Week 6: Mapping with Bowtie2 Version 2

Rika Anderson

Abstract

Citation: Rika Anderson Week 6: Mapping with Bowtie2. **protocols.io**
dx.doi.org/10.17504/protocols.io.g5zby76
Published: 04 May 2017

Protocol

Step 1.

Boot your computer as a Mac and use the Terminal to ssh in to Liverpool.

Mapping with a toy dataset

Step 2.

We're going to start by mapping the sequencing reads from a genome sequence of a type of archaeon (*Sulfolobus acidocaldarius*-- you've seen it before) against a scaffold from a very closely related species.

The sequencing reads from from one strain of *Sulfolobus acidocaldarius*, and the reference sequence that they are mapping to is from a very closely related strain of *Sulfolobus acidocaldarius*.

Make a new directory called “mapping,” then change into that directory.

cmd COMMAND

```
mkdir mapping
cd mapping
```

Mapping with a toy dataset

Step 3.

Copy this week's toy datasets into your directory. You're going to copy:

- the reference sequence (toy_dataset_contig_for_mapping.fasta)
- the reads that you will map to the reference (toy_dataset_reads_for_mapping.fasta)

cmd COMMAND

```
cp /usr/local/data/toy_dataset_reads_for_mapping.fasta .
cp /usr/local/data/toy_dataset_contig_for_mapping.fasta .
```

Mapping with a toy dataset

Step 4.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

The first thing you have to do is prepare an index of your reference dataset so that the mapping software can map to it.

- bowtie2-build is the program that indexes your reference.
- The first argument gives the reference dataset name.
- The second argument provides the name you want to give to the index.

cmd COMMAND
```
bowtie2-build toy_dataset_contig_for_mapping.fasta toy_dataset_contig_for_mapping.btindex
```

Mapping with a toy dataset

Step 5.

Now, map! First you make what is called a SAM file. It's a human-readable version of a BAM file, which we learned about previously in class.

- Bowtie2 is the name of the mapping program.
- “-x” is the flag that provides the name of the index you just made.
- “-f” means that the reads you are mapping are in fasta, not fastq, format.
- “-U” means that the reads are not paired.
- “-S” provides the name of your output file, which is in SAM format.

cmd COMMAND
```
bowtie2 -x toy_dataset_contig_for_mapping.btindex -f -
U toy_dataset_reads_for_mapping.fasta -S toy_dataset_mapped_species1.sam
```

Mapping with a toy dataset

Step 6.

If you look at the output file with less, you can see that it is human-readable (sort of). This can sometimes be useful if you want to parse it yourself with your own scripts-- but there's a whole suite of tools in a package called *samtools* that we'll rely on to do that next.

cmd COMMAND
```
less toy_dataset_mapped_species1.sam
```

Mapping with a toy dataset

Step 7.

Now you will use a package called samtools to convert the SAM file into a non-human-readable BAM file. You've heard of BAM files before-- now you get to make one.

- samtools is a package used to manipulate and work with mapping files. samtools view is one program within the whole samtools package.
- The flag “-bS” is not BS. It means convert a bam file to a sam file. (Bioinformatics jokes = not very funny.)

cmd COMMAND
```
samtools view -bS toy_dataset_mapped_species1.sam > toy_dataset_mapped_species1.bam
```

Mapping with a toy dataset

Step 8.

And because this is so fun, we get to do some more bookkeeping. Sort your bam file so that later programs have an easier time parsing it:

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

- “samtools sort” is the name of the program used for sorting
- The first argument provides the name of the bam file you want to sort
- The “-o” flag gives the name of the output file you want.

cmd COMMAND
```
samtools sort toy_dataset_mapped_species1.bam -o toy_dataset_mapped_species1_sorted.bam
```

Mapping with a toy dataset

Step 9.

In order to visualize your mapping, you have to index your reference genome. Yes, again. This time with samtools instead of bowtie2.

- Samtools faidx is the name of the program that indexes the reference.
- The first argument provides the name of the index, which should be your reference file.

cmd COMMAND
```
samtools faidx toy_dataset_contig_for_mapping.fasta
```

Mapping with a toy dataset

Step 10.

Almost there! Now you index the bam file that you just made:

- Samtools index is the name of the program that indexes the bam files.
- The first argument provides the name of a sorted bam file.

cmd COMMAND
```
samtools index toy_dataset_mapped_species1_sorted.bam
```

Mapping with a toy dataset

Step 11.

Now we're going to visualize this. Copy the entire "mapping" folder over to your local computer using either FileZilla or scp.

cmd COMMAND
```
scp -
r [your username]@liverpool.its.carleton.edu:/Accounts/[your username]/toy_dataset_director
y/mapping/ ~/Desktop
```
Remember, if you use scp, you should open a new Terminal window that is NOT logged in to liverpool. The above command copies the folder at toy_dataset_directory/mapping to your local computer's Desktop.

Mapping with a toy dataset

Step 12.

Find the IGV Viewer and open it. Click 'Genomes' --> 'Load Genome from File' and find your reference file. Then click 'File' --> 'Load from File' and open your sorted bam file. You should be able to visualize the mapping. Along the top, you'll see the coordinates of your reference sequence. Below that, you'll see a graph showing the coverage of each base pair along your reference sequence. Below that, you'll see each read mapped to each position. The arrows indicate the direction of the read; white reads are reads that mapped to two different locations in your reference. Single nucleotide variants in the reads are marked with colored letters; insertions are marked with a purple bracket, and deletions are marked with a horizontal black line. More information can be found at the link below.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

LINK:
http://software.broadinstitute.org/software/igv/AlignmentData

Mapping with a toy dataset

Step 13.

We're going to compare and contrast this mapping with another one. Now we're use the sequencing reads from a third very closely related strain of *Sulfolobus acidocaldarius*, and we're going to map those reads to the original reference sequence.

First, copy the second file to your directory (see below). Then, we will map these reads to the same reference file you used above, and then we will compare the mapping. You have already indexed the reference file, so you only need to repeat the steps that index the reads, and then map. All of those commands are listed below.

cmd COMMAND

```
cp /usr/local/data/toy_dataset_reads_for_mapping_species2.fasta .
bowtie2 -x toy_dataset_contig_for_mapping.btindex -f -
U toy_dataset_reads_for_mapping_species2.fasta -S toy_dataset_mapped_species2.sam
samtools view -bS toy_dataset_mapped_species2.sam > toy_dataset_mapped_species2.bam
samtools sort toy_dataset_mapped_species2.bam -o toy_dataset_mapped_species2_sorted.bam
samtools index toy_dataset_mapped_species2_sorted.bam
```

Mapping with a toy dataset

Step 14.

Copy the new data files to your local computer, and then visualize both of them in IGV viewer. Since you have already loaded the reference file and reads from your first mapping, all you have to do is click 'File' --> 'Load from File' and click on your new bam file. You should be able to see them side by side.

Questions to answer for this week's lab questions:

1. Describe the large-scale differences between the mapped reads from species 1 and species 2, and explain what this mapping tells us about the relative genome structure of the two genomes that we mapped. If we compared this genomic region in a dot plot, what would it look like? Describe at least one biological mechanism by which this may have occurred.

2. Do you see evidence of misassemblies? If so, describe where you see evidence for this and what this evidence looks like.

3. Do you see evidence of single nucleotide polymorphisms? If so, describe where you see evidence for this and what this evidence looks like.

Mapping your project datasets

Step 15.

Now we're going to map your project datasets. Remember that these are *metagenomes*, not a

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

genome, so the data will be a bit more complex to interpret.

We're going to map your raw reads against your assembled datasets. Why would we do this, you ask? A few reasons:

- to look for single nucleotide variants in specific genes.
- to quantify the relative abundances of different genes, and determine whether specific genes have better coverage than others.
- to quantify the relative abundances of specific taxa, and determine whether specific taxa are more abundant than others.

As you consider this, answer the following question for this week's lab questions:

4. If you wanted to quantify the relative abundances of specific genes in your sample, why couldn't you simply count the number of times your gene appears in your assembly?

Mapping your project datasets

Step 16.

Change directory into your project dataset directory folder. We're going to map your raw reads against your assemblies (not your ORFs). Make sure you know where your project assembly is and where your raw reads are. Follow the instructions to map your raw reads back to your assembled datasets. For example, if you were mapping the dataset ERR599166_1mill_sample.fasta and your assembly was called ERR599166_assembled.fa, you might do something like this. *Please be sure to use the assembled files that you've already run through anvi-script-reformat-fasta, which you should have done in our first computer lab.*

An example set of commands is shown below. Remember to replace the datasets here with your own project datasets!

cmd COMMAND

```
bowtie2-build ERR599166_assembled.fa ERR599166_assembled.btindex
bowtie2 -x ERR599166_assembled.btindex -f -U ERR599166_1mill_sample.fasta -
S ERR599166_mapped.sam
samtools view -bS ERR599166_mapped.sam > ERR599166_mapped.bam
samtools sort ERR599166_mapped.bam -o ERR599166_mapped_sorted.bam
samtools faidx ERR599166_assembled.fa
samtools index ERR599166_mapped_sorted.bam
```

Note that this command assumes that your raw reads and your assembly are in the same directory you're in. If they are not, you will need to either copy them over or use the correct path in your commands. (A reminder: the path simply gives directions to the computer for where to find a file.) For example, if you are in a mapping directory, your reads file is one directory up in the file hierarchy, and your assembled reads are in your assembly file, you might have to type something like this: bowtie2 -x ../assembly/ERR599166_assembled.btindex -f -U

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

../ERR599166_1mill_sample.fasta -S ERR599166_mapped.sam

Mapping your project datasets

Step 17.

When you visualize this in TGV, remember that you have multiple contigs. So you have to click the drop-down menu at the top and choose which contig you wish to visualize.

Question to answer for this week's lab questions:

5. Do you see evidence of single nucleotide variants? Provide an example (which contig, which position, what percent of reads had each base call). Biologically speaking, what does this indicate?

Mapping your project datasets

Step 18.

You were able to visualize the mappings in IGV, but sometimes you just want to have a number: for example, you might want to know the average coverage across a specific gene, and compare that to the average coverage of another gene in order to compare their relative abundances in the sample.

Go back to your Interproscan files and find two ORFs that you're interested in. Choose one that you think might be really abundant in a sample (a housekeeping gene, for example, that might be really common) and choose one that you think might be more specialized and only found in specific types of microbes.

Mapping your project datasets

Step 19.

Now we're going to make what's called a bed file. We will use it to find the average coverage of every single open reading frame in your dataset. Please make sure that your contigs have names that are something like c_00000000001 and your ORFs have names that are something like c_00000000001_1.

This will create a bed file that ends in .bed. You can take a look at it if you wish-- it should have the contig name, the coordinates of your ORF, and the name of your ORF.

cmd COMMAND

```
make_bed_file_from_ORF_file.py [your ORF file]
```

for example: make_bed_file_from_ORF_file.py ERR599166_assembled_ORFs.faa

Mapping your project datasets

Step 20.

Now run a script that will use your bed file and will calculate the read depth for every single ORF in your ORF file.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

cmd COMMAND

```
samtools bedcov [your bed file] [your sorted bam file] > [ an output file that ends in _ORF
_coverage.txt]
```

for example: samtools bedcov ERR599166_assembled.bed ERR599166_mapped_sorted.bam > ERR599166_ORF_coverage.txt

Mapping your project datasets

Step 21.

Open the file that it spits out. It should give the name of your contig, the start coordinate, the stop coordinate, the name of your open reading frame, and then the sum of the per-base coverage. To get the average coverage, open this document in Excel, and divide that number by the difference between the stop and start coordinates.

Now you have the average coverage of every ORF for this particular bam file, and you should be able to match the name of your ORF from Interproscan to the ORF in this output file.

Keep in mind that if you map a *different* metagenome against this same reference sequence (i.e. your own project assembly), you'll need to run this script again to get the coverage for *that* metagenome.

ANNOTATIONS

Dustin Michels 28 Oct 2017

If Excel doesn't automatically load your file correctly, with each field ending up in its own column, you can import it manually.

* open a blank excel document

* clck the 'data' tab, then the 'from text' button.

* select your file, and tweak the import settings until values are properly seperated.

To enter your formula and propogate it down the entire document:

* To to cell F1 (first row, right of the total coverage column.)

* Enter formula: `=E1/(C1-B1)`. That is total coverage divided by the different between the coordinates. You can click the desired cells rather than typing their names.

* Double click the green square in the bottom right of your cell to propogate that formula down the entire column.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Mapping your project datasets

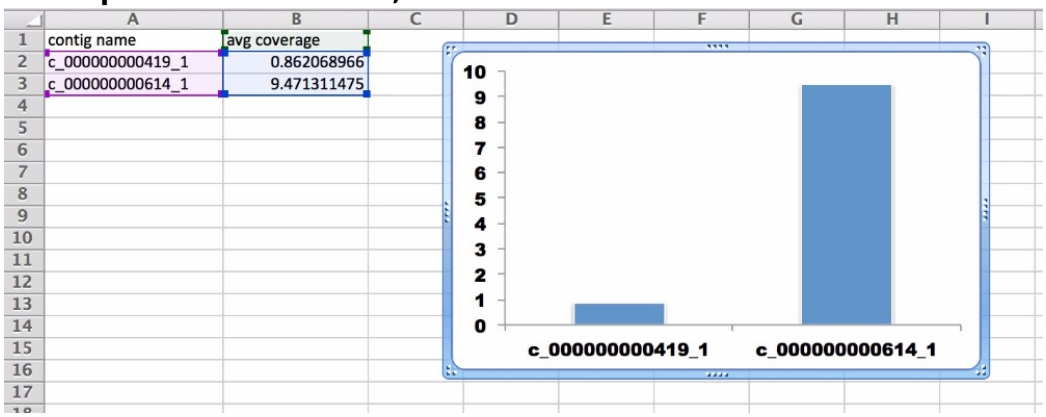
Step 22.

This is a really common type of analysis for 'omics-based studies-- you can compare the coverage of specific genes of interest. For example, you might compare the coverage of genes related to photosythesis, respiration, and nitrogen fixation if you're interested in how abundant those metabolisms are in the community. Or you might be interested in the relative coverage of ribosomal genes from archaea vs bacteria, for example, if you're interested in how abundant archaea are vs. bacteria.

For the writeup for lab this week, describe:

6. What ORFs did you choose to focus on? Which one did you expect to have higher coverage? Was your prediction correct?

7. Make a bar graph showing your results and submit it as 'Figure 1' for this week's lab writeup. I did one in Excel, like this:



8. As you scroll through the data file reporting the average coverage of all of your ORFs, which ORF had the *highest* coverage? What did it encode, according to your Interproscan file? Speculate on why that gene may have had the highest coverage of all the genes in your dataset.

Mapping your project datasets

Step 23.

Now, you're going to map the raw reads from each of the *other* datasets within your project region back to your assembled contigs. Basically, you'll be using the raw data that your other group members have been working with. This could be an interesting way for you to compare samples within your project study region with other members of your group. Each of you will map the raw reads of the *other* samples against your *own* assembled dataset.

Copy the file of raw reads from the *other* datasets over to your project directory, and map them to *your* assembled contigs using the commands above. So your reference sequences will be the same, but your reads for mapping will be different. Talk to your TA or to Rika if you need help.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Mapping your project datasets

Step 24.

Compare the mappings in IGV.

Question to answer for this week's lab questions:

9. Take a look at the mapping to the first contig. (Please include a screenshot of this mapping so I can see what you're looking at.)

a. Descibe patterns you observe in terms of differences in the relative coverage between the two metagenomes, and explain what this means biologically.

b. Provide an example of different patterns of single nucleotide variants between the metagenomes, and explain what this means biologically. What might this tell us about microbes of the same species living in these different habitats?

Mapping your project datasets

Step 25.

Mapping is often used as a way to compare the relative abundance of specific *types* of genes between regions. In the world of metagenomics, this can tell us something about how abundant a specific metabolism or function is within your microbial community. Use your Interproscan results to search for all ORFs related to a specific metabolism, function, or type of organism of your choice. For example, perhaps you are interested in the nitrogen cycle-- you might use Pfam to search for all genes related to nitrogen fixation, and identify them in your Interproscan results. Or perhaps you are interested in viruses-- you might look for all ORFs that have Interproscan hits that appear to be related to viral function (viral capsids, reverse transcriptases, etc).

For this week's lab questions:

10a. Write either a question or generate a hypothesis about the *relative coverage* of this set of genes with respect to your project datasets. This question/hypothesis should include a comparison between your own project dataset and another dataset within your project region, and it should be couched within the larger ecological context.

You do have some information to act as a basis for your question or hypothesis. For example, the KEGG metabolisms page you saw last week provides good information about which genes are used in specific pathways. You also have some metadata related to your project sample in the Project Dataset Excel spreadsheet on the Moodle.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Example #1:

I hypothesize that there will be lower coverage of genes related to photosynthesis (i.e. the psb genes) in the mesopelagic zone relative to the surface. This is because at the surface there will be more organisms that photosynthesize compared to the mesopelagic zone, where less light is available. Therefore, a lower proportion of genes in the microbial community in the mesopelagic zone will be related to photosynthesis compared to the surface, and therefore, fewer reads will map to photosynthesis genes in the mesopelagic zone.

Example #2:

I hypothesize that there will be higher coverage of genes that use nitrate as a substrate (i.e. the nar genes) in my sample relative to other samples in my project region because there is more nitrate in my sample (the surface) compared to the other depths in our project region (the deep chlorophyll maximum and the mesopelagic zone).

Once you've identified the set of genes related to a specific metabolism/function/type of organism, and you have written a question or generated a hypothesis, calculate the average coverage to each of those ORFs in your dataset. Then, contrast that with the average coverage of those genes from the mapping of at least one another dataset from within your region. Keep in mind that you are mapping *other* samples to your *own* assembly. Use bedtools and get_ORF_coverage.py to calculate the relative coverage of the genes you are investigating.

10b. Describe your results (including any caveats or variation you observed within your results) and create a graph to visualize those results. This should represent a mini 'Results' section in a lab report or paper.

10c. Interpret your results within the context of the ecosystem you are investigating. This should represent a mini 'Discussion' section in a lab report or paper.

Submit your responses to these questions on the Moodle by lab time next week.

NOTE!

Some of you are finding that other people's reads are NOT mapping well to your own

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

contigs. If you prefer, you can instead compare other people's mappings to their own datasets to your own mappings to your own dataset. (For example, if Sylvia has dataset A and Joyce has dataset B, they could look at the coverage of reads from A mapping to contigs from A, and reads from B matching to contigs from B, rather than trying to look at the coverage of reads from B mapping to contigs from A.)

This will require you all to share data with each other. So, please:

1) Make a directory in your Tara oceans project file with your name (i.e. /usr/local/data/Tara_datasets/Arabian_Sea/Rika, for example)

2) Copy ALL of your: bam files, bai files, and your ORF_coverage.txt files over to your new directory

3) Make sure your assembled contigs, ORF files, and Interproscan .tsv files are already in your Tara project folder. (They should be there already.)

Now you'll have access to each other's datasets in case you'd like to use their datasets to compare relative coverage.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited