

# MacVector 17

for Mac OS X

## **RNA-Seq Human Transcriptome Tutorial**

*MacVector, Inc.*  
Software for Scientists

## Copyright statement

Copyright **MacVector, Inc**, 2019. 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 the RNA-Seq Human Transcriptome Tutorial was published in August 2019.

# **Contents**

<b>CONTENTS</b>	<b>3</b>
<b>INTRODUCTION</b>	<b>4</b>
<b>SAMPLE FILES</b>	<b>4</b>
<b>STRATEGY</b>	<b>4</b>
<b>MACHINE REQUIREMENTS</b>	<b>5</b>
<b>TUTORIAL</b>	<b>5</b>
Download and Prepare Human Transcriptome	5
RNA-Seq Alignments using Bowtie	7
Exporting Data Into Microsoft Excel	11
Identifying Transcripts	14
Modifying Bowtie Parameters	17

## Introduction

MacVector with Assembler is capable of analyzing gene expression levels using the popular Next Generation Sequencing (NGS) “RNA-Seq” approach (also known as “whole transcriptome shotgun sequencing”). RNA from a source is isolated (potentially with enrichment for the type of RNA of interest), reverse-transcribed, then randomly sequenced using a high throughput NGS platform, such as Illumina HiSeq or MiSeq. The entire set of reads are then computationally aligned to an annotated reference genome and the relative abundance of transcripts from each transcribed gene determined by software.

There is an existing MacVector tutorial that uses a short bacterial reference genome and RNA-Seq data to illustrate the basic concept – see RNA-Seq Expression Analysis Tutorial.pdf.

This tutorial extends the concept to show how it is easily possible to analyze human RNA-Seq data using MacVector, even on a fairly modest laptop computer.

## Sample Files

The data used in this tutorial is not included in a standard MacVector installation because of the size of some of the data files. You can download the appropriate files using this link;

<https://macvector.net/humantranscriptomesampleddata.zip>

## Strategy

While it is (just) possible to align RNA-Seq data against the complete human genome with MacVector, that does require a fairly high-end machine with a LOT of RAM and even then, the analysis usually needs to be split into multiple tasks. It is far quicker, and requires less computational resources, to run the analysis against the human transcriptome i.e. just the known transcripts. There are several sources for this – our example will use the data collated by [GENCODE](#).

The steps we will use are;

- (a) Download the latest GENCODE human transcript data.
- (b) Concatenate the individual transcripts in that data to create a single reference sequence with the location of each transcript annotated appropriately.
- (c) Align a pair of RNA-Seq Illumina reads against the transcriptome reference using the popular *Bowtie* algorithm.
- (d) Ask MacVector to create a table listing how many reads aligned to each transcript along and calculate some basic statistical analysis.
- (e) Import the data into *Microsoft Excel* for further analysis.

## Machine Requirements

The total CPU time is noted for each major computational step during the tutorial. Initial timings were generated using a fairly high end (as of June 2019) 15” MacBook Pro with 32 GB RAM and a 6-core 2.9 GHz Intel Core i9 processor. While we recommend using machines with as much RAM as you can afford, as this is often the limiting factor, this entire workflow can be carried out on much more modest machines. 16 GB RAM is probably the practical lower end for human transcriptome analysis, but CPU speed is of less concern. Most of the analyses shown here used less than 6 MB RAM when running except where noted.

## Tutorial

### Download and Prepare Human Transcriptome

If you downloaded the combined zip file for this tutorial, then you have the MacVector file, all ready to go; `gencode.v26.pc_transcripts.fa.nucl`

This is how to generate an updated version of that file, or adapt this to your favorite transcriptome (GENCODE has a mouse version, for example, and other sites have versions for many other species).

In a browser, navigate to; <https://www.gencodegenes.org/human/> and scroll down the page to the **Fasta files** section and click on the Fasta download link for “Transcript Sequences”. This contains all known human RNA transcripts . Currently, this points to [ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_30/gencode.v30.transcripts.fa.gz](ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_30/gencode.v30.transcripts.fa.gz). This usually gets uncompressed during download.

The file is currently (June 2019) 354 MB.

Create a new folder somewhere on your hard drive where you can store data files and run the analysis. E.g. create one in your home folder called `HumanTranscriptome`. Move the downloaded `Xxxtranscripts.fa` file to this folder.

Now we are ready to convert this fasta file into a GenBank formatted file that we can import into MacVector.

Copy the `FASTAtoAnnotatedGB.pl` script file into the folder you created.

You most likely downloaded this file from the MacVector website, along with the data for this tutorial. Alternatively, for MacVector 17.1 and later, it is located in the `/Applications/MacVector/Applescripts/` folder.

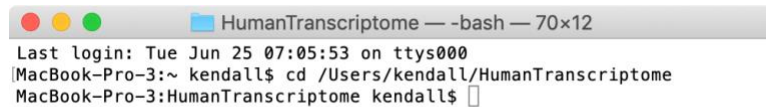
It is not strictly speaking necessary to copy the script to this location, but it simplifies the command line arguments below if you are not familiar with the use of `Terminal.app`.

Open `Terminal.app` (you can find this in `/Applications/Utilities/`).

We now want to change the current directory to the directory you created above. If you are familiar with command lines on the Mac, simply “cd” to that directory, otherwise, follow these instructions;

In the `Terminal.app` window, type “cd” then press the `<space>` bar. Switch to the Apple *Finder* application, navigate to the parent of the folder you created, select the new folder and drag it onto the Terminal window.

You should see something like this;



```
HumanTranscriptome — -bash — 70x12
Last login: Tue Jun 25 07:05:53 on ttys000
MacBook-Pro-3:~ kendall$ cd /Users/kendall/HumanTranscriptome
MacBook-Pro-3:HumanTranscriptome kendall$
```

Now you are ready to run the conversion script to create a new GenBank file containing all of the transcripts concatenated into a single large annotated sequence.

Type (or copy/paste) the following command (substitute your specific filename as appropriate) and press `<return>`;

```
./FASTAtoAnnotatedGB.pl gencode.v30.transcripts.fa -
sort=forward
```

A single prompt should appear describing what will happen – accept appropriately, unless something appears way off. During processing, you may get warnings about missing *Description* values. You can ignore these. On a MacBook Pro, processing takes about two minutes.

Open MacVector. Choose **File | Open** and navigate to your transcriptome folder and select the new `...transcripts.fa.gb` file that should be in there and click **OK**.

It will take some time to open this file. Initially it will appear as if you had not clicked the **OK** button as MacVector tries to parse the contents of the file. However, you will be prompted to confirm that you want to continue importing the sequence because the sequence is so large with very many features. On a MacBook Pro, the entire import took about one minute.

You now have a file with a ~323 Mbp concatenated sequence containing ~210,000 “gene” features, each of which is annotated with a `/dbxref` qualifier that references the identifiers of the original transcript.

Choose **File | Save As...** and save the file with a suitable filename. For the data used in this tutorial, that is `gencode.v30.transcripts.fa.nucl`

The final file takes about 45 seconds to save on the MacBook Pro and is about 395 GB on disk.

That's it! You now have a reference sequence that contains every known human transcript that you can use for RNA-seq experiments.

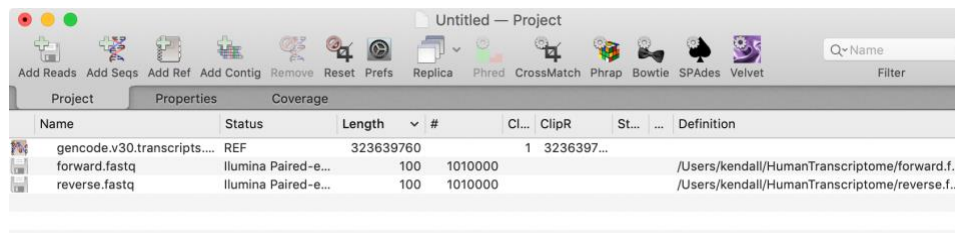
## RNA-Seq Alignments using Bowtie

The next step is to align the sample RNA-Seq reads to our new human transcriptome so that we can evaluate expression levels.

For this we are going to use a very reduced sample set to reduce processing time. You should have downloaded this sample set with the files required to run this tutorial.

First choose **File | New | Assembly Project** to create a new project. Then click on the **Add Ref** button and select your new `gencode.v30.transcripts.fa.nucl` reference sequence to import it. Finally, click on the **Add Reads** toolbar button, navigate to the location of the sample files, select `forward.fastq` and `reverse.fastq` (hold down the `<option>` key for multiple selections) and click **OK** to add them to the project.

You should have a project looking something like this;



Select all of the files and click on the **Bowtie** toolbar item.

The defaults will work fine for this example. Note in particular that *Hit Reporting* is set to **Report Best Hit only**;

**Bowtie Options**

Preset: Sensitive

Type of Alignment: Local

No gaps within first bases of read:

Number of Threads:

**Read pre-processing**

Discard reads less than  nt

Trim ends with quality less than

Trim N's from ends

Discard short reads that contain any N's

Use paired-end alignments

Minimum insert size:

Maximum insert size:

Orientation: Forward - Reverse

Generate child contigs

Check this box if you are using the Reference as a scaffold to assemble related reads, or if you want to 'drill down' into individual alignments. For other tasks (e.g. SNP analysis or RNA-Seq expression analysis) leave this unchecked.

**Hit Reporting**

Report Best Hit only

Number of hits to report:

Report all alignments

?
Defaults
Cancel
OK

This setting means that each read will only align at a single location on the reference sequence. However, because of splice-site variations, pseudogenes and possible duplicated entries in the reference dataset, this may mean that some valid alignments will be missed. The implications will be discussed later with suggestions for alternative settings.

Click **OK** and wait.....

The sample files have about 1 million x100nt reads each. On the MacBook Pro, this takes about 20 minutes to align. Once complete, a new job object appears in the project window.

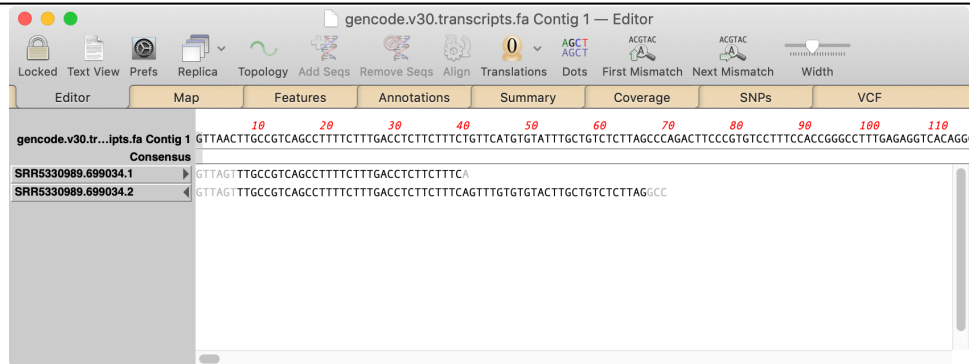
Untitled — Project									
Project		Properties		Coverage					
Name	Status	Length	#	ClipL	ClipR	Start	Stop	Definition	
gencode.v30.transcripts.fa	REF	323639760			1	3236397...			
forward.fastq	Illumina Paired-e...	100	1010000						/Volumes/D...
reverse.fastq	Illumina Paired-e...	100	1010000						/Volumes/D...
▼ Bowtie 1 - 12:29 - Jun 27, 2...									
Unaligned_Reads_1_1.fq.gz	Illumina Paired-e...	100	437743						/var/folders...
Unaligned_Reads_1_2.fq.gz	Illumina Paired-e...	100	437743						/var/folders...
gencode.v30.transcripts...		323639958	11567028						

Note that the reads that did not align are shown as a pair of `Unaligned_Reads` files. There are times when these might be exactly the files you want e.g. if you wanted to filter out all human RNA sequences in order to enrich for bacterial or virus sequences in blood samples.

The aligned reads are included in the contig object called `gencode.v30.transcripts.fa Contig 1`.



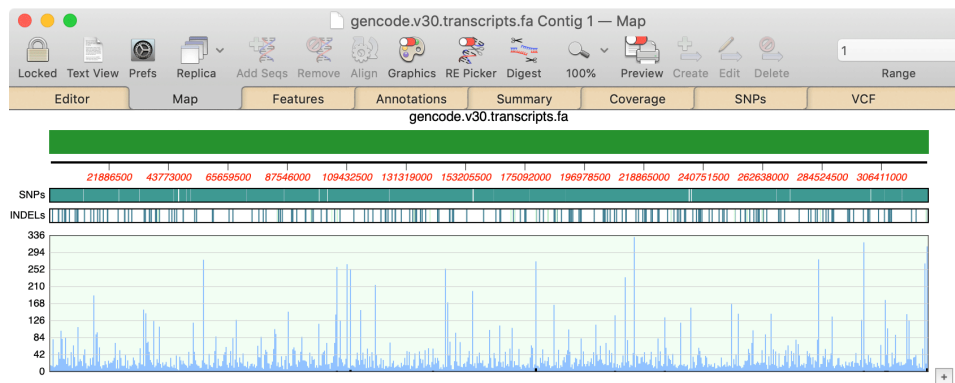
Double-click on the **Contig 1** object to open the *Contig Editor*



If you like, you can scroll through the entire assembly, viewing the actual aligned sequences, but there is typically little need to do this for these types of experiments.

Click on the **Map** tab.

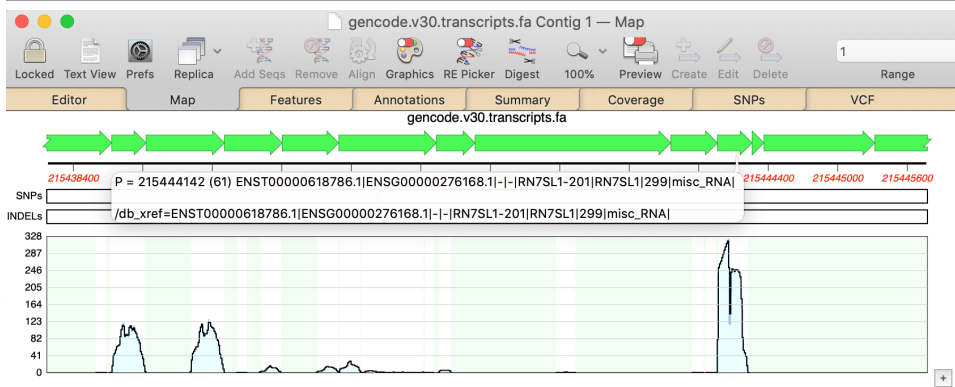
The **Map** tab can take a few seconds to calculate and draw due to the size of the sequence. It typically completes in less than 30 seconds;



The green bar is actually a graphic showing all 210,000 transcripts superimposed on top of each other. The lower graph is the distribution of aligned reads across the reference.

You can “zoom in” to view the coverage in more detail.

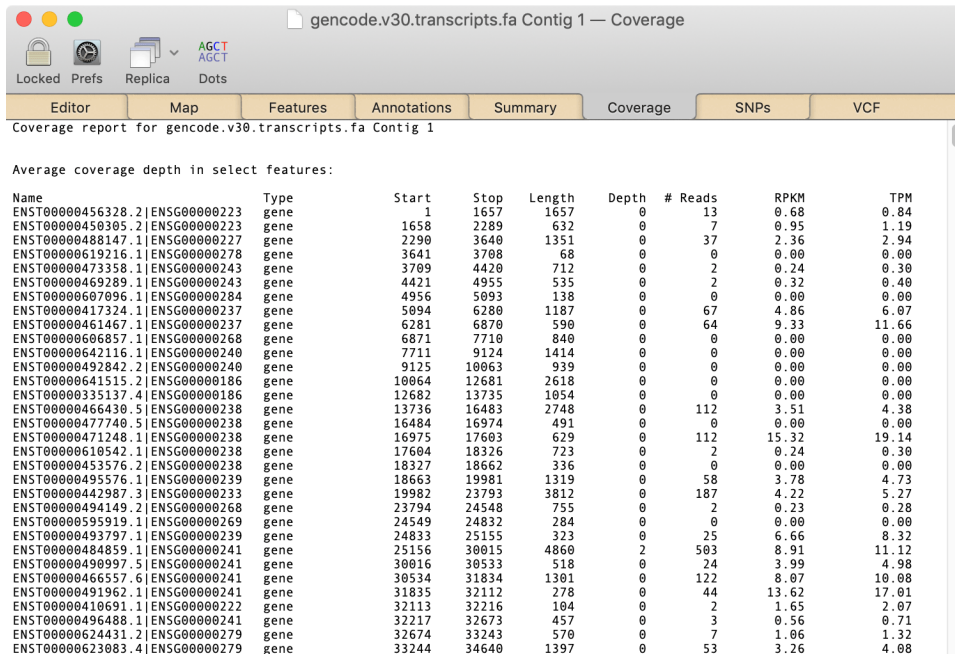
Pick one of the taller peaks and carefully select a short section on either side of it by clicking and dragging with the mouse or trackpad. The response may be a little sluggish at first due to the size of the reference. Repeat the drag until you can clearly see the individual reads.



If you hover the pointer over one of the green arrows, a tooltip appears with the annotation for that transcript.

Click on the **Coverage** tab.

This is the tab that lists the coverage for each transcript. There is a lot of data to process, so the content of the tab can take some time to appear (it is only generated when you click on the tab to save processing time for those cases where this information is of no interest). On the MacBook Pro, the tab took 90 seconds to calculate and display. But, once calculated, you can switch back and forth between tabs and it does not need to be calculated again while the Contig 1 window remains open.



The columns are;

- **Name** – the first 30 characters of the transcript name.
- **Type** – the type of feature. The FastatoAnnotatedGB.pl script assigns gene features to each transcript, but for annotated genomes these may be other types.

- **Start** – the start location of the transcript in the concatenated reference sequence.
- **Stop** – the stop position
- **Length** – the length of the reference transcript
- **Depth** – the average depth of coverage, rounded down. These reads are 100 nt in length so e.g. 13 reads across a 1,657 nt transcript still comes out to a coverage of <1x.
- **# Reads** – the number of reads that aligned to the transcript.
- **RPKM** - Reads Per Kilobase of transcript, per Million mapped reads. This is a normalized unit of transcript expression that scales by transcript length to compensate for the fact that most RNA-Seq protocols will generate more sequencing reads from longer RNA molecules.
- **TPM** – Transcripts Per Kilobase Million. When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a transcript across different samples.

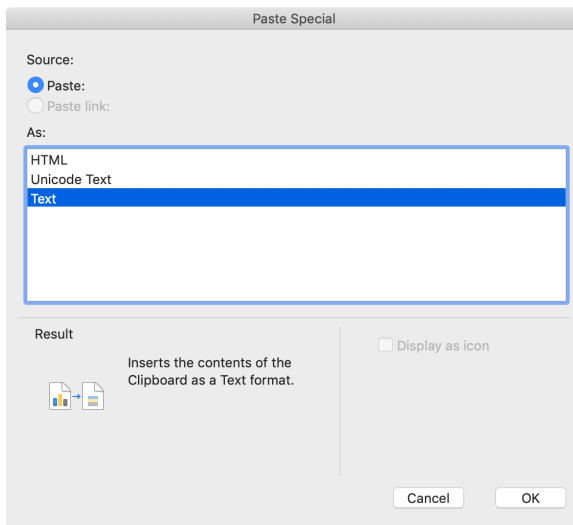
## Exporting Data Into Microsoft Excel

While MacVector does not currently have an interface to compare and analyze RNA-Seq samples from multiple experiments, it's easy to export the data from the coverage tab as it is displayed in tab separated format.

Carefully select the first few lines of the **Coverage** tab data, starting at the *Name* column header. Now scroll to the bottom of the data, hold down the <shift> key and click just after the last TPM data item and the entire text table should select. Choose the **Edit | Copy** menu item.

This will copy the text to the clipboard, so we can now switch to *Microsoft Excel* to paste the data. This tutorial used *Excel* version 16.26 as distributed with *Office 365*. This approach should also work with other spreadsheet-like applications.

Switch to *Microsoft Excel*. Create a new blank workbook and click in cell A1. Choose **Edit | Paste Special...** then select the *Text* option in the resulting dialog and press **OK**.



The data should get pasted into *Excel* with each data item in its own separate cell and column headers just as they appeared in MacVector.

	A	B	C	D	E	F	G	H	I
1	Name	Type	Start	Stop	Length	Depth	# Reads	RPKM	TPM
2	ENST000004	gene	1	1657	1657	0	8	0.96	1.26
3	ENST000004	gene	1658	2289	632	0	2	0.63	0.83
4	ENST000004	gene	2290	3640	1351	0	24	3.54	4.65
5	ENST000006	gene	3641	3708	68	0	0	0	0
6	ENST000004	gene	3709	4420	712	0	0	0	0
7	ENST000004	gene	4421	4955	535	0	0	0	0
8	ENST000006	gene	4956	5093	138	0	0	0	0
9	ENST000004	gene	5094	6280	1187	0	3	0.5	0.66
10	ENST000004	gene	6281	6870	590	0	10	3.37	4.44
11	ENST000006	gene	6871	7710	840	0	0	0	0
12	ENST000006	gene	7711	9124	1414	0	0	0	0
13	ENST000004	gene	9125	10063	939	0	0	0	0
14	ENST000006	gene	10064	12681	2618	0	0	0	0
15	ENST000003	gene	12682	13735	1054	0	0	0	0
16	ENST000004	gene	13736	16483	2748	0	13	0.94	1.24
17	ENST000004	gene	16484	16974	491	0	0	0	0
18	ENST000004	gene	16975	17603	629	0	17	5.38	7.08
19	ENST000006	gene	17604	18326	723	0	0	0	0
20	ENST000004	gene	18327	18662	336	0	0	0	0

It is trivial to sort the data in any column;

Click on the *TPM* header cell then select the **Data | Auto Filter** menu item

The headers should each now have a button at the right side;

	A	B	C	D	E	F	G	H	I
1	Name	Type	Start	Stop	Length	Depth	# Reads	RPKM	TPM
2	ENST000004	gene	1	1657	1657	0	8	0.96	1.26
3	ENST000004	gene	1658	2289	632	0	2	0.63	0.83
4	ENST000004	gene	2290	3640	1351	0	24	3.54	4.65
5	ENST000006	gene	3641	3708	68	0	0	0	0

Click on the button next to the *TPM* header and select **Descending** in the resulting dialog.

The screenshot shows the Microsoft Excel interface. The ribbon includes Home, Insert, Draw, Page Layout, Formulas, Data, Review, and View. The spreadsheet data is as follows:

	A	B	C	D	E	F	G	H	I	
1	Name	Type	Start	Stop	Length	Depth	# Reads	RPKM	TPM	
2	ENST00000631211.1	ENSG00000280	gene	300054534	300055456	923	209	59904	12914.81	16997.79
3	ENST00000625598.1	ENSG00000280	gene	300065104	300066026	923	212	59904	12914.81	16997.79
4	ENST00000627981.1	ENSG00000281	gene	300067716	300068638	923	209	59904	12914.81	16997.79
5	ENST00000629969.1	ENSG00000281	gene	300057148	300058070	923	172	59858	12904.89	16984.74
6	ENST00000490232.3	ENSG00000274	gene	215525009	215525308	300	188	6435	4268.36	5617.79
7	ENST00000618786.1	ENSG00000276	gene	215443962	215444260	299	184	6404	4262.01	5609.43
8	ENST00000620465.4	ENSG00000251	gene	179214480	179214812	333	18	7097	4240.96	5581.73
9	ENST00000612781.1	ENSG00000251	gene	179214813	179215046	234	0	4699	3995.98	5259.3
10	ENST00000610674.1	ENSG00000278	gene	215522697	215522995	299	5	5979	3979.16	5237.16
11	ENST00000600213.3	ENSG00000269	gene	64655675	64656723	1049	3	20235	3838.5	5052.03
12	ENST00000581458.2	ENSG00000265	gene	146149789	146150109	321	85	6178	3829.81	5040.59
13	ENST00000584058.2	ENSG00000263	gene	56969414	56969708	295	97	5655	3814.56	5020.52
14	ENST00000613376.1	ENSG00000251	gene	179215772	179215903	132	24	2294	3458.23	4551.53
15	ENST00000445125.2	ENSG00000225	gene	322801815	322803662	1848	86	30243	3256.54	4286.09
16	ENST00000389680.2	ENSG00000211	gene	323624440	323625393	954	222	15513	3235.8	4258.79
17	ENST00000618132.1	ENSG00000251	gene	179215047	179215771	725	142	11757	3226.95	4247.14
18	ENST00000387347.2	ENSG00000210	gene	323625463	323627021	1559	206	23819	3040.26	4001.43
19	ENST00000536684.2	ENSG00000255	gene	172331153	172332455	1303	2	19648	3000.6	3949.23
20	ENST00000361851.1	ENSG00000228	gene	323632079	323632285	207	248	3004	2887.78	3800.74

The list is sorted, and the most highly expressed genes are displayed at the top.

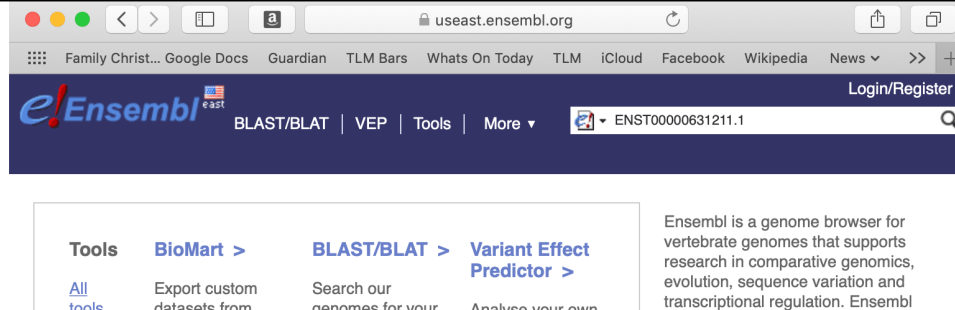
You can always get back to the original order by filtering **Ascending** on the *Start* header.

You can use this approach to compare results between different data sets. For example, you could paste results from a second data set into a second sheet, then copy the *TPM* column from that and paste next to the *TPM* column from the first sample. Then you could create a simple “*Delta*” column with the starting formula of (e.g.) “=I2–J2”, copied to all cells in the column. That would display the differences between the two sets of data. Then you could sort that column by **Descending** to identify those transcripts that had the highest drop off in expression from sample “I” to sample “J” and by **Ascending** to identify those that had the greatest increase in expression in sample “J”. In reality, you would probably want to use more sophisticated formulas to use a ratio of expression levels in the *Delta* column, but this gives a general concept of how to proceed.

## Identifying Transcripts

Once you have identified transcripts of interest, it would be nice to find out what they actually encode. By far the easiest way is to use a web browser and use the *Ensembl* genome browser.

First, carefully copy the text up to the first “|” in the top hit. In the case of this data it is accession number ENST00000631211.1. Open a browser and go to <https://ensembl.org>. In the top right corner is a *Search all species* edit box.



Paste the accession number into the search box as shown above and press <return>

**ENST00000631211.1**

2 results match ENST00000631211.1

[FP671120.4-201 \(Human Transcript\)](#)  
**ENST00000631211** 21:8210384-8211306:-1  
 Novel transcript, similar to YY1 associated myogenesis RNA 1 YAM1  
**ENST00000631211.1** (UCSC Stable ID) is an external reference matched to Transcript ENST00000631211  
[Location](#) • [External Refs.](#) • [cDNA seq.](#) • [Exons](#) • [Variant table](#) • [Population](#)

[FP671120.4 \(Human Gene\)](#)  
**ENSG00000280800** 21:8210384-8211306:-1  
 Novel transcript, similar to YY1 associated myogenesis RNA 1 YAM1  
**ENST00000631211.1** (UCSC Stable ID) is an external reference matched to Transcript ENST00000631211  
[Variant table](#) • [Phenotypes](#) • [Location](#) • [External Refs.](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

<< < 1 > >>

This immediately finds the appropriate references to the transcript. You can then click on the links to explore the transcript in more detail. If you'd like to download the region around the transcript location for more analysis in MacVector, here's how to do it;

Click on the top link that indicates “(Human Transcript)”

**Human (GRCh38.p12)**

Location: 21:8,210,384-8,211,306 Gene: FP671120.4 Transcript: FP671120.4-201

**Transcript-based displays**

- Summary
- Sequence
  - Exons
  - cDNA
  - Protein
- Protein Information
  - Protein summary
  - Domains & features
  - Variants
  - 3D Protein model
- Genetic Variation
  - Variant table
  - Variant image
  - Haplotypes
  - Population comparison
  - Comparison image
- External References
  - General identifiers
  - Oligo probes
- Supporting evidence
- ID History
  - Transcript history
  - Protein history

**Transcript: FP671120.4-201** ENST00000631211.1

**Description**  
novel transcript, similar to YY1 associated myogenesis RNA 1 YAM1

**Location**  
[Chromosome 21: 8,210,384-8,211,306](#) reverse strand.

**About this transcript**  
This transcript has [1 exon](#), is associated with [1 variant allele](#) and maps to [112 oligo probes](#).

**Gene**  
This transcript is a product of gene [ENSG00000280800](#) [Hide transcript table](#)

Show/hide columns (1 hidden)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq
FP671120.4-201	<a href="#">ENST00000631211.1</a>	923	No protein	lincRNA	-	

**Summary**

You will end up on a page with tabs for *Location*, *Gene* and *Transcript*.

Click on the *Location* tab

Location: 21:8,210,384-8,211,306 Gene: FP671120.4 Transcript: FP671120.4-201

**Location-based displays**

- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
  - Synten
  - Alignments (image)
  - Alignments (text)
  - Region Comparison
- Genetic Variation
  - Variant table
  - Resequencing
  - Linkage Data
- Markers
- Other genome browsers
  - UCSC
  - NCBI
  - Ensembl GRCh37

Configure this page

Custom tracks

**Export data**

Share this page

Bookmark this page

**Chromosome 21: 8,210,384-8,211,306**

Assembly exceptions  
Chr. 21

Region in detail

Chromosome bands  
Contigs  
Genes  
(Comprehensive set from GENCODE 27)

8.00 Mb 1.00 Mb 8.50 Mb Forward strand

6241.2 < KCNE1B  
F00019 < RF00614  
< SMIM34B  
SMIM11B >  
P236241.1  
< FAM243B

RF01518 > FP236383.9 >  
RF00002 > FP236383.5 >  
FP671120.8 > FP236383.8 >  
< FP671120.4 RF01518 > RPSAP6  
MIR3648-1 > < FP236383.3  
MIR6724-2 > RF01518 >  
MIR6724-1 > FP236383.6 >  
FP671120.1 > < FP236383.2  
< FP671120.5 RNAB-BSN1 >  
FP671120.9 > FP236383.4 >  
< FP671 20.3 MIR6724-3 >  
ENSG00000280800

This shows the genes and annotations in the region around the transcript. There are a lot of customization options in the browser which you can explore.

To export the data in a format MacVector can use, click on the **Export Data** button.

Export data

### Export Configuration - Feature List

Location to export: chromosome:GRCh38:21:8210384:8211306:1

Output: GenBank

Select location: 21 \* 8210384 \* 8211306 \* 1

5' Flanking sequence (upstream): 2000 \* (Maximum of 1000000)

3' Flanking sequence (downstream): 2000 \* (Maximum of 1000000)

Next >

Fields marked \* are required

### Options for GenBank

Select/deselect all:

Similarity features:

Repeat features:

Prediction features (genscan):

Contig Information:

Variant features:

Make sure you select **GenBank** as the output. Optionally add additional residues on each side of the location for context. Above we asked for an extra 2kb on each side. Click on the **Next** button.

A configuration window appears;

### Export Configuration - Output Format

Please choose the output format for your export

- [HTML](#)
- [Text](#)
- [Compressed text \(.gz\)](#)

Click on the **Text** link.

This displays the sequence in GenBank text format;



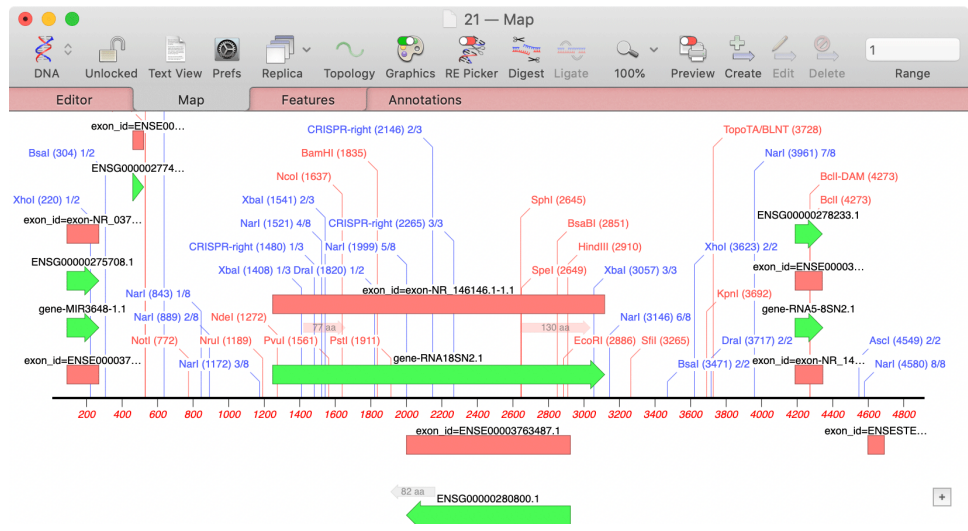
```

... Family:Chr1... Google Docs...
Chromosome 21: 8,210,384-8,211,306 - Region in detail - Homo... https://useast.ensembl.org/Homo_sapiens/Export/Output/L... +
LOCUS      21 4923 bp DNA HTG 27-JUN-2019
DEFINITION Homo sapiens chromosome 21 GRCh38 partial sequence 8208384..8213306 reannotated
            via Ensembl
ACCESSION  chromosome:GRCh38:21:8208384:8213306:1
VERSION    chromosome:GRCh38:21:8208384:8213306:1
KEYWORDS   .
SOURCE     human
ORGANISM   Homo sapiens
            Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria;
            Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata;
            Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha;
            Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria;
            Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini;
            Hominoidea; Hominidae.
COMMENT    This sequence was annotated by Ensembl(www.ensembl.org). Please visit the Ensembl
            or EnsemblGenomes web site, http://www.ensembl.org/ or
            http://www.ensemblgenomes.org/ for more information.
COMMENT    All feature locations are relative to the first (5') base of the sequence in this
            file. The sequence presented is always the forward strand of the assembly.
    
```

Its easy to get this into MacVector;

Either (a) carefully select from just before the LOCUS text at the top to just after the trailing // characters at the bottom of the window or (b) choose **Edit | Select All**. Choose **Edit | Copy**. Switch to MacVector. Select **File | New From Clipboard**.

You may get a warning about invalid features in the data – Ensembl does not always adhere particularly closely to the GenBank standard. Any features that cannot be parsed will be saved in the *COMMENT* section of the **Annotations** tab.



## Modifying Bowtie Parameters

While the above example tutorial does use a full-length Human Transcriptome sequence, it only uses a total of 2 million reads to speed things up for tutorial purposes. In addition, we use a *Bowtie* setting of **Report Best Hit only**. That means that if more than one transcript matches a particular read, only one will be reported as a match. In general, it is considered better to allow *Bowtie* to match multiple transcripts, though there are many scenarios where the single hit might be preferable. Let's look at the *Bowtie* dialog again;

The screenshot shows the Bowtie alignment software interface with the following settings:

- Bowtie Options:**
  - Preset: **Very Sensitive** (dropdown)
  - Type of Alignment: **Local** (dropdown)
  - No gaps within first bases of read: **4** (input)
  - Number of Threads: **8** (input)
- Read pre-processing:**
  - Discard reads less than **33** nt (input)
  - Trim ends with quality less than **20** (input)
  - Trim N's from ends
  - Discard short reads that contain any N's
- Use paired-end alignments**
  - Minimum insert size: **0** (input)
  - Maximum insert size: **500** (input)
  - Orientation: **Forward - Reverse** (dropdown)
- Generate child contigs**
- Hit Reporting:**
  - Report Best Hit only
  - Number of hits to report:** **8** (input)
  - Report all alignments

At the bottom, there is a help icon (?), a "Defaults" button, a "Cancel" button, and an "OK" button. A note below the "Generate child contigs" option reads: "Check this box if you are using the Reference as a scaffold to assemble related reads, or if you want to 'drill down' into individual alignments. For other tasks (e.g. SNP analysis or RNA-Seq expression analysis) leave this unchecked."

The **Preset** settings look as if they would make a significant difference to the speed of the alignment, but, in our hands, the differences are minimal. In general, there is less than a 10% difference in computation speed between *Very Fast* and *Very Sensitive*. However, *Very Sensitive* does tend to align more reads.

If you are using paired-end alignments and your insert size is significantly different than the defaults, you may want to change those settings.

The **Read pre-processing** section can generally be left unchecked. If you think you have a lot of failed reads, this might help clean up the data, but in general *Bowtie* will simply ignore bad reads.

The **Hit Reporting** section is the most critical for these types of alignments. Our example used **Report Best Hit only** and this does generate valid results. However, if you want reads to map to multiple transcripts, you need to choose one of the other options. The most obvious solution is to check **Report all alignments**. However, in our experience, with the Human Transcriptome, this has a huge effect on alignment time. With the sample data set, **Report Best Hit only** takes from 20 to 30 minutes to complete, no matter which **Preset** is used. But **Report all alignments** takes 6 to 7 hours and uses a maximum of ~14 MB RAM rather than the 4-6 MB RAM used with the other analyses. The upside of this is that many transcripts get many more reads aligned to them.

One alternative approach is to use the **Number of hits to report** option. This option limits the *Bowtie* search to give up after X number of hits are found. It turns out you can set this to a fairly high number and it still completes much faster than the **Report All Alignments** option. Let's look at a table with some timings;

		Time	Aligned Reads	Unaligned Reads
Best Hit Only	Very Sensitive	20:13	1199781	878786
Best Hit Only	Very Fast	21:53	1129472	928548
Number of hits = 4	Very Fast	28:22	3445495	923040
Number of hits = 8	Very Fast	22:47	5012563	922056
Number of hits = 12	Very Fast	21:32	5965105	921692
Number of Hits = 12	Very Sensitive	29:44	6784849	872834
Number of hits = 50	Very Sensitive	44:32	12992388	871288
Number of hits = 50	Very Fast	44:15	9961427	921034
All Alignments	Very Fast	6:26:36	79237491	920712

Here we can see that using the **All Alignments** option is around 20x slower than most of the other options. However, it does generate many more alignments (more than 8x as many as the equivalent 50 hits option), though it may be that many of these are spurious alignments rather than close-to-perfect matches. Similarly, using the **Very Sensitive** option leaves fewer unaligned reads, but, again, these may be imperfect matches.

Overall, the optimum parameters may depend on the actual questions you are asking of the data. As a trade-off between computational time, noise from spurious alignments and sensitivity, a good place to start would be **Number of hits = 12** with the **Very Sensitive** option.