MacVector 13

for Mac OS X

Tutorial: Next Generation Reference Alignments Using Bowtie

MacVector, Inc.

Software for Scientists

Copyright statement

This version of the NGS Reference Alignment tutorial was published in March 2014.

Contents

Introduction

Generating sequencing data is cheaper than it has ever been. However, with this increase in data has come a problem with analyzing this data using a desktop computer.

To help make this achievable, MacVector with Assembler can create reference assemblies from next generation sequencing data with just a few mouse clicks. Instead of sending millions of reads away to be assembled or delving into complicated software tools, you'll be able to align millions of NGS reads to multi megabase reference sequences in just a few minutes.

Assembler uses the popular Bowtie algorithm to create reference assemblies. Bowtie is a sequence aligner capable of extremely fast alignments of short sequences against a much larger reference sequence. Although it currently uses an ungapped alignment algorithm, what it loses in accuracy it makes up for in speed. You can assemble your data even on a low powered laptop computer with only two or three GB of RAM.

Assembler allows easy point and click assembly of reads against a reference using Bowtie. It will generate reports of SNPs and other variants. It supports the Variant Calling Format (VCF) and BAM file formats. Reads can be assembled against multiple references in a single analysis. Consensus and contig sequences can be exported in Fasta and Fastq formats for further analysis.

This tutorial will show how you can align a set of reads against a single reference sequence and how to analyze the results.

Sample Files

MacVector comes with a small set of tutorial sequences and reads. It's a "contrived" sample designed to assemble very quickly and small enough to download as part of the installer. The reference sequence is a single contig that is part of the *L. paracasei* genome and the reads are a small subset of the sequencing project that was used to sequence that genome using a 454 sequencer (from a single *de novo* assembled contig). They come from the SRR015575 set (the full set contains 6 such files). Note that although the reference sequence is the full contig sequence the reads will align only against the region from 1 to 427,800. Also be aware that the reads have been manipulated with two mutations.

The original reference sequence is available from the NCBI:

http://www.ncbi.nlm.nih.gov/nuccore/DS990486

You can also download the original full set of reads direct from the NCBI's Short Read Archive (SRA):

http://www.ncbi.nlm.nih.gov/sra/?term=SRR015575

These sample files we use for this tutorial are kept in the following folder after MacVector has been installed;

```
/Applications/MacVector 12.6/Tutorial Files/Contig Assembly/NextGen files/
```

Overview/Quickstart

To create a reference assembly, you need to have a reference sequence in any common format (MacVector, GenBank, EMBL, FastA etc) and Read files in either FastA or (preferably) FastQ format.

To quickly map reads against a reference, follow these steps;

1. Start MacVector with Assembler and choose **File | New | Assembly Project**
2. Click on the **Add Ref** button – in the resulting dialog, locate your reference sequence and click OK
3. Click on the **Add Seqs** button – in the resulting dialog, find the FastA or FastQ file(s) containing your Reads and click OK.
4. Select the reference and the read file in the Project window (hold down <shift> or use <command>-click to toggle selections). Click on the **Bowtie** icon on the toolbar.

Note: that if no sequences are selected, Bowtie will be run on ALL of the files in the project. However, if any sequences are selected then the reference sequence and at least one reads file must be selected.

5. Accept the default Bowtie options and click **OK**. Bowtie analyses can take some time depending on the reference and the number of Reads in your FastQ file(s). For example, assembly of 5 million paired end reads against a 5 Mb *E. coli* genome takes about 30 minutes on a Core 2 Duo laptop with 4 GB RAM.
6. You can close the progress window while the job completes.
7. When complete, a new "Reference Contig" will appear in your project window.

8. Double-click on the **<yoursequencename> Contig 1** item. This opens up the **Reference Contig Editor** – it shows the original reference sequence across the top and the aligned reads and consensus sequence below.
9. The **Summary** tab displays the summary statistics of the alignment. The **SNP** tab lists the differences between the Reference and the Consensus, along with any codon and amino acid changes if the SNP lies in a CDS feature. The **VCF** tab displays the raw VCF text data.
10. Back in the project window, click on the disclosure triangle next to the **<yoursequencename> Contig 1** item. This reveals all of the individual contigs within the reference assembly. You can double-click on any of these to open the corresponding **Contig Editor** – this shows just the consensus sequence and the aligned reads.

Tutorial

Creating and Populating a New Assembly Project

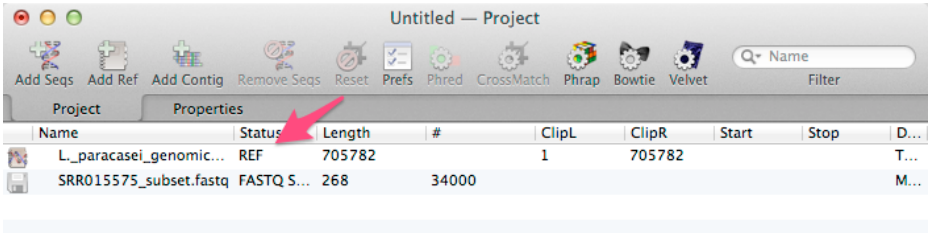
Start MacVector with Assembler and choose **File | New | Assembly Project.**

Click on the **Add Ref** button, select `L.paracasei genomic scaffold.nucl` and click **OPEN**.

Click on the **Add Seqs** button, select the `SRR015575_subset.fastq` and click **OPEN**.

You can also drag and drop Fastq reads files into the assembly project window. However, you must always use the **Add Ref** button to add reference sequences. Note that the **Add Contig** button is for adding existing BAM/SAM or .ace alignment files to a project.

Ensure that the `l.paracasei genomic scaffold.nucl` file has **REF** in the **Status** column (see screenshot below). This means it is a reference sequence.



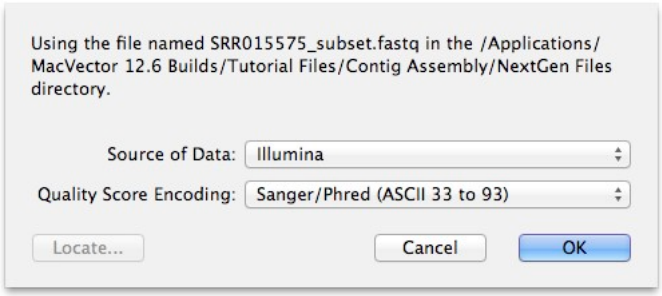
Now is a good opportunity to save the Assembly Project

Choose **File | Save As...** Select a suitable location and call the file `Tutorial Assembly`

MacVector uses a File Package to store the individual files of the Assembly Project. The actual project is saved as a BSML file, an XML-based format. This file will also contain trace files. Also note that the individual BAM files and BAM file indexes are also stored within the File Package. Right click and choose **Show Contents** to view the individual files.

Fastq files are added as a reference to the disk-based file rather than copying the entire file into the Assembly Project. If you double click on the file it will show the location of the original file. If the file has since been moved you need to use the **Locate** button to specify the location of the moved file.

Double click on the `SRR015575_subset.fastq` file in the project window.



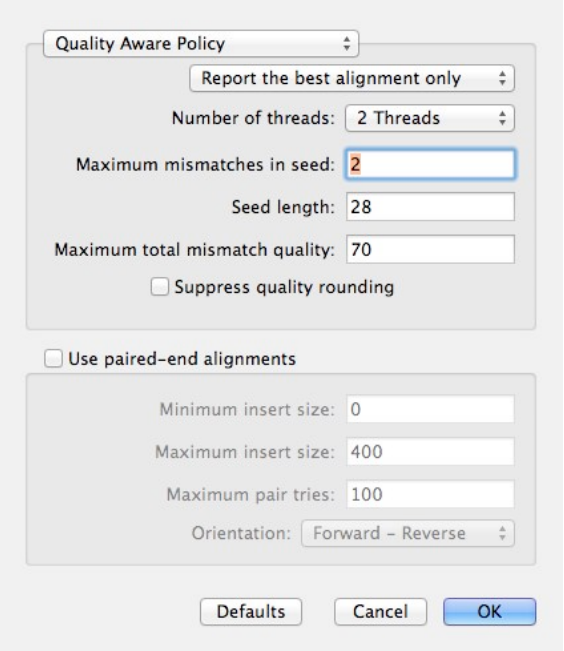
Click on **Cancel** to dismiss the dialog.

Running Bowtie

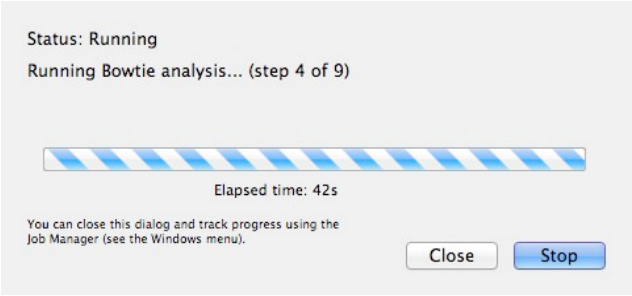
Select the reference and the read file in the Project window (hold down `<shift>` or use `<command>`-click to toggle selections). Click on the **Bowtie** icon on the toolbar.

You will see the Bowtie preference dialog. We will mostly use the defaults for this example. However, do read the section below called “Hit Reporting”. This contains important information on various settings when aligning against multiple references.

Change the popup menu to **Report the best alignment only**. If your Mac has a CPU with more than one core, you can change the **Number of threads** to a higher value (e.g. **2**). Click **OK**.



The Bowtie reference alignment goes through a number of steps. The first few steps will be performed very quickly but the assembly step and the generation of the VCF report (SNPs and other variants) may take a considerably longer time. The status of the alignment will be displayed in the status dialog or in the **Job Manager** if the status dialog is closed.



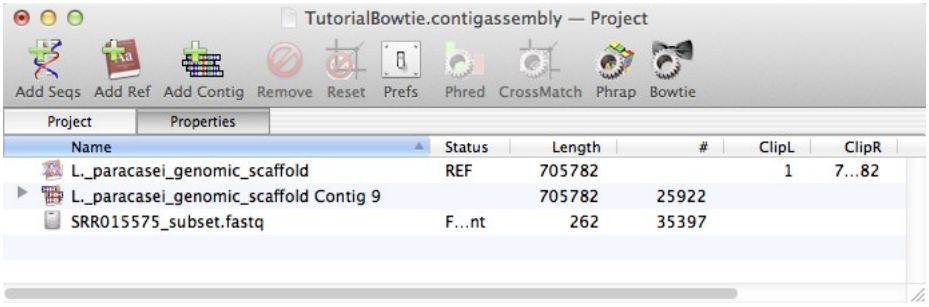
There are 9 steps to a Bowtie alignment;

1. **"Determining the read file encoding":** determines the common encoding for all the files.
2. **"Creating reference FASTA file":** creates the reference input sequence.
3. **"Creating read FASTQ files":** prepares the Fastq file(s) for submission to Bowtie
4. **"Running Bowtie analysis...":**
5. **"Extracting the consensus sequence and contigs"**: the `samtools` utility is run to create a consensus sequence.

6. **"Gathering unassembled reads"**: Any unmapped reads are placed in a single file.
7. **"Generating coverage data for <Reference Sequence>**
8. **"Generating contigs <Reference Sequence>"**
9. **"Generating SNP report for <Reference Sequence>:**

When complete click **view** in the Job Manager or Status Dialog and a new "Reference Contig" will appear in your project window. The original reference and read files are unchanged, so you can re-run the job with different parameters if you wish.

*Note that this in the screenshot below the **Name** field has been widened to accommodate the entire Reference Contig name:*



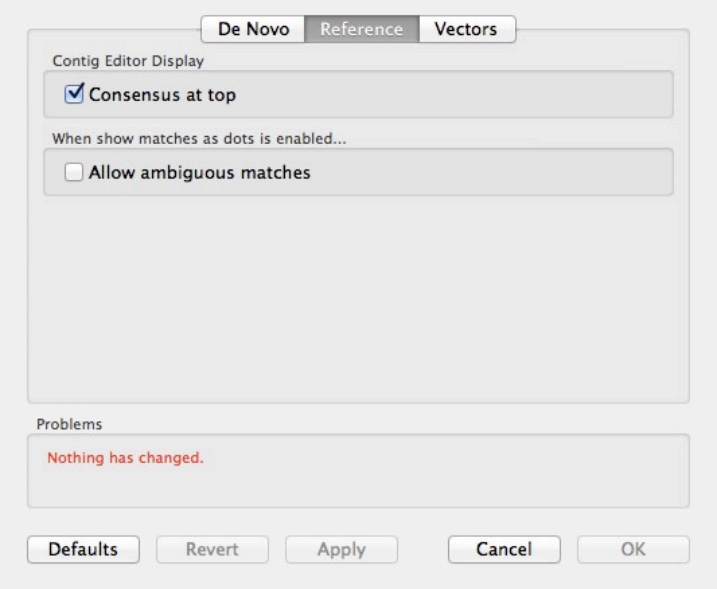
With the assembly project window active, choose **File I Save**

Now we will open the **Reference Contig** to see the results

Double-click on **L_paracasei_genomic_scaffold Contig 1**.

This opens up the **Reference Contig Editor** – it shows the original reference sequence across the top and the aligned reads and consensus sequence below.

Click on the **Prefs** toolbar button and check the **Consensus at top** box is checked and click **OK** or click **Cancel** if it is already checked.

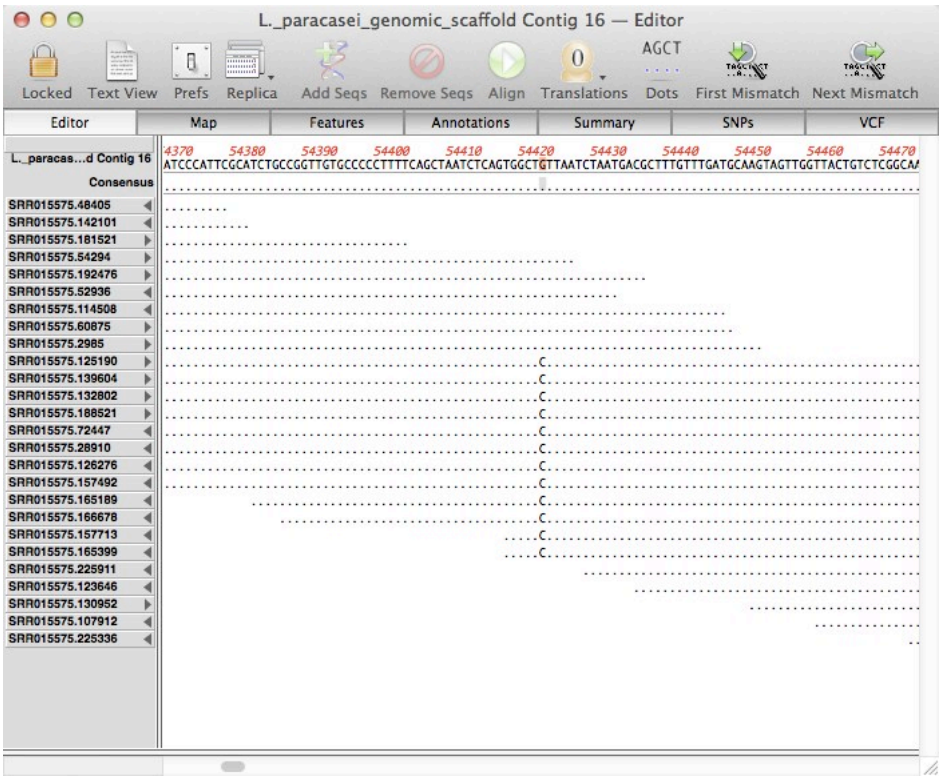


Click the **Dots** button in the toolbar

Now only residues that do not match the reference sequence will be shown.

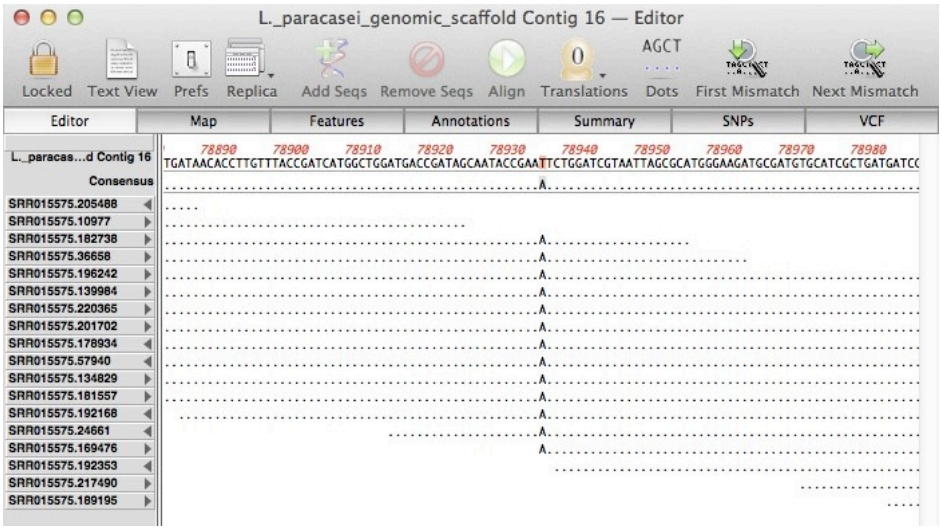
Now click **First Mismatch**

The cursor will move to and display the first location where the consensus does not agree with the reference sequence. In our example this is 54,420. Note that only a percentage of the reads are different at this point but sufficient to show a different base in the consensus.



Now click **Next Mismatch**

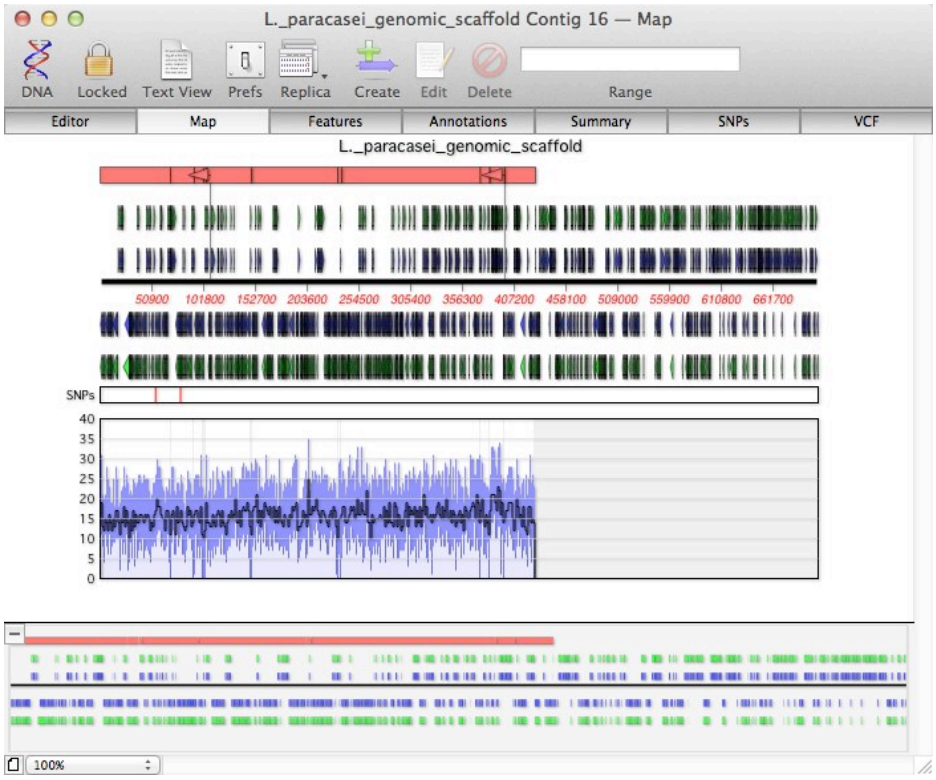
The cursor will move and display the next mismatch. This is at 78,934. Here all reads contain the different base.



Coverage Depth

Click on the **Map** tab

This shows a graphical representation of the reference contig. *NOTE:* for performance reasons individual reads are NOT shown in the **Map**.



Note the following features of the **Map** view from top to bottom:

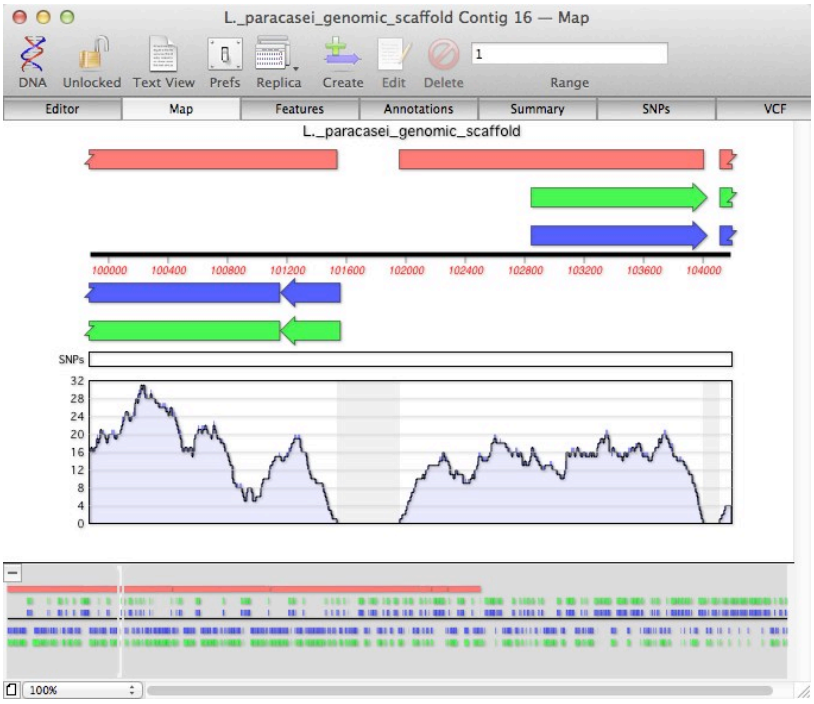
- Child contigs are annotated as **misc_feature** features in the reference contig. This will be saved if you export the reference contig as a single sequence MacVector file.
- The original annotation/features are shown above and below the sequence ruler line.
- Any SNPs found during the alignment are represented by a vertical line.
- The Coverage Depth plot shows the number of reads aligned against the template at that point (read depth). A single plot line (default color is black) shows a running average of the number of reads at that point, calculated using a moving window of dynamic length depending on the zoom level. The highest value in that window (default color is dark blue) and the lowest value (default color is light blue) are also shown.
- The Overview shows the full length of the contig with features and child contig features shown. If the main

sequence window is zoomed into a region this region will be marked on the overview.

Look at the region around 100,000. Note the two regions of zero coverage (in grey) although these regions may be as short as a single nucleotide in length, even at this level they are still visible.

Zoom into 100,000 to 104,000 by dragging the cursor along the map. Do this by positioning the cursor along the sequence line somewhere just 100,000. Hold down the mouse button, drag the cursor to just after the 104,000 marker then release the mouse button. You can do this multiple times until you have zoomed into the correct region. To reset the zoom level, simply double click on the white space background. You can also use the right and left cursor keys for fine adjustment of the viewed area.

The **Map** will now look similar to the following screenshot:

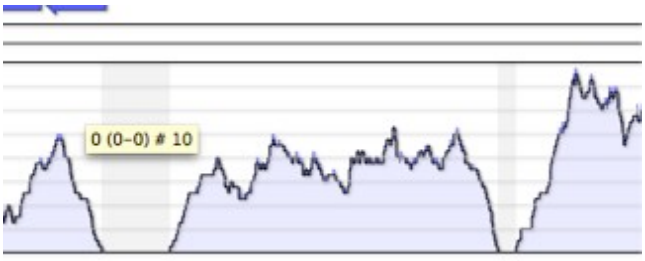


Note the following:

- The Overview now indicates the location of the region shown in the main window.
- The three plots in the coverage map have become very close as the window they are calculated from becomes shorter. Compare this with the initial coverage map showing the entire mapped region.
- There are two areas of zero coverage in the coverage map (shown with a grey background).

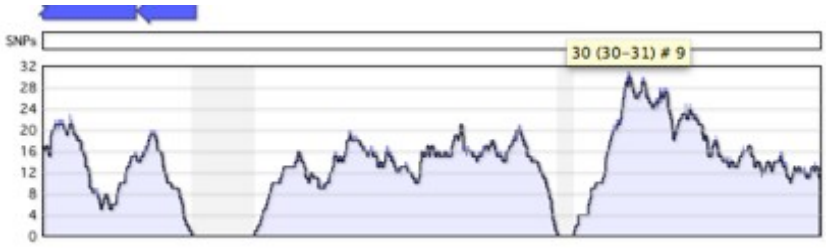
- The end of one child contigs, a full one and the start of a third are annotated (in red) on the reference contig. Note that these start and stop with the areas of zero coverage (a child contig is defined as a region of the reference sequence bounded by either end or a region of two or more bases with zero reads aligned.).
- Hovering the mouse over a child contig feature will show the start, stop and name of that child contig.
- The green features are genes that existed in the original reference sequence.
- If you hover the mouse over the coverage map it will give the exact number of reads at that position (for example X reads over base XX), along with the range of reads and the number of bases over which the average has been calculated.

Hover the mouse cursor over the larger area of zero coverage.



Note the tooltip showing zero values along with the window size.

Now hover the mouse over the highest peak



Note that the tooltip now shows the height of that peak along with the same window size.

Mapping Statistics

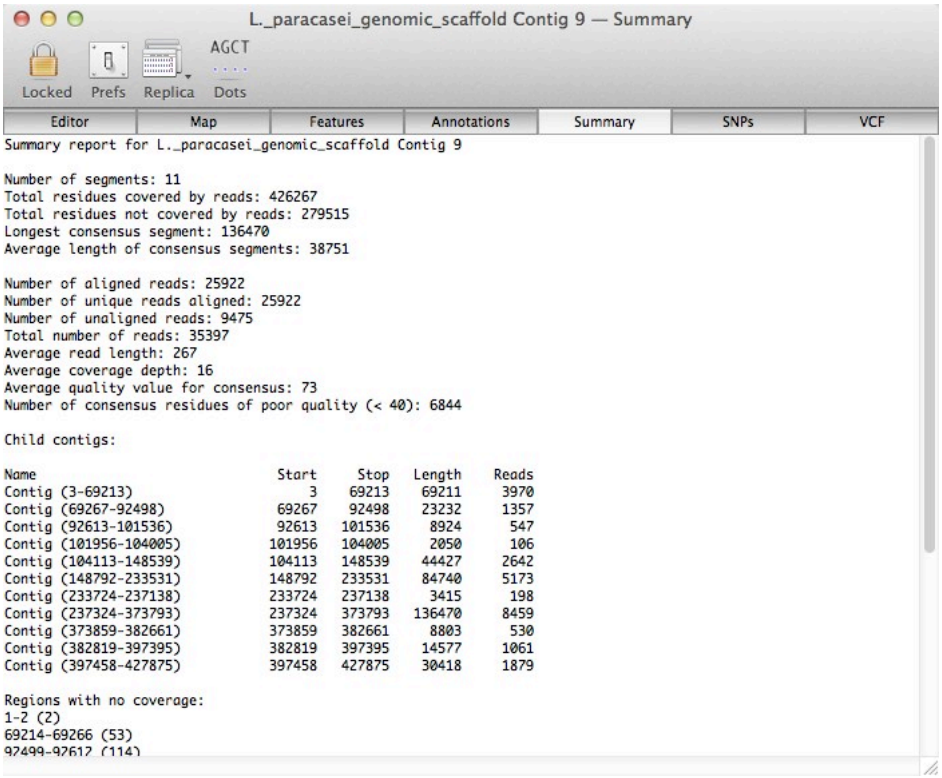
Click on the **Summary** tab.

The **Summary** tab displays a summary of the alignment.

Note the following:

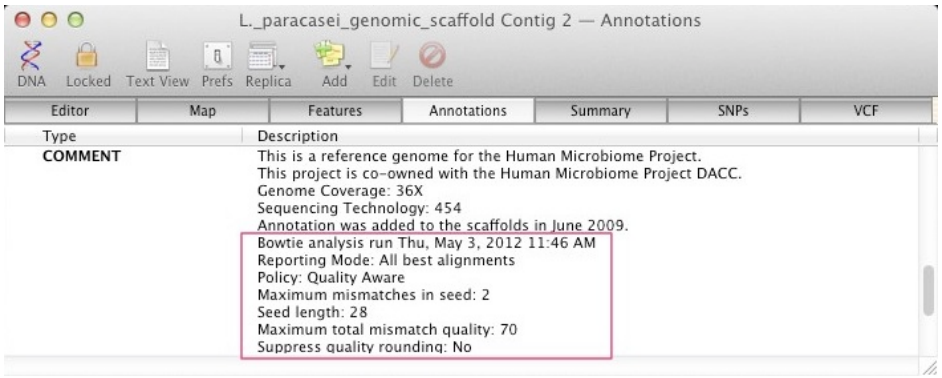
- All the child contigs are listed with details.
- Regions with no coverage are again listed.
- There is also a base composition table of the consensus.

This report can be saved using **File I Export As..**



Click on the **Annotations** tab and highlight the **COMMENT** annotation

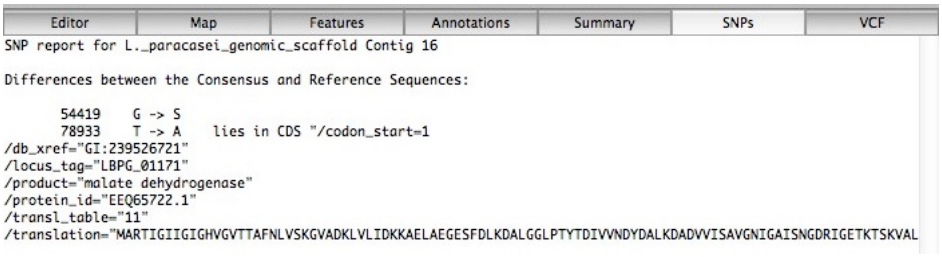
All the settings used to run Bowtie are added, along with the date and time of the run, to the Notes field of the Annotations tab.



Looking for SNPs and other variations

Assembler produces two reports on sequence variations found in the alignment.

Click on the **SNPs** tab



The SNP tab lists the differences between the Reference and the Consensus, along with any codon and amino acid changes if the SNP lies in a CDS feature.

Note that the two variations that we have already seen above are listed:

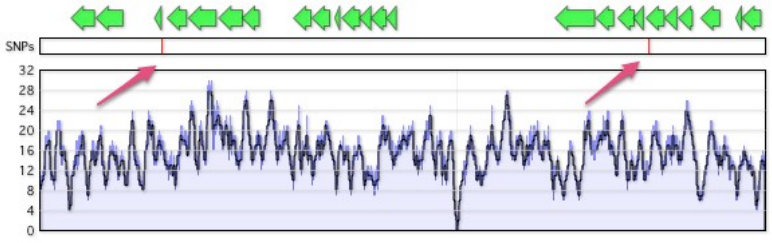
Click on the **VCF** tab

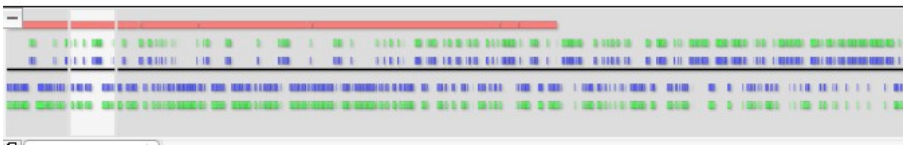
VCF or Variant Calling Format is a popular file format used to store and report variations found in a reference assembly. The VCF tab displays the raw VCF data that many other programs can use to evaluate SNPs in the data. Again our two SNPs are reported here.

You can export the VCF file using **File | Export As...** from this tab.

Switch back to the **Map** tab. Drag select to zoom into the section around the first vertical red line in the SNP line

Any reported SNPs are shown in the **Map** view.



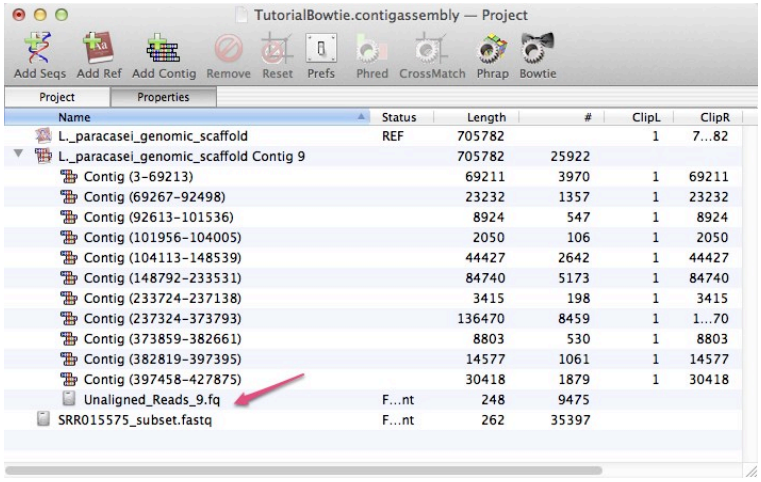


Child Contigs

Back in the Assembly Project window, click on the disclosure triangle next to the **L_paracasei_genomic_scaffold Contig** 1 item.

This reveals all of the individual child contigs within the reference assembly. You can double-click on any of these to open the corresponding **Contig Editor** – this shows jsut the consensus sequence and the aligned reads.

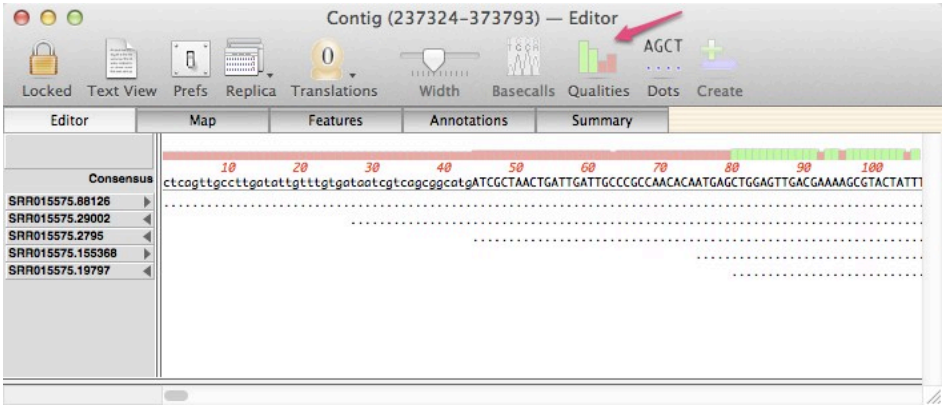
The unaligned reads are also shown (red arrow).

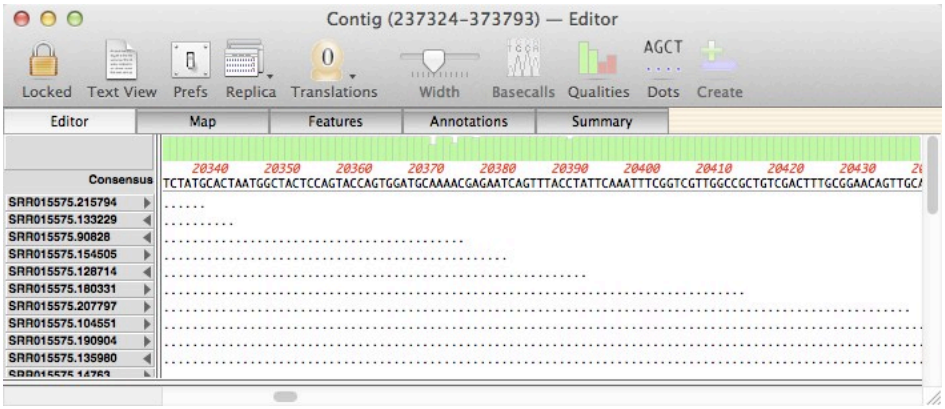


Double click on the longest contig (237,324 to 373,793).

Once open click on the **Qualities** button in the toolbar

Note that the beginning of the contig shows poor quality due to the low number of reads and their poor quality scores. However, as you scroll along the contig in this example the consensus will be shown to have a higher quality score due to the number of high quality reads that the consensus has been calculated from.





Further Analysis

Exporting Contigs

The **File | Export As..** menu option allows contigs and consensus sequences to be exported in FastA or FastQ formats.

- From the Project window if the reference contig is selected it will save a FastA or FastQ file containing all child contigs and no reference contig sequence or reference contig consensus. Selection of child contigs is ignored.
- From the Project window with only child contigs selected it will save a FastA or FastQ file containing all selected child contigs and no reference contig sequence or reference contig consensus.
- From an open Reference Contig it will save a multiple sequence FastA or FastQ file containing the reference sequence and the consensus.
- From an open Child Contig it will export a single sequence FastA or FastQ file with the consensus sequence.
- From the Project window with the unassembled reads file selected it will save a FastA or FastQ file containing all unaligned reads and no other sequences.

Working with sequences back in MacVector

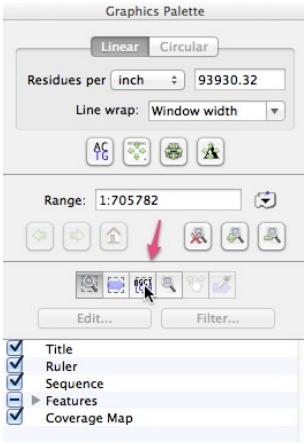
The coverage map makes it very easy to design primers for further sequence, for example, Sanger sequencing for hybrid assembly. Remember that you can run any MacVector analysis function directly on a contig and it will act as if you are running that analysis on a single sequence.

Here's how easy it is to design primers:

Zoom into an area of low coverage using the cursor in the reference contig.

First look for an area of low, or zero, coverage. Remember that areas of 2 or more bases with zero aligned reads are highlighted in grey and will be visible at all levels.

Click the **Select Sequence** button in the Graphics Palette



Now back in the Map View of the Reference Contig drag the cursor over the sequence spanning the low coverage region to select it.

Now select **Analyze | Primers | Design Primers (Primer3)....**

Check the popup menu in the dialog is set to **Amplify Feature/Region**. This will now take a 200bp region either side of your selected region and design primers to amplify this region.

Now you can amplify this sequence from your original sample, or instead design some sequencing primers and sequence it directly. For more information and practical examples of designing PCR and sequencing primers with MacVector, read the `Primer Design Tutorial.pdf` document in the `/MacVector 12.6/Documentation/` folder.

Important Concepts

Hit Reporting

In the dialog you'll see an important setting called Hit Reporting. Bowtie uses a concept of strata to score alignments. A stratum is defined by all reads that contain the same number of mismatches in the seed (the seed is the first "n" bases of a read which is given higher priority in scoring than the entire read). You can select from the following options;

All Alignments – each read can be aligned to multiple places in each reference sequence.

Report Best Alignment Only – each read can be aligned to just one location in one reference sequence.

Report All Best Alignments – each read can be aligned to just one location on each reference sequence, but can align to multiple references.

Which you choose depends on your aim in producing the alignment. If you are aligning a set of reads to a complete yeast genome with each chromosome supplied as a reference then you might choose **Report Best Alignment Only** as you would expect each read to come from one location somewhere on one of the chromosomes. However, if you are aligning a set of reads to a collection of genomic variant reference sequences then you might choose **Report All Best Alignments**. That is the equivalent of independently aligning the reads against each genome in turn. Choosing **All Alignments** ensures you don't miss any alignments, but you'll need to be aware that you may get over-reporting of hits to repeat regions such as rRNA operons or eukaryotic repetitive sequences.

Coverage Map

When you generate a reference contig with Bowtie, the **Map** view of a reference or child contig will show a plot of the depth of reads along the entire reference. This coverage map shows four statistics. A single plot line (default color is black) shows a running average of the number of reads at that point, calculated using a moving window of varying length depending on the zoom level. Such a plot is not sensitive when the window shows a large region of sequence at a high level, for example when viewing megabases of sequence). So, two shaded areas indicate the highest value (default color is dark blue) and the lowest value (default color is light blue) of the reads averaged for that window. When the coverage map is viewed at higher magnifications, the window from which the running average is calculated becomes shorter and so these three values will become closer. Eventually, when viewed at, or close to, residue level, these three plots will become identical.

Regions of zero coverage

Areas of zero coverage are shown in light grey. These areas are always displayed even when they are disproportionate to the level of magnification to ensure that you can spot even a single residue of no coverage in a 20 Mb contig.

Regions with low coverage

There are many reasons why regions will have lower than average coverage. These generally are caused by the base composition over that region. For example regulatory elements in a sequence, where proteins such as transcription factors bind, often display abnormally low coverage.

Regions with high coverage

Short regions with excessively high coverage are typically indicative of a repeated region, particularly if **All Alignments** is selected as the search mode. Reads will be piled up on one of the repeated sections rather than being spread out over each repeated region. Paired end reads can go some way to help detect these and allow correct alignment of reads.