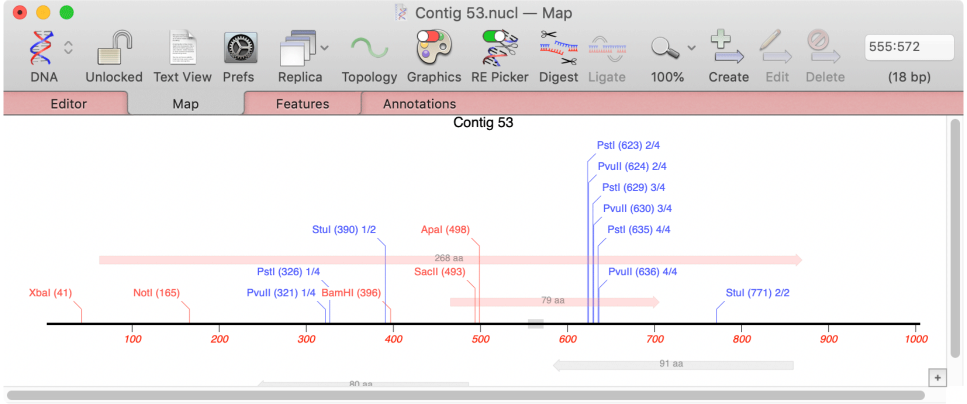
Now you can see that the match is to a potential translation on the plus strand (`GlnAlaLeuGlyGlyHis`). If you then click on **Find Next** in the **Find** dialog, a second domain should be selected;



```
uProHisArg ProGlyAlaA rgProAlaHi sArgLysSer SerAspAlaL euThrAla** ****ArgArg SerAspValS erCysSerGl yAspGluGly GlyCysArgA spAsp***Al
AlaSerSerP roTrpArgAl aAlaSerPro ProGlnIleP he***CysPr oHisArgLeu MetThrSerI le***ArgGl nLeuGlnArg ArgArgArgA rgLeuProAr gArgLeuSer
GCCTCATCGC CCTGGCGCGC GGCCAGCCCA CCGCAAATCT TCTGATGCCC TCACCGCCTG ATGACGTCGA TCTGACGTCA GCTGCAGCGG AGACGAAGGC GGCTGCCGCG ACGACTGAGC
                                                                                                                                 480
lnHisTyrLy sCysSerVal CysGlyLysA laPheGlySe rTyrGlnAla LeuGlyGlyH isLysAlaSe rHisLysLys LeuValLeuP roProAlaSe rAlaAspAsp GlnHisSerA
aAlaLeuGln ValLeuGlyL euArgLysGl yLeuArgIle LeuProSerA laArgArgSe rGlnGlyGln ProGlnLysA laArgProSe rThrArgLeu ArgArg***P roAlaGlnAr
SerThrThrS erAlaArgSe rAlaGluArg ProSerAspP roThrLysAr gSerAlaVal ThrArgProA laThrLysSe rSerSerPhe HisProProP roProMetTh rSerThrAla
AGCACTACAA GTGCTCGGTC TGCGGAAAGG CCTTCGGATC CTACCAAGCG CTCGGCGGTC ACAAGGCCAG CCACAAAAAG CTCGTCCTTC CACCCGCCTC CGCCGATGAC CAGCACAGCG
                                                                                                                                 600
laSerSerTh rAlaGlyPro ThrSerGlyA rgValHisGl nCysSerVal CysLeuLysT hrPheAlaSe rGlyGlnAla LeuGlyGlyH isLysArgCy sHisTyrAsp GlyThrIleG
gLeuIleAsp ArgGlyProH isGluArgGl nGlyProPro ValLeuCysL euProGlnAs pPheArgVal GlySerGlyA laGlyArgPr oGlnAlaVal ProLeu***A rgAspAspAr
ProHisArgP roArgAlaPr oArgAlaAla GlySerThrS erAlaLeuPh eAlaSerArg LeuSerArgA rgValArgAr gTrpAlaAla ThrSerGlyA laThrMetTh rGlyArgSer
CCTCATCGAC CGCGGGCCCC ACGAGCGGCA GGGTCCACCA GTGCTCTGTT TGCCTCAAGA CTTTCGCGTC GGGTCAGGCG CTGGGCGGCC ACAAGCGGTG CCACTATGAC GGGACGATCG
                                                                                                                                 720
lySerGlyTh rGlyValAla AlaAlaAlaA laAlaAlaAl aAlaAlaAla AlaGlySerG lyValThrSe rAlaSerAla AlaSerGluG lyAlaIleSe rValThrAsn HisArgGlyP
gGluArgAsn TrpSerCysC ysSerCysSe rCysSerCys GlyCysGlyG lyTrpLysTr pSerTyrGlu ArgLeuArgC ysIleArgGl yGlyAspLeu GlyAsnLysP roGlnGlyPh
GlyAlaGluL euGluLeuLe uGlnLeuGln LeuGlnLeuA rgLeuArgAr gLeuGluVal GluLeuArgA laProProLe uHisProArg GlyArgSerA rg***GlnTh rThrGlyVal
GGAGCGGAAC TGGAGTTGCT GCAGCTGCAG CTGCAGCTGC GGCTGCGGCG GCTGGAAGTG GAGTTACGAG CGCCTCCGCT GCATCCGAGG GGGCGATCTC GGTAACAAAC CACAGGGGTT
                                                                                                                                 840
heAspLeuAs pLeuAsnLeu ProAlaMetP roGluPheAl aAlaPheGlu AlaSerProA laAlaArgAr gCysValAla AlaProAspG luGluGluVa lGlnSerPro LeuAlaPheL
```

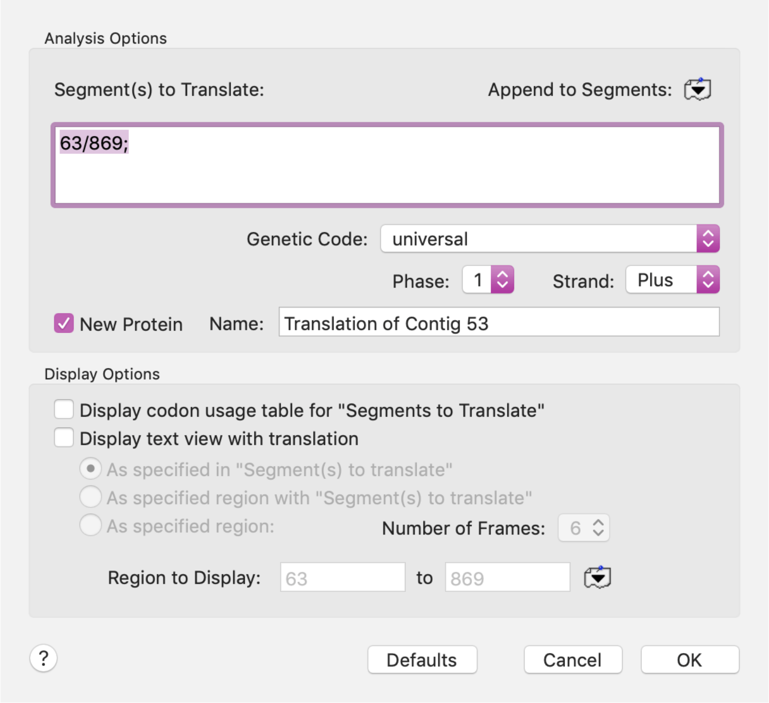
This demonstrates that `Contig 53` has not just one but TWO potential QALGGH domains. But, are they actually in a potential protein coding open reading frame?

To evaluate this, we need to look at the potential open reading frames in the contig. There are a number of approaches we can used for this. However, we have to be aware that the contig may not contain the entire coding region of the protein. In fact, it may not even include the AUG start codon. Click on the **Map** tab to view the graphical map of the sequence (note that this can also be done directly from the Contig **Editor** window);



Here we can see a pink arrow representing a potential protein-coding open reading frame on the plus strand.

Let's create a new protein sequence with the translation. Click on the pink arrow to select it, then choose **Analyze | Translation**... and click on the **New Protein** checkbox.



Analysis Options

Segment(s) to Translate: 63/869;

Append to Segments:

Genetic Code: universal

Phase: 1 Strand: Plus

☑ New Protein Name: Translation of Contig 53

Display Options

☐ Display codon usage table for "Segments to Translate"

☐ Display text view with translation

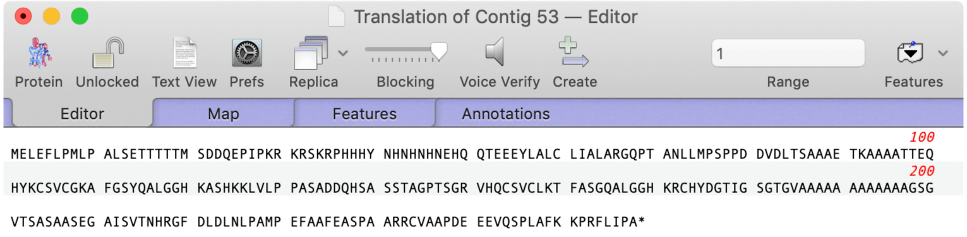
● As specified in "Segment(s) to translate"

○ As specified region with "Segment(s) to translate"

○ As specified region: Number of Frames: 6

Region to Display: 63 to 869

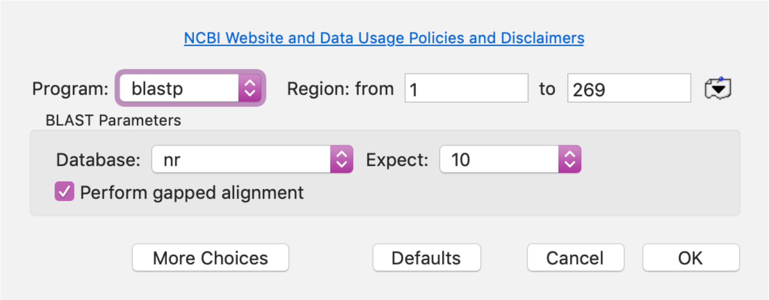
Note how the **Segment(s) to Translate** edit box is already set to the appropriate range and the **Strand** is set to **Plus**. When you click **OK**, a new protein sequence is created.



```
                                                                                                100
MELEFLPMLP ALSETTTTTM SDDQEPIPKR KRSKRPHHHY NHNHNHNEHQ QTEEEYLALC LIALARGQPT ANLLMPSPPD DVDLTSAAAE TKAAAATTEQ
                                                                                                200
HYKCSVCGKA FGSYQALGGH KASHKKLVLP PASADDQHSA SSTAGPTSGR VHQCSVCLKT FASGQALGGH KRCHYDGTIG SGTGVAAAAA AAAAAAAGSG

VTSASAASEG AISVTNHRGF DLDLNLPAMP EFAAFEASPA ARRCVAAPDE EEVQSPLAFK KPRFLIPA*
```

Now we can use **Edit | Find | Find**... to once again search for QALGGH, but this time the source is a protein sequence. You should be able to see that there are two QALGGH domains in the protein.

Finally, lets run an online BLAST search to confirm that this is indeed a zinc finger protein. With `Translation of Contig 20` the active sequence, choose **Database | Online Search for Similar Sequences (BLAST)**.



NCBI Website and Data Usage Policies and Disclaimers

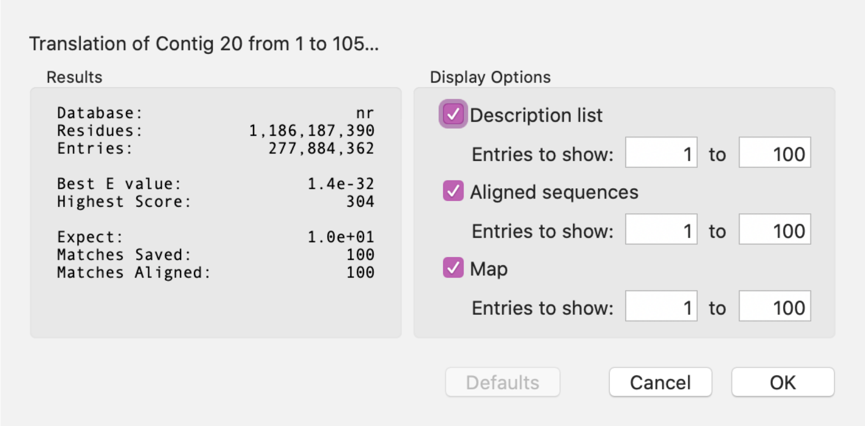
Program: blastp Region: from 1 to 269

BLAST Parameters

Database: nr Expect: 10

☑ Perform gapped alignment

In this case, the default settings are fine, so click **OK** to initiate the search. This runs an external search at the NCBI in Maryland, USA. Sometimes this completes within a few seconds, other times it may take a number of minutes depending on the server load and other factors. Once complete, press the **View** button if the result dialog does not automatically appear.



Translation of Contig 20 from 1 to 105...

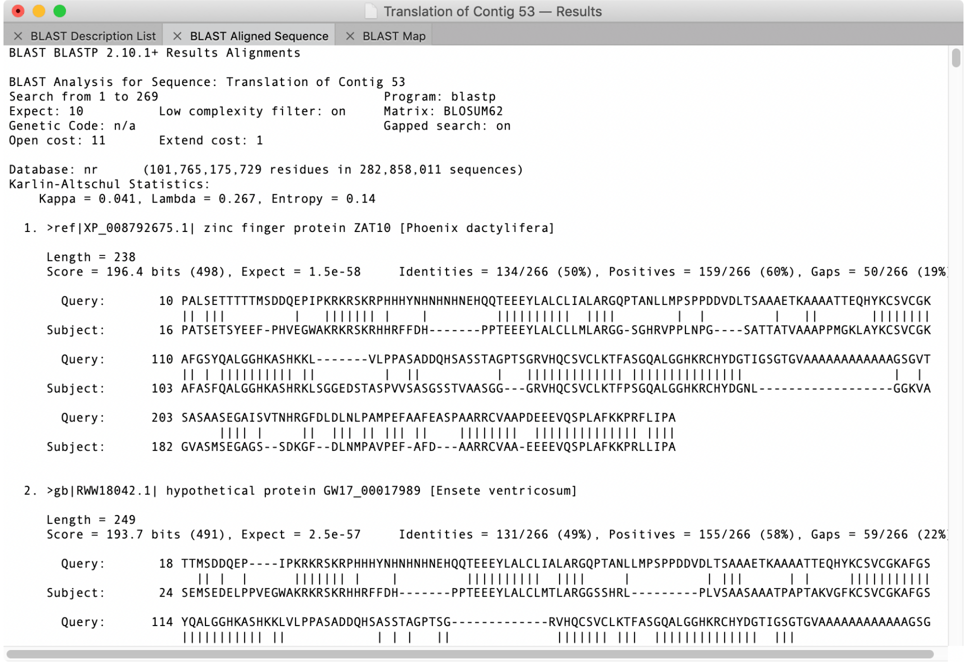
Results

```
Database:                     nr
Residues:          1,186,187,390
Entries:             277,884,362

Best E value:            1.4e-32
Highest Score:               304

Expect:                  1.0e+01
Matches Saved:               100
Matches Aligned:             100
```

Display Options

☑ Description list
Entries to show: 1 to 100
☑ Aligned sequences
Entries to show: 1 to 100
☑ Map
Entries to show: 1 to 100

Select all three **Display Options** and click **OK** to view the alignments.



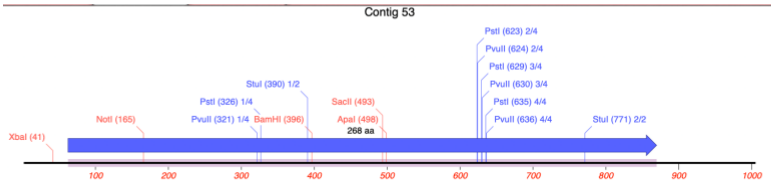
```
BLAST BLASTP 2.10.1+ Results Alignments

BLAST Analysis for Sequence: Translation of Contig 53
Search from 1 to 269                              Program: blastp
Expect: 10          Low complexity filter: on     Matrix: BLOSUM62
Genetic Code: n/a                                 Gapped search: on
Open cost: 11       Extend cost: 1

Database: nr      (101,765,175,729 residues in 282,858,011 sequences)
Karlin-Altschul Statistics:
    Kappa = 0.041, Lambda = 0.267, Entropy = 0.14

  1. >ref|XP_008792675.1| zinc finger protein ZAT10 [Phoenix dactylifera]

     Length = 238
     Score = 196.4 bits (498), Expect = 1.5e-58     Identities = 134/266 (50%), Positives = 159/266 (60%), Gaps = 50/266 (19%

       Query:       10 PALSETTTTTMSDDQEPIPKRKRSKRPHHHYNHNHNHNEHQQTEEEYLALCLIALARGQPTANLLMPSPPDDVDLTSAAAETKAAAATTEQHYKCSVCGK
                       || |||         |   ||||||| |    |         ||||||||||  ||||        |  |         |   ||       ||||||||
     Subject:       16 PATSETSYEEF-PHVEGWAKRKRSKRHHRFFDH-------PPTEEEYLALCLLMLARGG-SGHRVPPLNPG----SATTATVAAAPPMGKLAYKCSVCGK

       Query:      110 AFGSYQALGGHKASHKKL-------VLPPASADDQHSASSTAGPTSGRVHQCSVCLKTFASGQALGGHKRCHYDGTIGSGTVAAAAAAAAAAAAAGSGVT
                       || | |||||||||| ||         |  ||          |   ||||||||||||| |||||||||||||||                    |  |
     Subject:      103 AFASFQALGGHKASHRKLSGGEDSTASPVVSASGSSTVAASGG---GRVHQCSVCLKTFPSGQALGGHKRCHYDGNL------------------GGKVA

       Query:      203 SASAASEGAISVTNHRGFDLDLNLPAMPEFAAFEASPAARRCVAAPDEEEVQSPLAFKKPRFLIPA
                            |||| |     ||  ||| || ||| ||    ||||||||  |||||||||||||| ||||
     Subject:      182 GVASMSEGAGS--SDKGF--DLNMPAVPEF-AFD---AARRCVAA-EEEEVQSPLAFKKPRLLIPA


  2. >gb|RWW18042.1| hypothetical protein GW17_00017989 [Ensete ventricosum]

     Length = 249
     Score = 193.7 bits (491), Expect = 2.5e-57     Identities = 131/266 (49%), Positives = 155/266 (58%), Gaps = 59/266 (22%

       Query:       18 TTMSDDQEP----IPKRKRSKRPHHHYNHNHNHNEHQQTEEEYLALCLIALARGQPTANLLMPSPPDDVDLTSAAAETKAAAATTEQHYKCSVCGKAFGS
                         || |  |      ||||||| |    |         ||||||||||  ||||     |          | |||      | |     |||||||||||
     Subject:       24 SEMSEDELPPVEGWAKRKRSKRHHRFFDH-------PPTEEEYLALCLMTLARGGSSHRL---------PLVSAASAAATPAPTAKVGFKCSVCGKAFGS

       Query:      114 YQALGGHKASHKKLVLPPASADDQHSASSTAGPTSG-------------RVHQCSVCLKTFASGQALGGHKRCHYDGTIGSGTVAAAAAAAAAAAAAGSG
                       ||||||||||| ||            | | |   ||              ||||||| |||  ||||||||||||||  |||
```

At the time this tutorial was written, the most significant hits were all zinc finger proteins from *Phoenix dactylifera* (Date Palm), *Ensete ventricosum* (Ethiopian banana) and *Musa balbisiana* (Plantain), indicating we have likely “cloned” a zinc finger protein from *Aloe vera*.

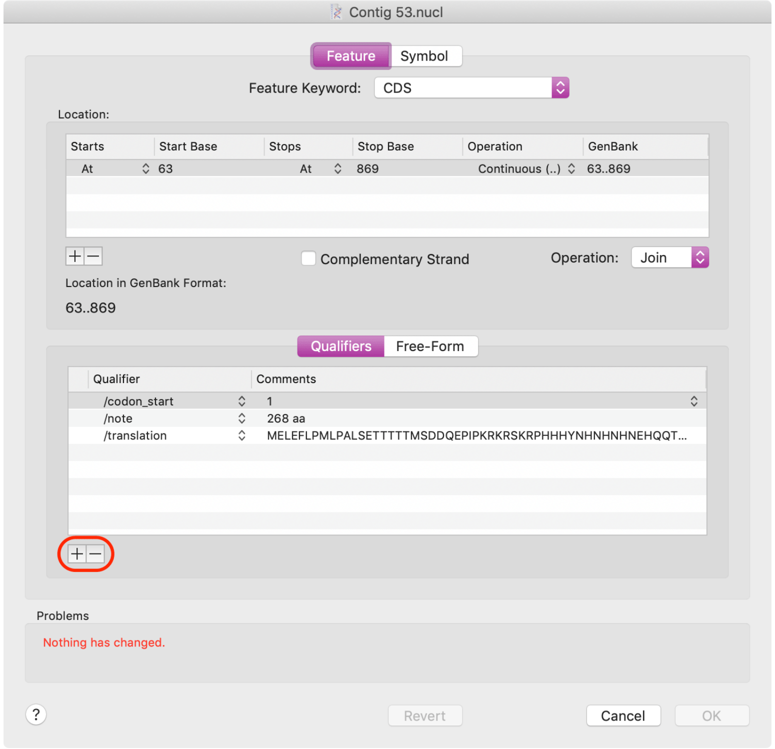
Annotating the Sequences

Let’s now add some simple annotation to the sequences.

Open `Contig 53.nucl`. Click on the single long plus strand ORF to select it. Then right-click on the selected graphic (or `<ctrl>`-click if you don’t have a right mouse button) and choose **Create CDS Feature** from the popup menu. This creates a new CDS feature which appears by default as a blue arrow.



Double-click on the new annotation and switch to the **Feature** tab.


Contig 53.nucl

Feature Keyword: CDS

Location:

Starts	Start Base	Stops	Stop Base	Operation	GenBank
At	63	At	869	Continuous (..)	63..869

☐ Complementary Strand Operation: Join

Location in GenBank Format:

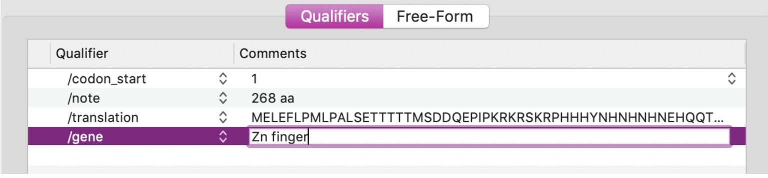
63..869

Qualifier	Comments
/codon_start	1
/note	268 aa
/translation	MELEFLPMLPALSETTTTTMSDDQEPIPKRKRSKRPHHHYNHNHNHNEHQQT...

Problems

Nothing has changed.

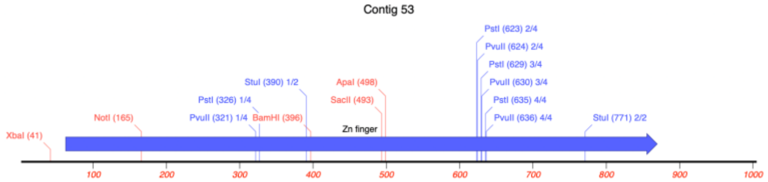
Note that MacVector has already added a */translation* qualifier, a */codon_start* qualifier and a */note* that lists the length of the coding region. Let's add some additional common qualifiers. Click on the **+** button in the lower left corner. This creates a new qualifier using */note*. Click on the new */note* entry and a popup menu will appear listing all of the valid qualifiers for a *CDS* feature. Choose */gene*, then click in the adjacent *Comments* field and type in "Zn Finger".



Qualifier	Comments
/codon_start	1
/note	268 aa
/translation	MELEFLPMLPALSETTTTTMSDDQEPIPKRKRSKRPHHHYNHNHNHNEHQQT...
/gene	Zn finger

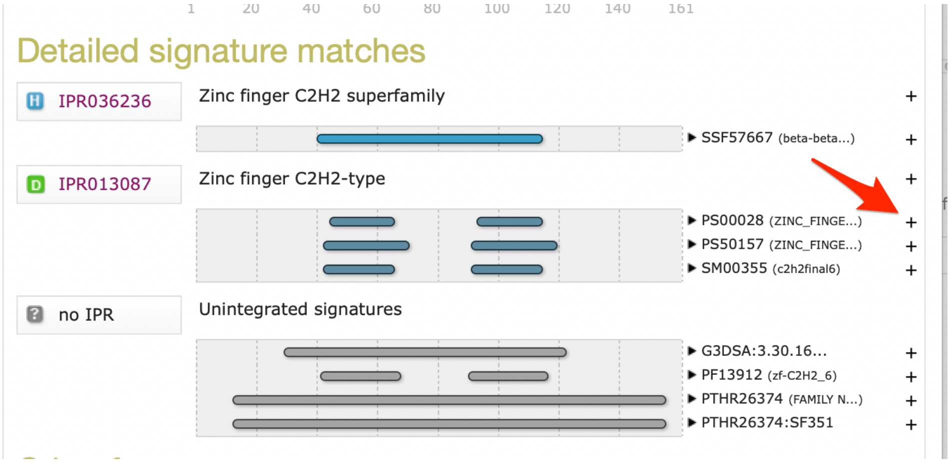
Another common qualifier is */product*, so we can add a new */product* qualifier with the value "Potential C2H2 zinc finger protein".

The **Map** tab will update to use the /gene qualifier as the new label;

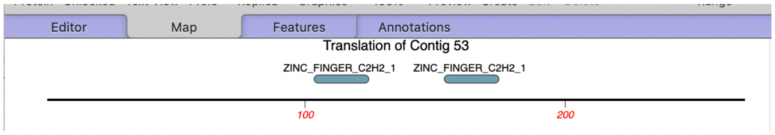


Next, lets annotate the protein. Open `Contig 53.prot` and choose **Database | Online Functional Domain Analysis (InterPro)**. This is an online service that will identify domains and other signatures in protein sequences hosted by the European Bioinformatics Institute. The first time you invoke the search you will be prompted to enter an e-mail address. This is purely to help reduce robot access and to help keep track of unique usage of the service for funding purposes– the EBI does NOT use your e-mail address for any marketing campaigns.

When complete (runs typically take from 30 seconds to 5 minutes) an interactive graphical results window appears.



There are many links in the window that you can click on to learn more about the domains, features and signatures in the protein sequence. These links will open in your default Internet Browser. You can add any identified domains as annotations to the protein by clicking on the small **+** button at the right-hand side of the window. Once clicked on, the **+** will change to a check mark. The domains will then appear in the **Map** tab.



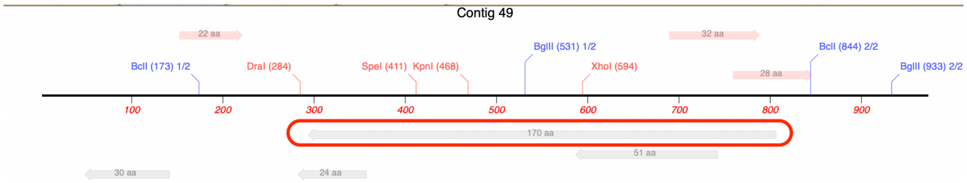
For this tutorial, we will annotate all of the proteins by clicking on the + button next to the PS00028 match to ensure all proteins are annotated consistently.

Identifying Additional Zinc Finger Encoding Genes

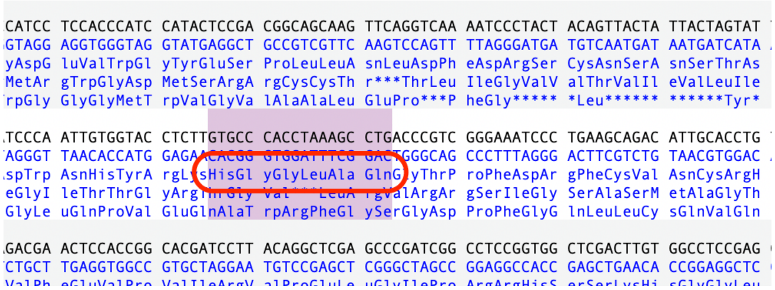
We can repeat the above steps using each of the other contigs from the initial assembly. Many will proceed exactly as described for `Contig 53`, but others may have issues that require additional analysis. Let's take a look at some of these cases.

Continuing down the list of contigs, sorted by length, `Contig 58` has a single large ORF on the plus strand and can be analyzed and annotated just like `Contig 53`.

`Contig 49` is a little different. In this case, the long ORF shows on the minus strand;

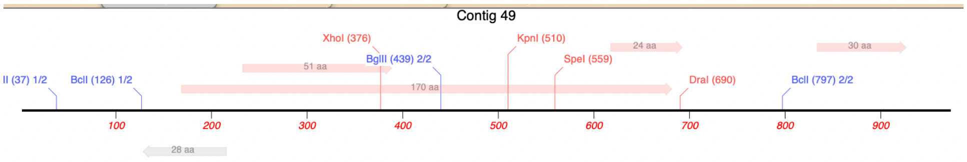


When running the Find for QALGGH you'll need to turn on the **Show Complementary Strand** and **Show Minus Strand Translations** from the **Display** toolbar button to see the `GlnAlaLeuGlyGlyHis` domains;



```
CATCC TCCACCCATC CATACTCCGA CGGCAGCAAG TTCAGGTCAA AATCCCTACT ACAGTTACTA TTACTAGTAT
GTAGG AGGTGGGTAG GTATGAGGCT GCCGTCGTTC AAGTCCAGTT TTAGGGATGA TGTCAATGAT AATGATCATA
yAspG luValTrpGl yTyrGluSer ProLeuLeuA snLeuAspPh eAspArgSer CysAsnSerA snSerThrAs
MetAr gTrpGlyAsp MetSerArgA rgCysCysTh r***ThrLeu IleGlyValV alThrValIl eValLeuIle
rpGly GlyGlyMetT rpValGlyVa lAlaAlaLeu GluPro***P heGly***** *Leu****** ******Tyr*

TCCCA ATTGTGGTAC CTCTTGTGCC CACCTAAAGC CTGACCCGTC GGGAAATCCC TGAAGCAGAC ATTGCACCTG
AGGGT TAACACCATG GAGAACACGG GTGGATTTCG GACTGGGCAG CCCTTTAGGG ACTTCGTCTG TAACGTGGAC
spTrp AsnHisTyrA rgLysHisGl yGlyLeuAla GlnGlyThrP roPheAspAr gPheCysVal AsnCysArgH
eGlyI leThrThrGl yArgThrGly Val***LeuA rgValArgAr gSerIleGly SerAlaSerM etAlaGlyTh
GlyLe uGlnProVal GluGlnAlaT rpArgPheGl ySerGlyAsp ProPheGlyG lnLeuLeuCy sGlnValGln

GACGA ACTCCACCGG CACGATCCTT ACAGGCTCGA GCCCGATCGG CCTCCGGTGG CTCGACTTGT GGCCTCCGAG
CTGCT TGAGGTGGCC GTGCTAGGAA TGTCCGAGCT CGGGCTAGCC GGAGGCCACC GAGCTGAACA CCGGAGGCTC
ValPh eGluValPro ValIleArgV alProGluLe uGlyIlePro ArgArgHisS erSerLysHi sGlyGlyLeu
```

You can still directly translate the ORF from the minus strand by selecting the gray arrow and choosing **Analyze | Translation**... as MacVector will automatically identify that the selection is a minus strand ORF and set the parameters appropriately. However, its often easier to work with coding regions on the plus strand, so we can "flip" the sequence. If you are working in the single nucleic acid sequence editor, choose **Edit | Select All** followed by **Edit | Reverse & Complement** to reverse and complement the entire sequence. If you are working in the contig **Editor** window, simply choose **Edit | Reverse & Complement**.



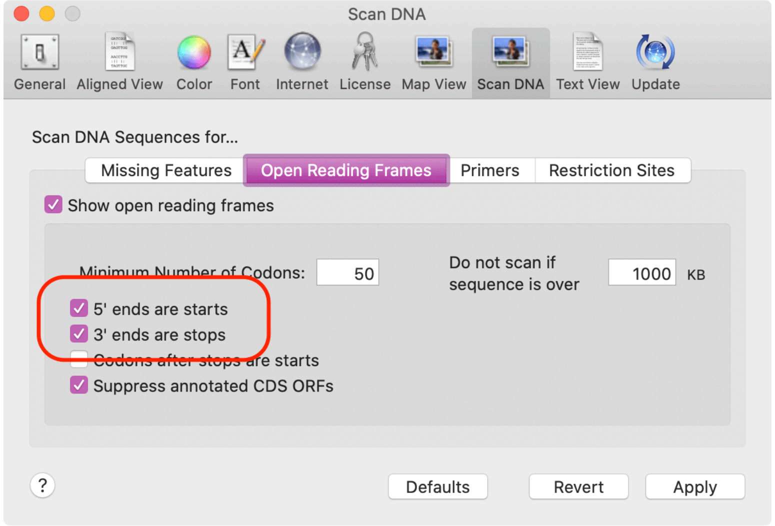
There is no need for you to perform this analysis yourself on all of the contigs as we've already done the hard work for you. You can download all of the annotated contigs and translated proteins from this link https://macvector.net/ContigsWith2Domains.zip..

For reference, the following Contigs have a single long zinc finger ORF on the plus strand: Contig 53, Contig 54, Contig 43, Contig 3, and these contigs have a single long zinc finger ORF on the minus strand: Contig 58, Contig 60, Contig 59, Contig 56, Contig 52, Contig 57, Contig 51, Contig 44, Contig 42, Contig 46, Contig 48. Other contigs require additional analysis, as described below.

Analysis of Contig 34

`Contig 34` (548 nt in length) shows no long ORFs in the **Map** tab, but the contig has two QALGGH coding domains on the minus strand. So, it seems like it should encode a zinc finger protein. One likely possibility is that we simply do not have reads to cover the entire coding region (this contig has just 22 reads whereas many of the others had hundreds or even thousands). It may also be that there is a long ORF that crosses through the sequence but is missing a suitable start and/or stop codon.

By default, MacVector scans all opened DNA sequences for the presence of open reading frames. However, because the start and stop codons may not be present in our truncated sequence, we need to adjust the defaults. Choose **MacVector | Preferences**... and switch to the **Scan DNA | Open Reading Frames** tab;



Scan DNA Sequences for...

Missing Features | **Open Reading Frames** | Primers | Restriction Sites

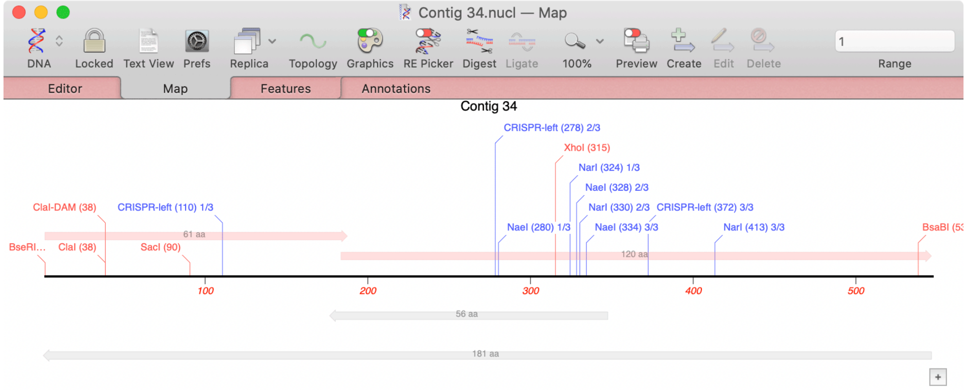
☑ Show open reading frames

Minimum Number of Codons: 50

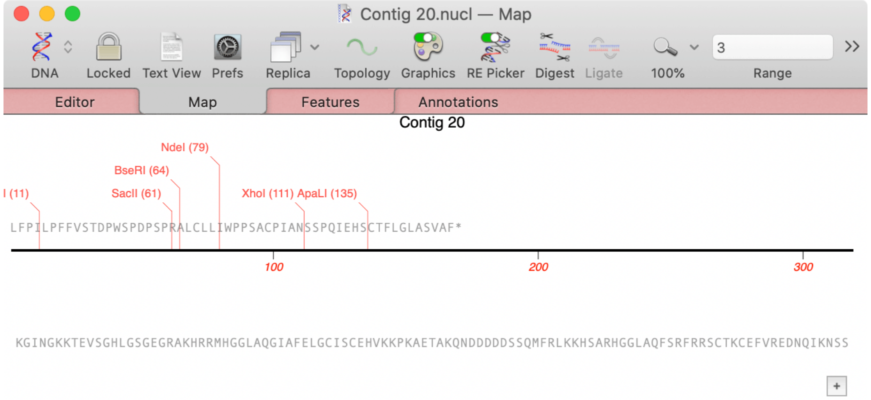
Do not scan if sequence is over 1000 KB

☑ 5' ends are starts
☑ 3' ends are stops
□ Codons after stops are starts
☑ Suppress annotated CDS ORFs

Select the **5' ends are starts** and **3' ends are stops** checkboxes then click **Apply**. This will allow the algorithm to identify ORFs that start and stop outside of the sequenced region. If we look at the `Contig 34.nucl` **Map** we should see that we do in fact have some open reading frames and that there is one on the minus strand that extends the full length of the sequence.



Note that if you have a sequence that is relatively short, or if your window is sized a bit larger than shown here, you may see the actual translation rather than just a grey arrow e.g.



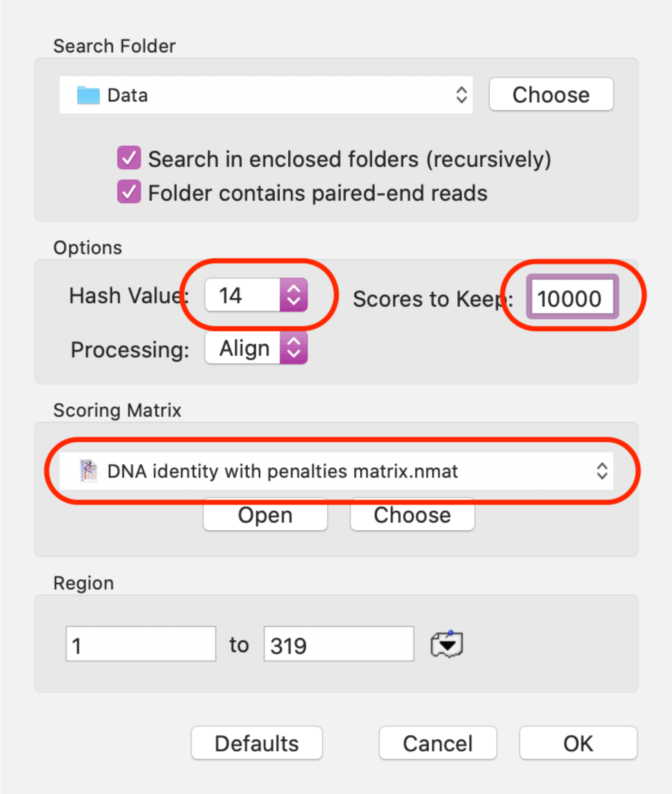
```
LFPILPFFVSTDPWSPDPSPRALCLLIWPPSACPIANSSPQIEHSCTFLGLASVAF*
```

```
KGINGKKTEVSGHLGSGEGRAKHRRMHGGLAQGIAFELGCISCEHVKKPKAETAKQNDDDDDSSQMFRLKKHSARHGGLAQFSRFRRSCTKCEFVREDNQIKNSS
```

Extending the Sequence of Contig 34

The next step is to try to extend the sequence of `Contig 34` to include the entire coding region of the zinc finger protein. Again, we can do this using **Align to Folder**. Here, we are hoping to identify reads that overlap the ends of `Contig 34` and that do not actually include the QALGGH coding region.

With `Contig 34` active, choose **Database | Align to Folder**. This time we are searching starting with a DNA sequence, so we will use different parameters;



Search Folder

Data

☑ Search in enclosed folders (recursively)

☑ Folder contains paired-end reads

Options

Hash Value: 14

Scores to Keep: 10000

Processing: Align

Scoring Matrix

DNA identity with penalties matrix.nmat

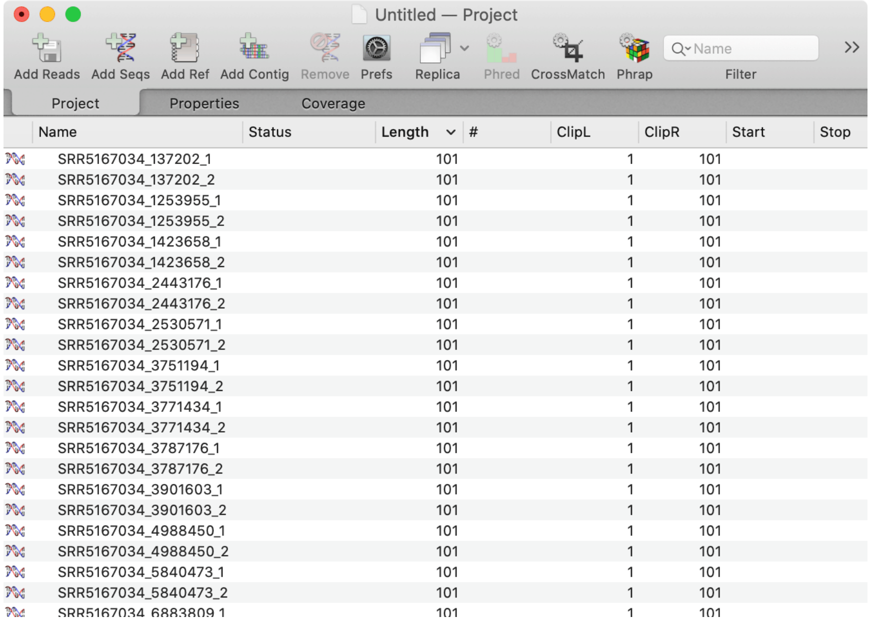
Region

1 to 319

The most critical change is that we will use the `DNA identity with penalties matrix.nmat` **Scoring Matrix**. To select this, click on the **Choose** button in the **Scoring Matrix** box and locate the file in `/Applications/MacVector/Scoring Matrices/`. This scoring matrix is tuned to identify perfect matches in the NGS reads, even if there is only a short overlap between the query sequence and a read. Make sure that your **Search Folder** is set to point to the fasta data and adjust the **Hash Value** to 12 or above (for speed) and the **Scores to Keep** to 10,000. Again, the search will take some time (3 to 8 hrs, depending on your machine) so you may want to let this go overnight.

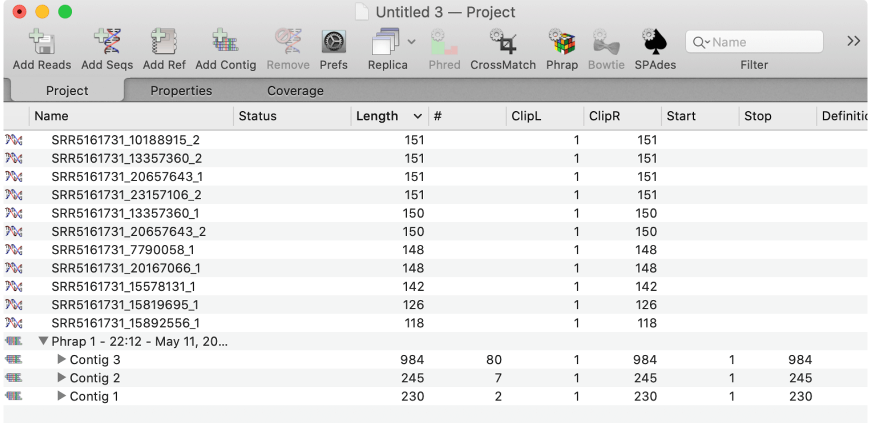
When the search completes, save all of the reads in the **Description List** tab – I called mine "`Contig34hits`".

Now we are ready to assemble them. Use **File | New | Assembly Project** to create a new assembly project. With smaller projects like this (~50 pairs of reads) it is generally better to use the **Add Seqs** button to add the reads to the project, rather than **Add Reads**. This adds the reads as individual sequences, rather than as a disk-based fastq collection.



Name	Status	Length	#	ClipL	ClipR	Start	Stop
SRR5167034_137202_1		101		1	101		
SRR5167034_137202_2		101		1	101		
SRR5167034_1253955_1		101		1	101		
SRR5167034_1253955_2		101		1	101		
SRR5167034_1423658_1		101		1	101		
SRR5167034_1423658_2		101		1	101		
SRR5167034_2443176_1		101		1	101		
SRR5167034_2443176_2		101		1	101		
SRR5167034_2530571_1		101		1	101		
SRR5167034_2530571_2		101		1	101		
SRR5167034_3751194_1		101		1	101		
SRR5167034_3751194_2		101		1	101		
SRR5167034_3771434_1		101		1	101		
SRR5167034_3771434_2		101		1	101		
SRR5167034_3787176_1		101		1	101		
SRR5167034_3787176_2		101		1	101		
SRR5167034_3901603_1		101		1	101		
SRR5167034_3901603_2		101		1	101		
SRR5167034_4988450_1		101		1	101		
SRR5167034_4988450_2		101		1	101		
SRR5167034_5840473_1		101		1	101		
SRR5167034_5840473_2		101		1	101		
SRR5167034_6883809_1		101		1	101		

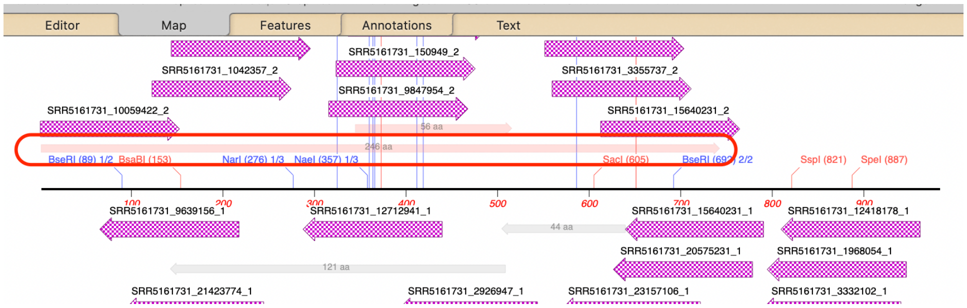
Again, we will use *phrap* to assemble the reads. You can either select all of the reads and then click on the **phrap** toolbar button, or you can make sure that no reads are selected and click on **phrap**. When nothing is selected, *phrap* operates on all of the reads in the project. Again, make sure you click on the **Short Read Defaults** option in the *phrap* dialog to make sure you are using the appropriate settings. When complete, you may have to scroll down the project to see the results;



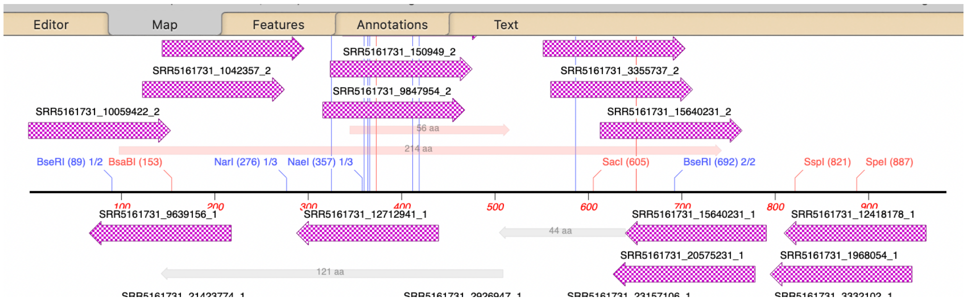
Name	Status	Length	#	ClipL	ClipR	Start	Stop	Definitio
SRR5161731_10188915_2		151		1	151			
SRR5161731_13357360_2		151		1	151			
SRR5161731_20657643_1		151		1	151			
SRR5161731_23157106_2		151		1	151			
SRR5161731_13357360_1		150		1	150			
SRR5161731_20657643_2		150		1	150			
SRR5161731_7790058_1		148		1	148			
SRR5161731_20167066_1		148		1	148			
SRR5161731_15578131_1		142		1	142			
SRR5161731_15819695_1		126		1	126			
SRR5161731_15892556_1		118		1	118			
▼ Phrap 1 - 22:12 - May 11, 20...								
▶ Contig 3		984	80	1	984	1	984	
▶ Contig 2		245	7	1	245	1	245	
▶ Contig 1		230	2	1	230	1	230	

Here we can see that 3 contigs were generated. `Contig 3` is 984 nt in length and contains 80 reads. This is undoubtedly the extended version of our original `Contig 34`. The other contigs are much shorter and only contain a few reads. If you open the short `Contig 1` and run a BLAST search you will find that it does appear to encode a short section of a QALGGH zinc finger protein. It is presumably the result of a few reads that came from a different zinc finger protein than `Contig 34` but were retrieved because they match the QALGGH domain. We will ignore this and focus on `Contig 3`.

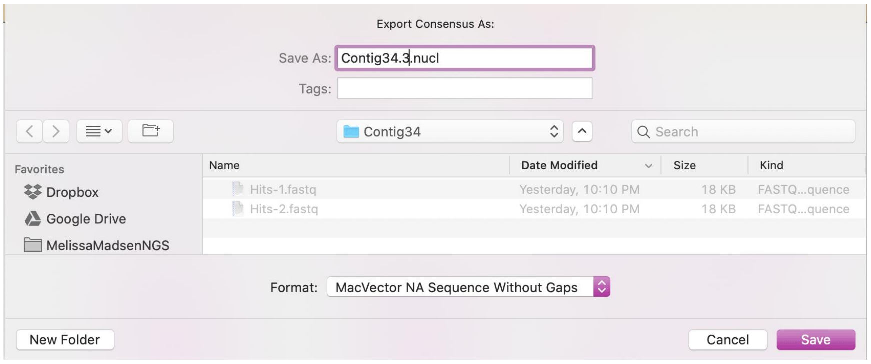
Double-click to open `Contig 3`. If you look at the **Map** tab, you'll see that there is a long ORF extending in from the 5' end of the plus strand.



The ORF appears to terminate at a stop codon and, if you return to the **Preferences | Scan DNA | Open Reading Frames** tab and turn off the **5' ends are starts** and **3' ends are stops** checkboxes, you will find that the ORF remains and appears to have a reasonable start codon ~100 nt into the contig.



Lets save `Contig 3` – the easiest way is to move to the `Contig 3` **Editor** tab and then select **File | Export Consensus As...** and choose *MacVector NA Sequence Without Gaps* in the **Format** field;


Export Consensus As:

Save As: Contig34.3.nucl

Tags:

Contig34

Name	Date Modified	Size	Kind
Hits-1.fastq	Yesterday, 10:10 PM	18 KB	FASTQ...quence
Hits-2.fastq	Yesterday, 10:10 PM	18 KB	FASTQ...quence

Format: MacVector NA Sequence Without Gaps

To keep track of the different contigs we might generate in this analysis, its useful to name them appropriately - in this case we can save the sequence as `Contig34.3` in a `Contig 34` folder where we can keep all of the sequences related to this contig.

If we had not retrieved enough reads to completely cover the `Contig 34` coding sequence, we could keep repeating the **Database | Align to Folder** analysis using this new `Contig 34.3.nucl` sequence and assemble the resulting hits until we finally cover the entire coding region. You can also repeat searches using just a hundred residues or so from one end of the contig – these often complete much more quickly.

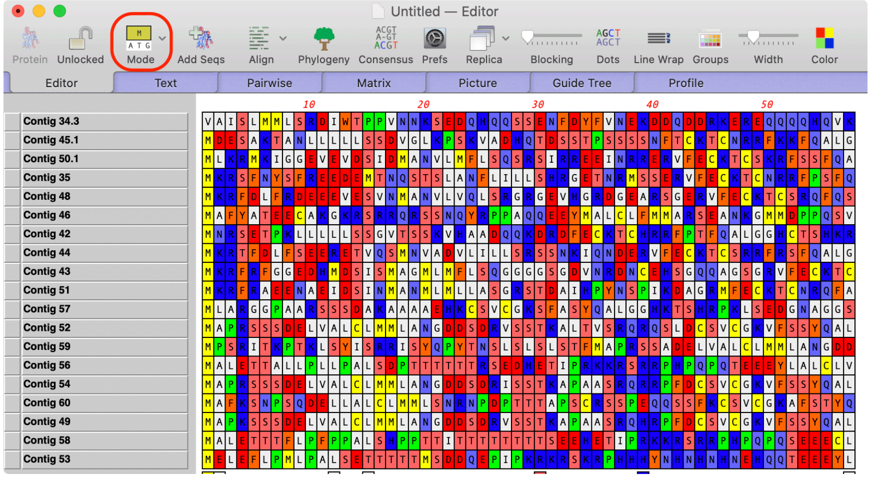
Select the pink open reading frame, choose **Analyze | Translation**... and create a new protein. If you BLAST the resulting 215 aa protein, you'll see that it has reasonable matches to other plant zinc finger proteins and that, most crucially, the numbering and length of the matching protein sequences are very similar, giving us confidence that we have indeed recovered the entire coding sequence.

Finally we can **File | Save As...** the protein as `Contig34.3.prot` and then run *InterProScan* on the protein to annotate the QALGGH domains as we did previously. You can use similar approaches to extend many of the other contigs from the initial assembly.

Aligning Proteins to View Common Functional Domains

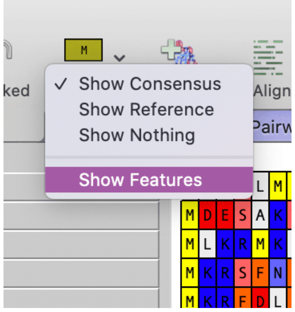
If you follow the above protocols for all of the contigs you will eventually find a total of 19 contigs that encode proteins containing two QALGGH domains (there are many more that contain just a single QALGGH domain). The data for this can be downloaded from https://macvector.net/ContigsWith2Domains.zip. Some immediate questions that come to mind are: (a) are these all unique proteins or are some duplicates, and (b) how are they all related? Each of the proteins in the sample set were annotated using *InterProScan* as described above, specifically choosing the PS00028 match. Let's align them and view the shared QALGGH domains;

First choose **File | New | Protein Alignment** to create a new MSAP multiple alignment window, then click on the **Add Seqs** toolbar button, select all of the files in the `/Proteins/` folder of the sample set and click on the **Open** button. A new MSAP window appears;

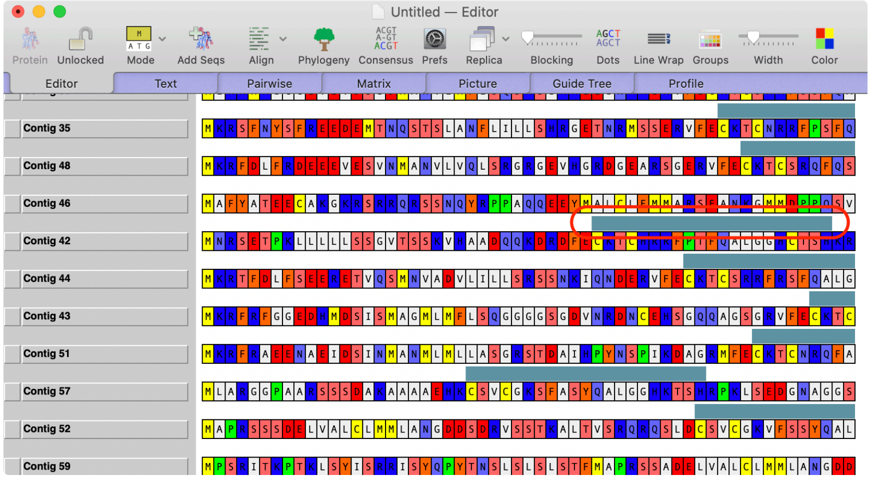


```
                   10        20        30        40        50
Contig 34.3  VAISLMMLSRDIWTPPVNNKSEDQHQQSSENFDYFVNEKDDQDDRKEREQQQQHQVK
Contig 45.1  MDESAKTANLLLLLSSDVGLKPSKVADHQTDSSTPSSSSNFTCKTCNRRFKKFQALG
Contig 50.1  MLKRMKIGGEVEVDSIDMANVLMFLSQSRSIRREEINRRERVFECKTCSKRFSSFQA
Contig 35    MKRSFNYSFREEDEMTNQSTSLANFLILLSHRGETNRMSSERVFECKTCNRRFPSFQ
Contig 48    MKRFDLFRDEEEVESVNMANVLVQLSRGRGEVHGRDGEARSGERVFECKTCSRQFQS
Contig 46    MAFYATEECAKGKRSRRQRSSNQYRPPAQQEEYMALCLFMMARSEANKGMMDPPQSV
Contig 42    MNRSETPKLLLLLSSGVTSSKVHAADQQKDRDFECKTCHRRFPTFQALGGHCTSHKR
Contig 44    MKRTFDLFSEERETVQSMNVADVLILLSRSSNKIQNDERVFECKTCSRRFRSFQALG
Contig 43    MKRFRFGGEDHMDSISMAGMLMFLSQGGGGSGDVNRDNCEHSGQQAGSGRVFECKTC
Contig 51    MKRFRAEENAEIDSINMANMLMLLASGRSTDAIHPYNSPIKDAGRMFECKTCNRQFA
Contig 57    MLARGGPAARSSSDAKAAAAEHKCSVCGKSFASYQALGGHKTSHRPKLSEDGNAGGS
Contig 52    MAPRSSSDELVALCLMMLANGDDSDRVSSTKALTVSRQRQSLDCSVCGKVFSSYQAL
Contig 59    MPSRITKPTKLSYISRRISYQPYTNSLSLSLSTFMAPRSSADELVALCLMMLANGDD
Contig 56    MALETTALLPLLPALSDPTTTTTTRSEDHETIPRKKRSRRPHPQPQTEEEYLALCLV
Contig 54    MAPRSSSDELVALCLMMLANGDDSDRISSTKAPAASRQRRPFDCSVCGKVFSSYQAL
Contig 60    MAFKSNPSQDELLALCLMMLSNRNPDPTTTAPSCRSSPEQQSSFKCSVCGKAFSTYQ
Contig 49    MAPKSSSDELVALCLMMLANGDDSDRVSSTKAPAASRQHRPFDCSVCGKVFSSYQAL
Contig 58    MALETTTFLPFPPALSHPPTTITTTTTTTTSEEHETIPRKKRSRRPHPQPQSEEECL
Contig 53    MELEFLPMLPALSETTTTTMSDDQEPIPKRKRSKRPHHHYNHNHNHNEHQQTEEEYL
```

Click on the **Mode** toolbar button and choose **Show Features** from the popup menu;

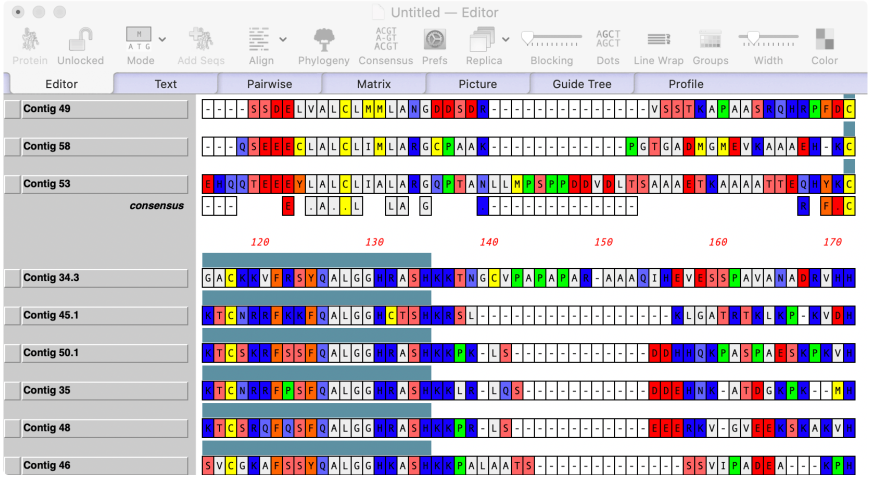


The window updates to show the features in the proteins as simple blocks above the residues;



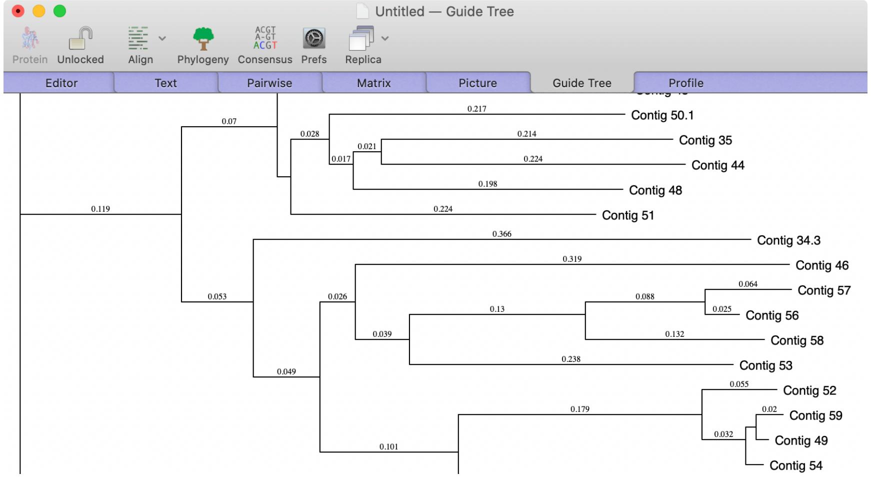
```
Contig 35   MKRSFNYSFREEDEMTNQSTSLANFLILLSHRGETNRMSSERVFECKTCNRRFPSFQ
Contig 48   MKRFDLFRDEEEVESVNMANVLVQLSRGRGEVHGRDGEARSGERVFECKTCSRQFQS
Contig 46   MAFYATEECAKGKRSRRQRSSNQYRPPAQQEEYMALCLEMMARSEANKGMMDPPQSV
Contig 42   MNRSETPKLLLLLSSGVTSSKVHAADQQKDRDFECKTCHRRFPTFQALGGHCTSHKR
Contig 44   MKRTFDLFSEERETVQSMNVADVLILLSRSSNKIQNDERVFECKTCSRRFRSFQALG
Contig 43   MKRFRFGGEDHMDSISMAGMLMFLSQGGGGSGDVNRDNCEHSGQQAGSGRVFECKTC
Contig 51   MKRFRAEENAEIDSINMANMLMLLASGRSTDAIHPYNSPIKDAGRMFECKTCNRQFA
Contig 57   MLARGGPAARSSSDAKAAAAEHKCSVCGKSFASYQALGGHKTSHRPKLSEDGNAGGS
Contig 52   MAPRSSSDELVALCLMMLANGDDSDRVSSTKALTVSRQRQSLDCSVCGKVFSSYQAL
Contig 59   MPSRITKPTKLSYISRRISYQPYTNSLSLSLSTFMAPRSSADELVALCLMMLANGDD
```

Align the sequences by clicking on the **Align** toolbar button and selecting **ClustalW** from the dropdown menu. Accept the defaults and click **OK**. After a few seconds, the display updates;

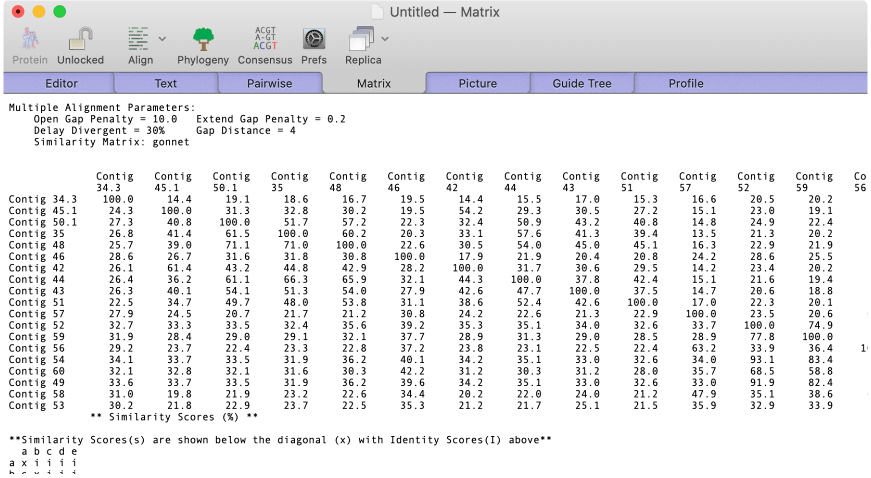


```
Contig 49   ----SSDELVALCLMMLANGDDSDR-------------VSSTKAPAASRQHRPFDC
Contig 58   ---QSEEECLALCLIMLARGCPAAK-------------PGTGADMGMEVKAAAEH-KC
Contig 53   EHQQTEEEYLALCLIALARGQPTANLLMPSPPDDVDLTSAAAETKAAAATTEQHYKC
  consensus ---    E .A..L  LA G    .-------------              R F.C

                 120       130       140       150       160       170

Contig 34.3 GACKKVFRSYQALGGHRASHKKTNGCVPAPAPAR-AAAQIHEVESSPAVANADRVHH
Contig 45.1 KTCNRRFKKFQALGGHCTSHKRSL-----------------KLGATRTKLKP-KVDH
Contig 50.1 KTCSKRFSSFQALGGHRASHKKPK-LS-------------DDHHQKPASPAESKPKVH
Contig 35   KTCNRRFPSFQALGGHRASHKKLR-LQS------------DDEHNK-ATDGKPK--MH
Contig 48   KTCSRQFQSFQALGGHRASHKKPR-LS-------------EEERKV-GVEEKSKAKVH
Contig 46   SVCGKAFSSYQALGGHKASHKKPALAATS-------------SSVIPADEA---KPH
```

If you click on the **Guide Tree** tab, you can see a very basic phylogenetic tree showing the relationships between the proteins. Note that this is NOT a true phylogeny, just a rough guide that *ClustalW* used to work out the order to assemble the multiple alignment. However, it clearly shows that none of the proteins are identical;

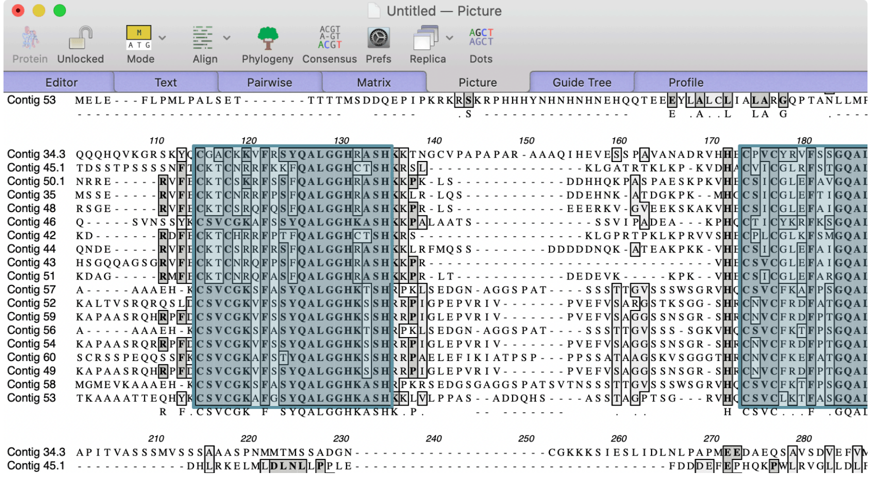


This can be seen in numerical form in the **Matrix** tab – this displays a table of the relatedness of each sequence in the alignment in a spreadsheet-like form. You can see that on the diagonal where each sequence is obviously identical to itself, but all of the other values are less than 100%, indicating that we have 19 distinct proteins;



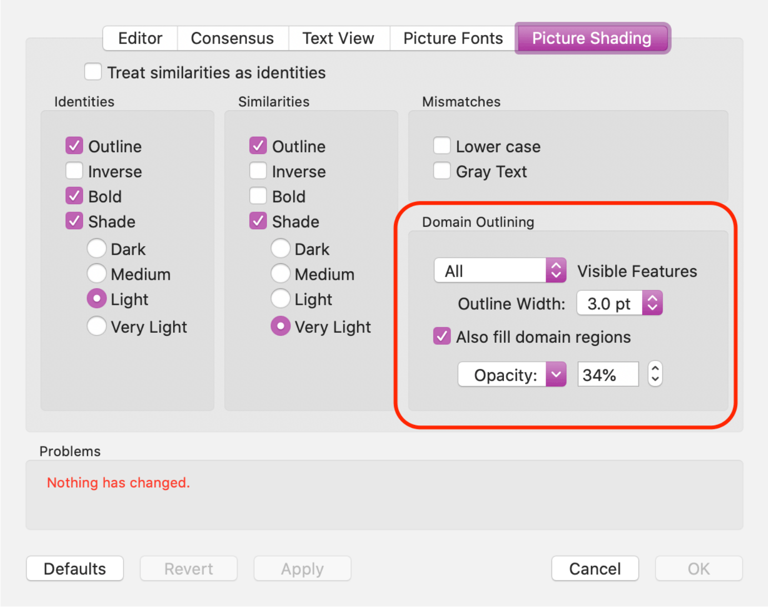
```
Multiple Alignment Parameters:
    Open Gap Penalty = 10.0   Extend Gap Penalty = 0.2
    Delay Divergent = 30%     Gap Distance = 4
    Similarity Matrix: gonnet


              Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Contig   Co
              34.3     45.1     50.1     35       48       46       42       44       43       51       57       52       59       56
Contig 34.3    100.0     14.4     19.1     18.6     16.7     19.5     14.4     15.5     17.0     15.3     16.6     20.5     20.2
Contig 45.1     24.3    100.0     31.3     32.8     30.2     19.5     54.2     29.3     30.5     27.2     15.1     23.0     19.1
Contig 50.1     27.3     40.8    100.0     51.7     57.2     22.3     32.4     50.9     43.2     40.8     14.8     24.9     22.4
Contig 35       26.8     41.4     61.5    100.0     60.2     20.3     33.1     57.6     41.3     39.4     13.5     21.3     20.2
Contig 48       25.7     39.0     71.1     71.0    100.0     22.6     30.5     54.0     45.0     45.1     16.3     22.9     21.9
Contig 46       28.6     26.7     31.6     31.8     30.8    100.0     17.9     21.9     20.4     20.8     24.2     28.6     25.5
Contig 42       26.1     61.4     43.2     44.8     42.9     28.2    100.0     31.7     30.6     29.5     14.2     23.4     20.2
Contig 44       26.4     36.2     61.1     66.3     65.9     32.1     44.3    100.0     37.8     42.4     15.1     21.6     19.4
Contig 43       26.3     40.1     54.1     51.3     54.0     27.9     42.6     47.7    100.0     37.5     14.7     20.6     18.8
Contig 51       22.5     34.7     49.7     48.0     53.8     31.1     38.6     52.4     42.6    100.0     17.0     22.3     20.1
Contig 57       27.9     24.5     20.7     21.7     21.2     30.8     24.2     22.6     21.3     22.9    100.0     23.5     20.6
Contig 52       32.7     33.3     33.5     32.4     35.6     39.2     35.3     35.1     34.0     32.6     33.7    100.0     74.9
Contig 59       31.9     28.4     29.0     29.1     32.1     37.7     28.9     31.3     29.0     28.5     28.9     77.8    100.0
Contig 56       29.2     23.7     22.4     23.3     22.8     37.2     23.8     23.1     22.5     22.4     63.2     33.9     36.4     1
Contig 54       34.1     33.7     33.5     31.9     36.2     40.1     34.2     35.1     33.0     32.6     34.0     93.1     83.4
Contig 60       32.1     32.8     32.1     31.6     30.3     42.2     31.2     30.3     31.2     28.0     35.7     68.5     58.8
Contig 49       33.6     33.7     33.5     31.9     36.2     39.6     34.2     35.1     33.0     32.6     33.0     91.9     82.4
Contig 58       31.0     19.8     21.9     23.2     22.6     34.4     20.2     22.0     24.0     21.2     47.9     35.1     38.6
Contig 53       30.2     21.8     22.9     23.7     22.5     35.3     21.2     21.7     25.1     21.5     35.9     32.9     33.9
             ** Similarity Scores (%) **

**Similarity Scores(s) are shown below the diagonal (x) with Identity Scores(I) above**
  a b c d e
a x i i i i
b c x i i i
```

Finally, if you click on the **Picture** tab, you should see the QALGGH "C2H2" zinc finger domains outlined something like this;



```
Contig 53    MELE---FLPMLPALSET-------TTTTMSDDQEPIPKRKRSKRPHHHYNHNHNHNEHQQTEEEYLALCLIALARGQPTANLLMF
             ------------------------------  -----    .S -----------------    E .A..L  LA G     .----

                     110       120       130       140       150       160       170       180
Contig 34.3  QQQHQVKGRSKYQCGACKKVFRSYQALGGHRASHKKTNGCVPAPAPAR-AAAQIHEVESSPAVANADRVHHECPVCYRVFSSGQAL
Contig 45.1  TDSSTPSSSSNFTCKTCNRRFKKFQALGGHCTSHKRSL-----------------KLGATRTKLKP-KVDHACVICGLRFSTGQAL
Contig 50.1  NRRE-----RVFECKTCSKRFSSFQALGGHRASHKKPK-LS------------DDHHQKPASPAESKPKVHECSICGLEFAVGQAL
Contig 35    MSSE-----RVFECKTCNRRFPSFQALGGHRASHKKLR-LQS-----------DDEHNK-ATDGKPK--MHCCSICGLEFTIGQAL
Contig 48    RSGE-----RVFECKTCSRQFQSFQALGGHRASHKKPR-LS------------EEERKV-GVEEKSKAKVHECSICGLEFAIGQAL
Contig 46    Q-----SVNSSYKCSVCGKAFSSYQALGGHKASHKKPALAATS-------------SSVIPADEA---KPHQCTICYKRFKSGQAL
Contig 42    KD-------RDFECKTCHRRFPTFQALGGHCTSHKRS------------------KLGPRTPKLKPRVVSHECPLCGLKFSMGQAL
Contig 44    QNDE-----RVFECKTCSRRFRSFQALGGHRASHKKLRFMQSS--------DDDDDNQK-ATEAKPKK-VHECSICGLEFAIGQAL
Contig 43    HSGQQAGSGRVFECKTCNRQFPSFQALGGHRASHKKPR-------------------------------VHECSVCGLEFAIGQAL
Contig 51    KDAG-----RMFECKTCNRQFASFQALGGHRASHKKPR-LT------------DEDEVK-----KPK--VHECSICGLEFARGQAL
Contig 57    A-----AAAEH-KCSVCGKSFASYQALGGHKTSHRPKLSEDGN-AGGSPAT----SSSTTGVSSSWSGRVHQCSVCFKAFPSGQAL
Contig 52    KALTVSRQRQSLDCSVCGKVFSSYQALGGHKSSHRRPIGPEPVRIV-------PVEFVSARGSTKSGG-SHRCNVCFRDFATGQAL
Contig 59    KAPAASRQHRPFDCSVCGKVFSSYQALGGHKSSHRRPIGLEPVRIV-------PVEFVSAGGSSNSGR-SHRCNVCFRDFPTGQAL
Contig 56    A-----AAAEH-KCSVCGKSFASYQALGGHKTSHRPKLSEDGN-AGGSPAT----SSSTTGVSSS-SGKVHQCSVCFKTFPSGQAL
Contig 54    KAPAASRQRRPFDCSVCGKVFSSYQALGGHKSSHRRPIGLEPVRIV-------PVEFVSAGGSSNSGR-SHRCNVCFRDFPTGQAL
Contig 60    SCRSSPEQQSSFKCSVCGKAFSTYQALGGHKSSHRRPAELEFIKIATPSP---PPSSATAAGSKVSGGGTHRCNVCFKEFATGQAL
Contig 49    KAPAASRQHRPFDCSVCGKVFSSYQALGGHKSSHRRPIGLEPVRIV-------PVEFVSAGGSSNSGR-SHRCNVCFRDFPTGQAL
Contig 58    MGMEVKAAAEH-KCSVCGKSFASYQALGGHKASHRPKRSEDGSGAGGSPATSVTNSSSTTGVSSSWSGRVHQCSVCFKTFPSGQAL
Contig 53    TKAAAATTEQHYKCSVCGKAFGSYQALGGHKASHKKLVLPPAS--ADDQHS----ASSTAGPTSG---RVHQCSVCLKTFASGQAL
                      R F.CSVCGK F SYQALGGHKASHK.P.     -- -------     . ..        H CSVC...F .GQAL

                     210       220       230       240       250       260       270       280
Contig 34.3  APITVASSSMVSSSAAASPNMMTMSSADGN---------------------CGKKKSIESLIDLNLPAPMEEDAEQSAVSDVEFVM
Contig 45.1  ------------DHLRKELMLDLNLLPPLE----------------------------------FDDDEFEPHQKPWLRVGLLDLF
```

If your display doesn't look exactly like this, click on the **Prefs** toolbar button. The shading and colored domain outlining are controlled by the **Picture Shading** tab;



□ Treat similarities as identities

Identities

☑ Outline
□ Inverse
☑ Bold
☑ Shade
○ Dark
○ Medium
● Light
○ Very Light

Similarities

☑ Outline
□ Inverse
□ Bold
☑ Shade
○ Dark
○ Medium
○ Light
● Very Light

Mismatches

□ Lower case
□ Gray Text

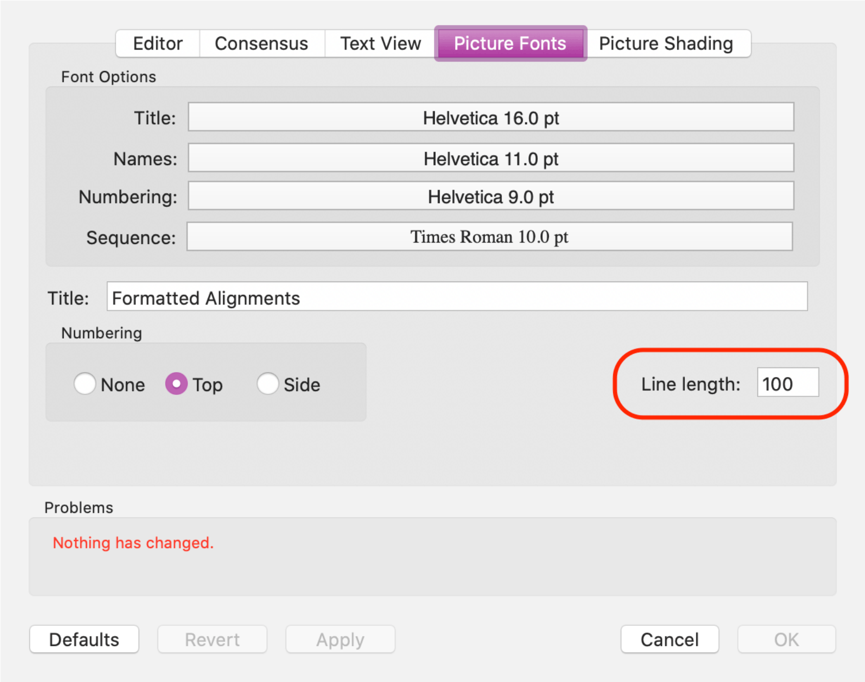
Domain Outlining

All Visible Features
Outline Width: 3.0 pt
☑ Also fill domain regions
Opacity: 34%

Problems

Nothing has changed.

The line length and other font information is controlled by the **Picture Fonts** tab.



Font Options

Title: Helvetica 16.0 pt
Names: Helvetica 11.0 pt
Numbering: Helvetica 9.0 pt
Sequence: Times Roman 10.0 pt

Title: Formatted Alignments

Numbering

○ None ● Top ○ Side

Line length: 100

Problems

Nothing has changed.

Conclusions

This tutorial demonstrates how you can use MacVector to “clone” interesting genes from RNA-Seq data in the NCBI short read archive. One key is the use of the **Database | Align to Folder** function to identify the few reads from a large data set that encode proteins of interest. This dramatically simplifies the assembly process, allowing even fairly weakly expressed genes to be retrieved and assembled. Without this enrichment step, it is likely that there would not be enough overlapping reads present to allow fast NGS assemblers like *Velvet* and *SPAdes* to successfully assemble all of the potential genes.