MacVector 17

for Mac OS X

RNASeq Gene Expression Analysis Tutorial

MacVector, Inc.

Software for Scientists

Copyright statement

This version of the RNASeq Analysis tutorial was published in January 2019.

Contents

Introduction

One common use of Next Generation Sequencing (NGS) technology is to analyze the relative expression levels of all known genes of an organism in a single experiment. mRNA is extracted from the organism and randomly sequenced using NGS to generate millions of reads. These can then be aligned against a sequenced genome to determine how many reads align to each known gene, resulting in data that can be used to estimate the relative expression levels of each gene.

MacVector incorporates the popular Bowtie algorithm which is a blazingly fast assembler than can align millions of reads to a reference genome in just a few minutes with minimal memory requirements. This tutorial shows you how to perform one of these analyses with sample data that is installed along with MacVector.

Sample Files

You can find the data files for this tutorial in the following location;

`/Applications/MacVector/Tutorial Files/Contig Assembly/RNASeq/`

The data can also be downloaded from;

https://macvector.com/downloads.html

Tutorial

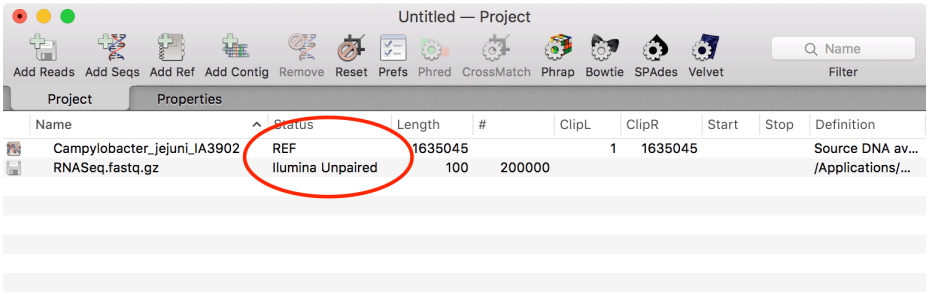
Create and Populate an Assembly Project

You must have the *Assembler* module enabled for this tutorial. You can check if you have an active *Assembler* license by choosing **MacVector | About MacVector**. You should see a "splash screen" something like this;

If the logo simply reads "*MacVector*" and not "*MacVector* with *Assembler*" then you do not have a current license to run *Assembler* and you should contact us at support@macvector.com to obtain a temporary license.

Select **File | New | Assembly Project** to create an empty Assembly Project document. Click on the Add Ref toolbar button, navigate to the `/Applications/MacVector/Tutorial Files/Contig Assembly/RNASeq/` folder and choose `Campylobacter jejuni IA 3902`. Then click on the Add Reads toolbar button and choose the file `RNASeq.fastq.gz` from the same folder.

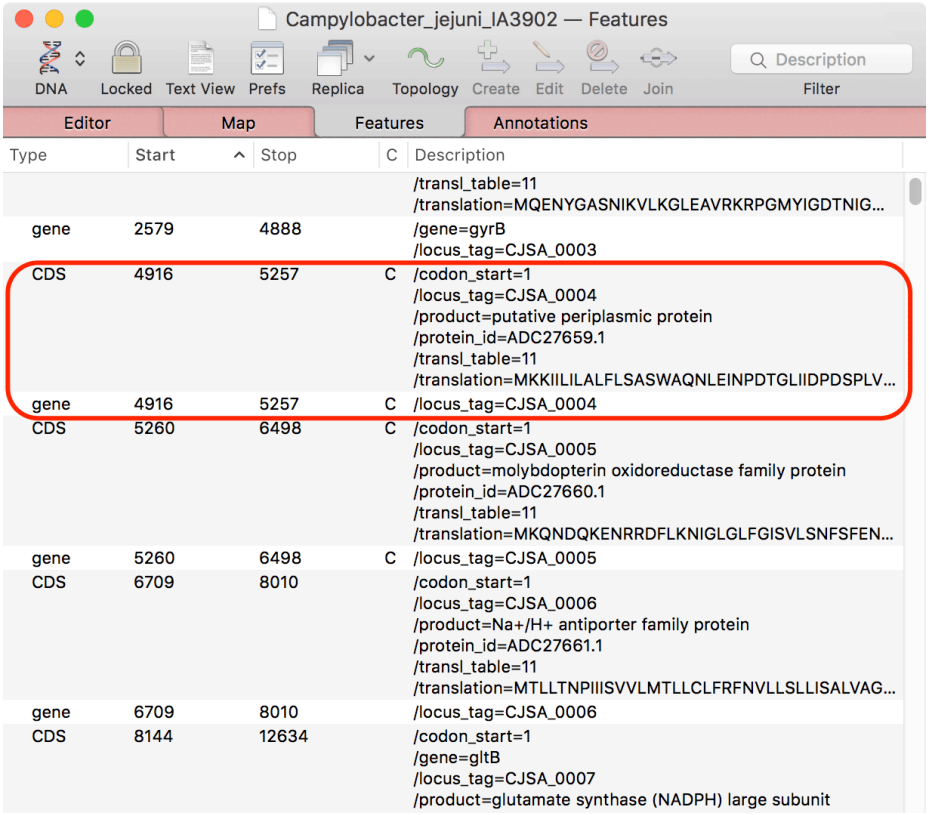
Your project should look like this;



Note that the Status column indicates the type of data contained in each imported file with ***REF*** indicating a reference sequence. By default, imported fastq files are assign as ***Illumina*** files (MacVector will also automatically identify and flag paired-end read files). If your data comes from a different source (IonTorrent, PacBio, Oxford Nanopore etc) you can simply double-click on the row to change the data source.

Also note that MacVector can directly read gzipped files (these typically have a .gz extension). There is no need to unzip the files prior to analysis. This can save a lot of disk space with large data sets.

Double-click on the *Campylobacter jejuni* item to open up a sequence document window. Switch to the Features tab.

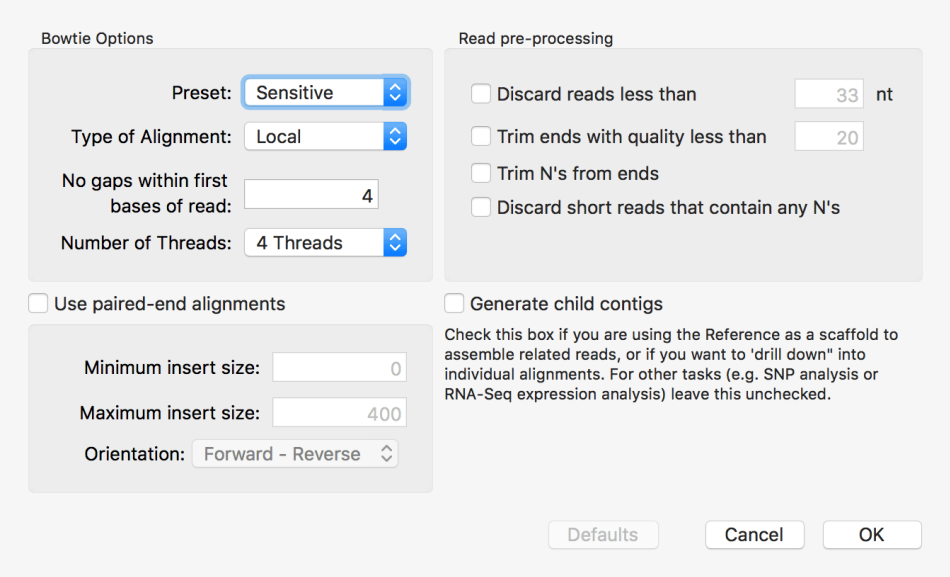


Type	Start	Stop	C	Description
				/transl_table=11 /translation=MQENYGASNIKVLKGLEAVRKRPGMYIGDTNIG...
gene	2579	4888		/gene=gyrB /locus_tag=CJSA_0003
CDS	4916	5257	C	/codon_start=1 /locus_tag=CJSA_0004 /product=putative periplasmic protein /protein_id=ADC27659.1 /transl_table=11 /translation=MKKIILILALFLSASWAQNLEINPDTGLIIDPDSPLV...
gene	4916	5257	C	/locus_tag=CJSA_0004
CDS	5260	6498	C	/codon_start=1 /locus_tag=CJSA_0005 /product=molybdopterin oxidoreductase family protein /protein_id=ADC27660.1 /transl_table=11 /translation=MKQNDQKENRRDFLKNIGLGLFGISVLSNFSFEN...
gene	5260	6498	C	/locus_tag=CJSA_0005
CDS	6709	8010		/codon_start=1 /locus_tag=CJSA_0006 /product=Na+/H+ antiporter family protein /protein_id=ADC27661.1 /transl_table=11 /translation=MTLLTNPIIISVVLMTLLCLFRFNVLLSLLISALVAG...
gene	6709	8010		/locus_tag=CJSA_0006
CDS	8144	12634		/codon_start=1 /gene=gltB /locus_tag=CJSA_0007 /product=glutamate synthase (NADPH) large subunit

Note how the reference sequence is annotated with **CDS** and **gene** features. This will be important later when we need to calculate the number of RNASeq reads that align to each feature.

Run a Bowtie Assembly

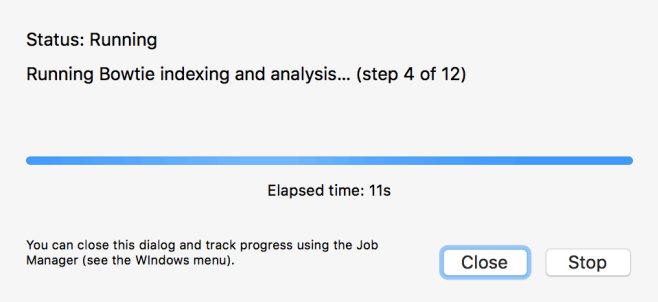
Close the sequence window and return to the Assembly Project. Select both items in the list (hold down the `<shift>` key to select the second item), then click the Bowtie toolbar button.



This is where you can change the parameters for the Bowtie analysis. Typically, the defaults settings work just fine. In this case, our reads are not paired-end, but if you do have paired-end reads, make sure the appropriate checkbox is selected.

Click on the OK button to start the Bowtie alignment.

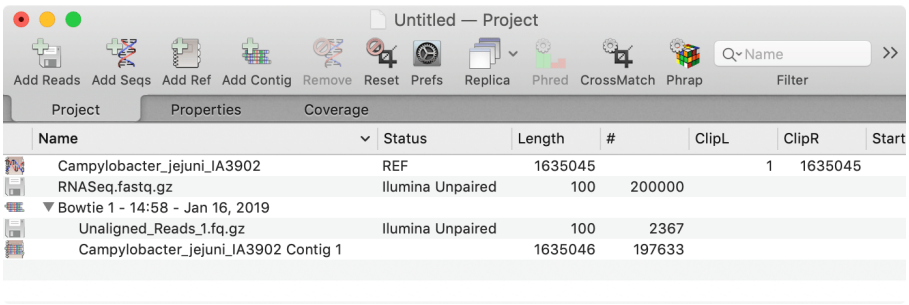
During the analysis, a job progress sheet will open;



The job should complete in less than a minute on a reasonably modern machine.

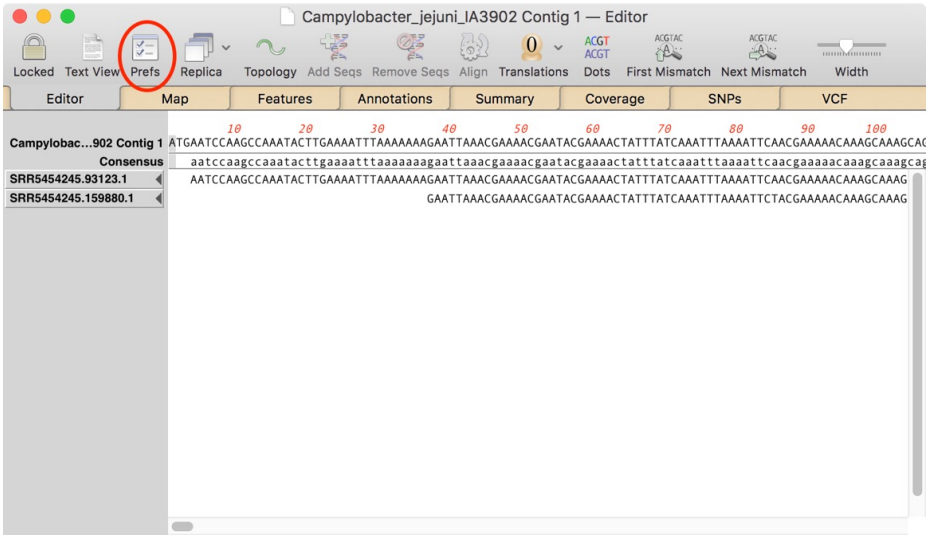
Click on the View button to close the dialog and show the results.

The Bowtie results are encapsulated in a named job item in the project. If you click on the disclosure item to the left of the job name you will see (a) a "Reference Contig" that you can click on to open up the alignment in the ***Contig Editor*** and (b) any reads that do not align are collected into an `Unaligned_Reads` file. While these are typically failed or contaminant reads, there may be times when these are the reads you want e.g. if your reference contained just rRNA and tRNA genes are you wanted to generate a fastq file enriched in mRNA.



Analyzing the Bowtie Results

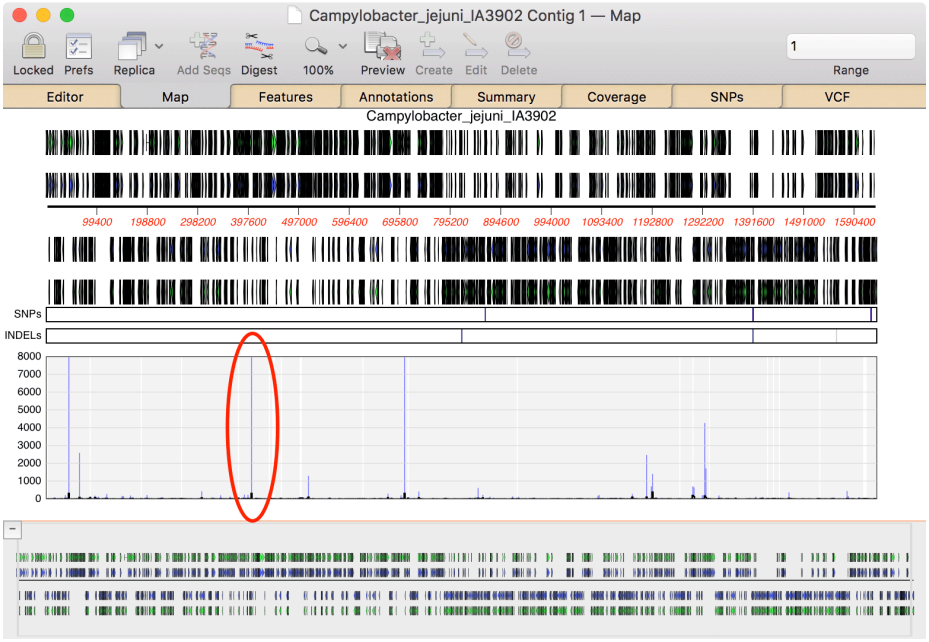
Double-click on the **Contig 1** item to open the ***Reference Contig*** document window. Switch to the Editor tab if necessary.



By default, the consensus is shown in the middle of the window. If you want to see the consensus at the top, immediately underneath the reference sequence (as shown above), click on the Prefs toolbar button.

The reads are shown aligned to the reference. You can scroll through the entire assembly (1.6 Mbp and ~198,000 reads) if you wish.

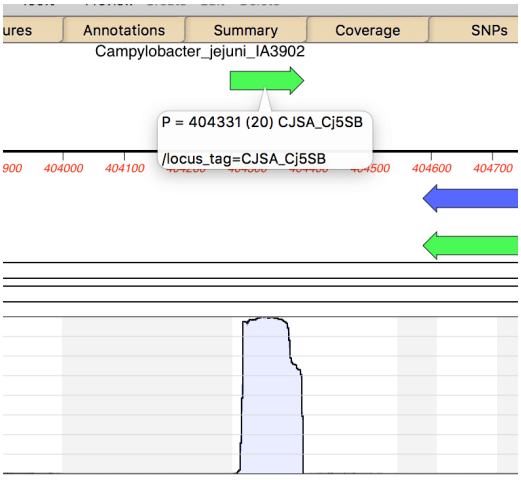
Select the Map tab.



This shows a map of the entire genome, along with coverage information for the reads. You can see there are several spikes of high coverage. Lets zoom into the circled area.

Carefully click to the left of the circled area, hold down the mouse button, drag to the right of the area and let go.

If you repeatedly drag, eventually you will see that there is a peak of coverage immediately under a single green **gene** feature;

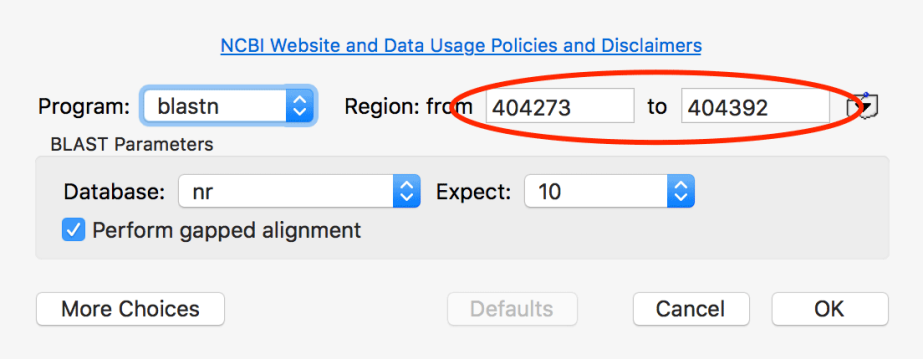


Let the mouse pointer hover over the green arrow representing the gene. The annotations for the **gene** are displayed in a tooltip.

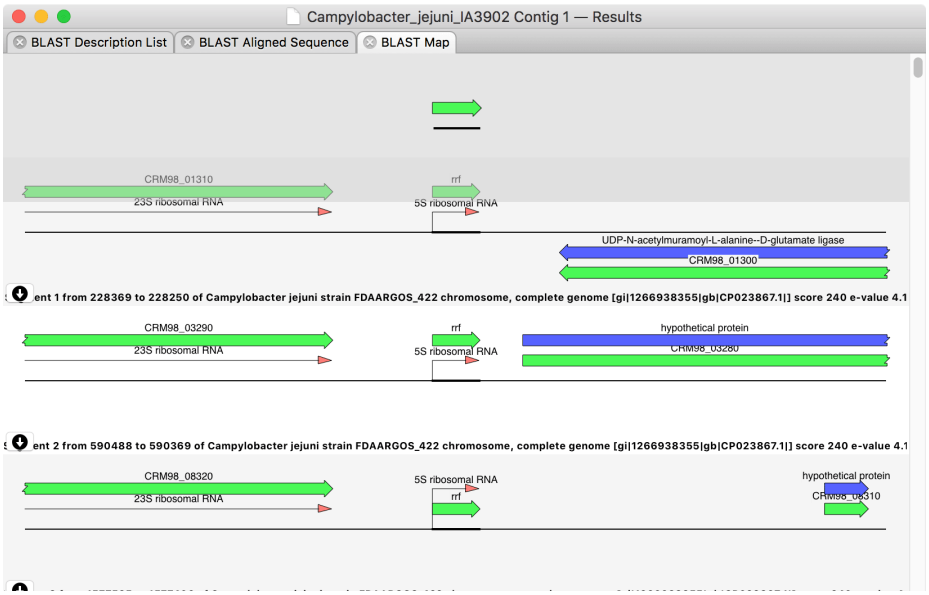
In this case, the annotation is not too informative, as there is simply a cryptic /locus_tag qualifier with the value “CJSA_Cj5SB”. However, it is simple to run an internet BLAST search to find out more information;

Click on the green Feature arrow so select it, then choose the **Database | Online Search for Similar Sequences (BLAST)** menu item. Set up the dialog as shown below.

Note that MacVector remembers the region you selected when you clicked on the green arrow,

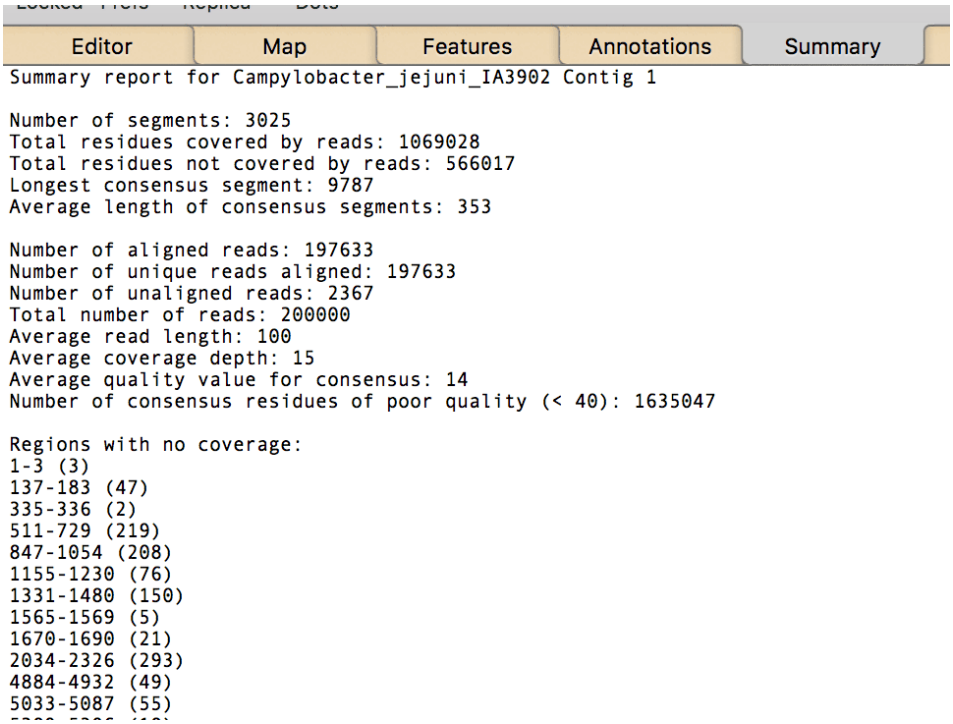


Click OK, wait for the job to complete, then click OK in the resulting sheet. Switch to the BLAST Map results tab.



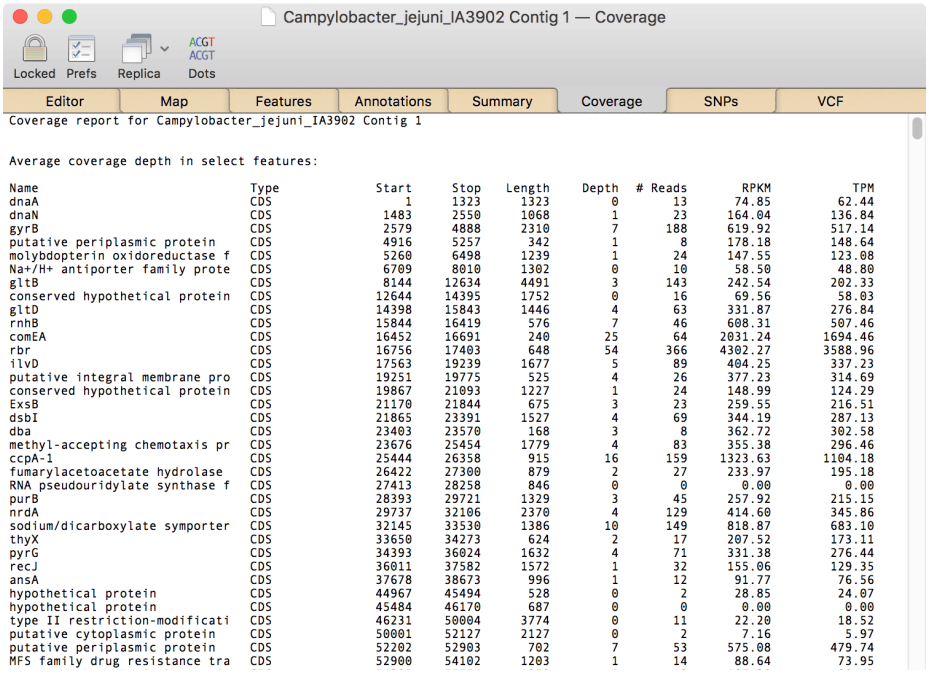
The BLAST Map displays a graphical representation of the region around the primary alignment. You can see the gene we selected at the top and the *High-scoring Segment Pairs* of the matching database sequences aligned underneath, complete with annotations on the matching region along with ~2kb on either side. It is immediately apparent that the gene aligns to 5s rRNA genes, exactly the sort of gene we would expect to express high levels of RNA in the experiment.

Close the BLAST results window and switch to the Summary tab of the reference contig window.



This tab summarizes the results of the Bowtie assembly. Because the data is RNASeq, the entire genome does not have coverage, as you would expect. In fact, there were 3,025 separate aligned segments, representing just over 1 million of the 1.6 million bases in the genome,

Switch to the Coverage tab.



This is by far the most useful tab for RNASeq expression analysis. There are a number of columns;

Name: this is the preferred name of the feature. For **CDS** features it is typically the contents of the `/gene` qualifier, but MacVector will use other qualifiers if `/gene` is not present.

Type: the type of feature. By default MacVector only displays **CDS** and **gene** features but other feature types can be requested.

Start: the start location of the feature.

Stop: the stop location of the feature.

Length: the length of the feature.

Depth: the average depth of coverage across the entire length of the feature (rounded down).

 # Reads: the total number of reads that aligned to the feature.

RPKM: Reads Per Kilobase of transcript per Million mapped reads. This is a common calculation used to normalize the data to facilitate comparison of expression levels between genes. It is calculated as follows;

- Count up the total reads in a sample and divide that number by 1,000,000 – this is our "per million" scaling factor.
- Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
- Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

TPM: Transcripts Per Kilobase Million. This is a variation on RPKM that is calculated slightly differently;

- Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
- Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
- Divide the RPK values by the "per million" scaling factor. This gives you TPM.

The advantage of using TPM is that this normalizes the data between different experiments so that you can directly compare the values for the same gene between different runs.

Exporting Data to Excel

The data in the Coverage tab is formatted to simplify exporting the columns into Microsoft Excel for further analysis. Specifically, the

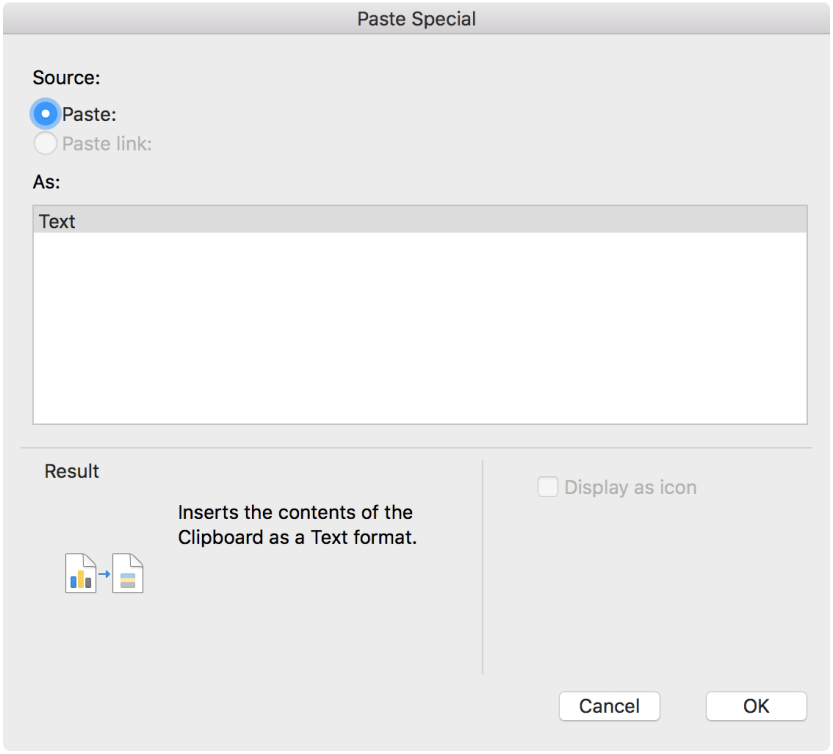
columns are tab-separated so that when you copy and paste into Excel, each value gets pasted into a separate cell.

Carefully select the text starting at the **Name** column header. Hold down the `<shift>` key and scroll to the bottom of the window. Still holding down `<shift>`, click just after the last character of the bottom line, so that the entire table gets selected. Choose **Edit | Copy**

This copies the entire text table to the clipboard. Now we can switch to Microsoft Excel and paste the data into a new workbook.

Open Microsoft Excel and (if necessary) create a new blank workbook. Select the top-left cell. Choose **Edit | Paste Special…**

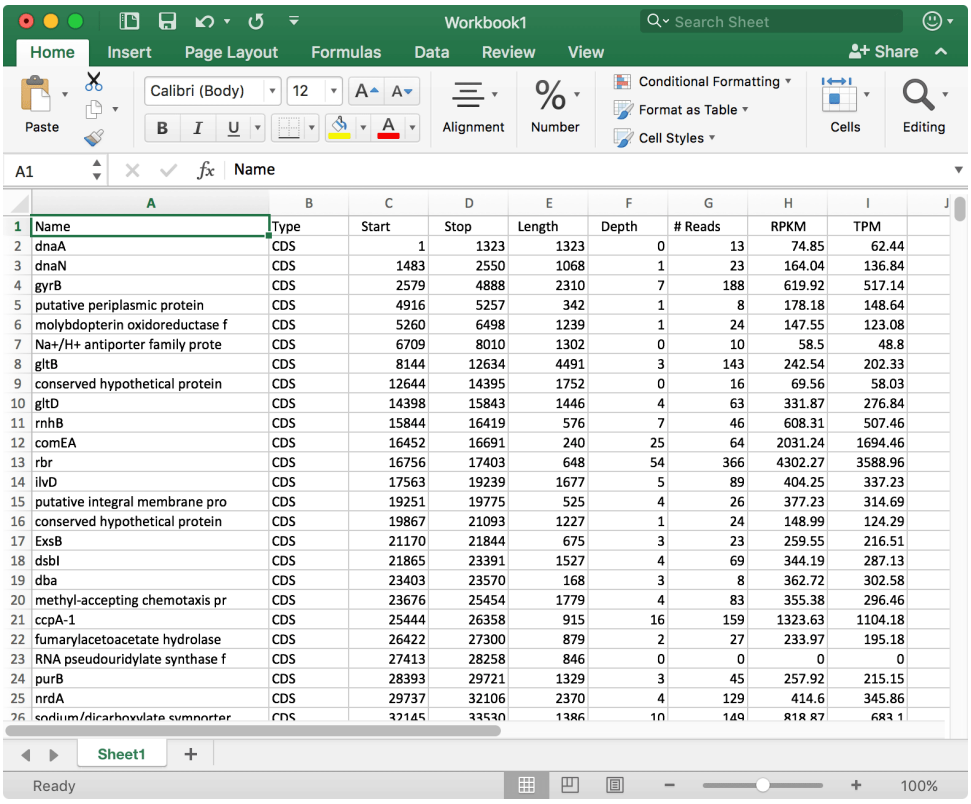
Older versions of Microsoft Excel (e.g. Office 2008) would correctly paste the tab-separated values directly, but Excel 2016 and later do not paste correctly, requiring you to go through this workaround;



Even though "Text" is the only possible option, you must go through this dialog to get the required behavior.

Click OK

The data gets pasted into the workbook, with each data point in a separate cell;



	A	B	C	D	E	F	G	H	I
1	Name	Type	Start	Stop	Length	Depth	# Reads	RPKM	TPM
2	dnaA	CDS	1	1323	1323	0	13	74.85	62.44
3	dnaN	CDS	1483	2550	1068	1	23	164.04	136.84
4	gyrB	CDS	2579	4888	2310	7	188	619.92	517.14
5	putative periplasmic protein	CDS	4916	5257	342	1	8	178.18	148.64
6	molybdopterin oxidoreductase f	CDS	5260	6498	1239	1	24	147.55	123.08
7	Na+/H+ antiporter family prote	CDS	6709	8010	1302	0	10	58.5	48.8
8	gltB	CDS	8144	12634	4491	3	143	242.54	202.33
9	conserved hypothetical protein	CDS	12644	14395	1752	0	16	69.56	58.03
10	gltD	CDS	14398	15843	1446	4	63	331.87	276.84
11	rnhB	CDS	15844	16419	576	7	46	608.31	507.46
12	comEA	CDS	16452	16691	240	25	64	2031.24	1694.46
13	rbr	CDS	16756	17403	648	54	366	4302.27	3588.96
14	ilvD	CDS	17563	19239	1677	5	89	404.25	337.23
15	putative integral membrane pro	CDS	19251	19775	525	4	26	377.23	314.69
16	conserved hypothetical protein	CDS	19867	21093	1227	1	24	148.99	124.29
17	ExsB	CDS	21170	21844	675	3	23	259.55	216.51
18	dsbI	CDS	21865	23391	1527	4	69	344.19	287.13
19	dba	CDS	23403	23570	168	3	8	362.72	302.58
20	methyl-accepting chemotaxis pr	CDS	23676	25454	1779	4	83	355.38	296.46
21	ccpA-1	CDS	25444	26358	915	16	159	1323.63	1104.18
22	fumarylacetoacetate hydrolase	CDS	26422	27300	879	2	27	233.97	195.18
23	RNA pseudouridylate synthase f	CDS	27413	28258	846	0	0	0	0
24	purB	CDS	28393	29721	1329	3	45	257.92	215.15
25	nrdA	CDS	29737	32106	2370	4	129	414.6	345.86
26	sodium/dicarboxylate symporter	CDS	32145	33530	1386	10	149	818.87	683.1

Now you can repeat this procedure with multiple datasets representing time-points, drug treatments, different growth conditions etc., and use the built-in functions of Excel for advanced analysis and comparison between runs.