Basic Read Alignment

Dan MacLean

2019-11-29

Contents

Chapter 1

About this course

In this short course we'll look at a method for running an alignment of sequence reads against a reference genome. The course is very brief and will show you how to use a program called `minimap2` and `samtools` to create a binary alignment file that you can use in further work.

I acknowledge that there are lots of other programs and methods - this course is *not* meant to be comprehensive, it is meant to get you being productive. Seek out further advice if you need to run other programs. Do be encouraged though, 99 % of what you learn here will be applicable to other tools for the same job (they all run in a *very* similar manner) so this is a good place to start.

The course has two main parts - first we'll learn to do this 'locally', that is to say on the computer that you are actually physically sitting at and have direct control over. Once we've done that and know how to run the actual programs then we shall switch to running the programs 'remotely' on a HPC environment using a submission system that you have to log in to.

1.1 Prerequisites

This course assumes that you are familiar with the basics of running stuff from a command-line. You'll need to have some experience not lots. If you've done the TSL command-line course you'll know plenty

For the first 'local' part of this you'll need some software on your machines. Most bioinformatics software has to run on Unix style computers, which for most of us means Macs. The installation instructions below only apply to Macs.

For the second part you'll need an HPC account. See the Bioinformatics Team to get one of these.

1.1.1 Local Software Installation

We need to install `minimap2` and `samtools`. Installing bioinformatics software is often *not* straightforward so we'll take a path of least resistance and install some tools that manage software for us. This is a roundabout way of doing things, but it greatly simplifies the hard parts and means that we can isolate our installations from the rest of our computer and not mess up anything already installed.

1. **Get `conda`** `conda` is an environment and package manager for software projects (initially Python - hence the name). Its purpose is to create a sandbox area on your computer where you can safely install software without it interfering or overwriting any of the existing software already on there. This safe space is called a conda 'environment'. To install `conda`:
    1. Click this link https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.pkg and wait for the package to download. When it has double-click it and go through the installation process.
2. **Get `bioconda`** `bioconda` is plugin that makes `conda` aware of the bioinformatics software we will need. To make `conda` aware of `bioconda`
    1. Open `Terminal` and type the following:

```
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

3. **Create a new environment** We can now create our new safe space environment. Type in the following.

```
conda create --name alignment_env
```

This step can take a moment or two and nothing seems to happen for a while - be patient. Accept the defaults (answer `y` when questioned).

4. **Activate the environment** Everytime we want to use the new environment, including to install something into it, we have to activate it. That means each time we leave or start a new Terminal we need to type this into it

```
conda activate alignment_env
```

You should see the name at the `$` prompt change, telling you that this Terminal is in the `alignment_env` environment. If you ever can't find programs that you're sure you installed, it means that you probably didn't activate the right environment.

5. **Install the alignment software `minimap2` and `samtools`** All the steps up to now have been so we can do this one! Install each of the pieces of software in turn by using `conda` with:

```
conda install minimap2
```

```
conda install samtools
```

These steps can also take a while. Again accept the defaults (answer `y`).

Now you are done! Everything is installed ready for you to work with. Next we need to get the sample data

1.1.2 Sample reference genome and reads

You'll need this E.coli reference genome file and this set of paired end reads. Download them and put them into a folder on your machine. I suggest `Desktop/align_tut`. I'll assume that is the directory you will work from.

That's all you need to do the lesson. If you have any problems getting this going, then ask someone in the Bioinformatics Team and we'll help.

Chapter 2

The Alignment Pipeline

In this chapter we'll look at an overview standard paired-end read run of `minimap2`, what it outputs and how to manipulate the output with `samtools`.

The overall pipeline is very straightforward…

1. Align all reads to a reference
2. Filter badly scoring reads
3. Sort and compress output

2.1 Align all reads to a reference

This is the main step, and with `minimap2` it can be accomplished with a single command-line. In this step each read is aligned against the reference, and its best aligning position found. That position, along with a metric of the quality of the single alignment is reported in a `SAM` format file.

2.2 Filter out badly scoring reads

This is the quality control step. We remove reads that don't have a good alignment score because in most contexts it means the read is a bad read with bad sequence in it. Of course in some contexts it isn't - it depends what you're aligning to what, but for the RNAseq situation or SNP calling situation where we expect the reads to be very like the reference this is appropriate. This step is done with `samtools`

2.3 Sort and compress output

Once filtering is done and we have the set of reads we wish to retain we can take our output file and convert it to a sorted binary format that uses less disk space and is optimised for searching in downstream analysis. This step is a kind of housekeeping step that makes everything later easier. We do it with `samtools`

2.4 Further Reading

1. FastQ quality scores
2. Alignment scores
3. Why minimap2 and not bwa bowtie

Chapter 3

Running `minimap2`

Running `minimap2` takes only one step. Assuming we've already `cd`'d into the directory with the reads and reference we can use this command

```
minimap2 -ax sr ecoli_genome.fa ecoli_left_R1.fq ecoli_right_R2.fq > aln.sam
```

Try running that and see what happens… You should get an output file in the working directory called `aln.sam`. On my machine this takes just a few seconds to run.

Let's look at the command in detail.

3.1 The `minimap2` command and options

First we get this

```
minimap2
```

which is the name of the actual program we intend to run, so it isn't surprising that it comes first. The rest of the command are options (sometimes called arguments) telling the program how to behave and what it needs to know. Next up is this

```
-ax sr
```

which gives option `a` meaning print out SAM format data. And option `x` meaningwe wish to use a preset parameter set. The preset we wish to use comes after `x` and is `sr`, which stands for `short reads` and tells `minimap2` to use settings for short reads against a long genome. Next is this

```
ecoli_genome.fa ecoli_left_R1.fq ecoli_right_R2.fq
```

which are the input files in the 'reference' 'left read' 'right read' order. Finally, we have

```
> aln.sam
```

which is the `>` output redirect operator and the name of an output file to write to. This bit specifies where the output goes.

So the structure of the `minimap2` command (like many other commands) is simply `program_name options input output`.

And this one command is all we need for a basic alignment with `minimap2`. We can now move on to the next step in the pipeline.

3.2 Further Reading

The `>` operator. `minimap2` further instructions and github

Chapter 4

Filtering Badly Aligned Reads

Once we have an alignment, the next step is often to throw out the reads that align badly or not in pairs as we we expect. To do this we need to look at the alignments and assess them one-by-one. We'll need first to have some understanding of the output from our alignment, in this case `aln.sam` a SAM format file.

4.1 SAM Format

Alignments are generally stored in SAM format, a standard for describing how each read aligned one-by-one. Each line carries the results for a single read. Let's examine a single reads alignment. Recall that we can look at one line in a file called `aln.sam` using `tail -n 1 aln.sam` (this gives the bottom line in the file). Running this prints the following

```
NC_011750.1_1004492_1005000_1:0:0_3:0:0_1869f   147 NC_011750.1 1004931 33  70M =   1004492 -509
```

On close inspection we can see this mess (which is only a single line) contains things like the read name, the position it maps to on the reference sequence, the read sequence, and lots of other strange things like `70M` and `de:f:0.0429`. The important thing to note is that these weird things are encoded quality information for this alignment, so we can - if we know how to manipulate those codes - select read alignment of the proper quality.

Thankfully the program `samtools` makes this easy for us.

4.2 samtools

We can accomplish read filtering with the following command.

```
samtools view -S -h -q 25 -f 3 aln.sam > aln.filtered.sam
```

Try running that and looking at the output file that is generated. You should have another SAM format file called `aln.filtered.sam` in your working directory.

Let's take a look at that command in detail

4.3 The samtools command and options

Straight away, the command seems to fit the familiar `program name options files` pattern. It starts with

```
samtools
```

which is the program name. Then we get the options

```
view -S -h -q 25 -f 3
```

The first option to `samtools` must be the name of the sub-program to run. There are lots of these as `samtools` is a suite of sub-programs. `view` is the option for working with alignments directly. The second option `-S` tells `samtools view` that we are handing it a SAM format file (soon we will hand it a different type) and `-h` tells it to show the header as well (each SAM file has a header that we sometimes don't want). The next two options are the important ones. `-q 25` will remove reads with a mapping quality (a measure of how well a read is aligned) lower than 25 (a reasonable score) and `-f 3` is a 'flag' a really complex way of encoding alignment attributes (see Further Reading for more details). The important thing is that `3` means `keep reads that are paired and whose pair is mapped too`.

At the end of the command is the input and output file information

```
aln.sam > aln.filtered.sam
```

which means the input file is our `aln.sam` and that the output should be redirected to `aln.filtered.sam`

4.4 Checking the filtering

As an exercise to show that we did filter stuff out lets compare the input `aln.sam` file with the output `aln.filtered.sam` file. Recall that `wc -l` will give us the number of lines in a text file. Run it like this, on both files at once

```
wc -l aln.sam aln.filtered.sam
```

I get this as output

```
  200002 aln.sam
  166905 aln.filtered.sam
  366907 total
```

The number of lines (alignments) in the filtered files is less than that in the unfiltered, so we can casually assume the command worked.

And that's all there is to getting the reads filtered. In real-life you have many options for filtering and you may choose to do it at other points (for instance, lots of RNAseq quantification programs will allow you to filter when you use them), but the process will be similar and take advantage of the same mapping quality and flag metrics you've been introduced to here.

4.5 Are we done?

On the face of it then, it looks like we've come to the end of what we intended to do - we did an alignment, and we've filtered out the poor ones. In practice though, we'll be dealing with many millions of reads, many files of many Gb size. This complicates the housekeeping we have to do, not the procedure we've learned *per se*, so before we jump to the HPC we need to look at that. That's the next chapter.

4.6 Further Reading

4.6.1 SAM Format

I only really alluded to the SAM format above, but there's a lot to it. This Wikipedia page gives a lot of detail.

4.6.2 Mapping Quality

A metric that describes how well overall the read aligned, it takes into account not just the alignment, but the nubmer of other possible alignments that were rejected. Consider that a read mapping well equally at a number of places in the genome cannot be said to be mapping well at all. Different aligners make arbitrary decisions about how to score such alignments. See this short summary for information on how it can be calculated.

4.6.3 Flags

The flags option is the most powerful way to describe a filter to `samtools view`, it is also really complicated. The number you pass (e.g `-f 3`) is calculated as a sum of lots of options. The way they're are described in the documentation is a bit more complex than I want to go into, but there are helpful web-apps that can simplify things - try this one

Chapter 5

Connecting Programs and Compressing output

Now that we've been through the whole alignment and filtering pipeline, let's look at the output. Specifically lets compare the sizes of the files we used. Recall that we can do that with `ls -alh`

On my folder I get this (some columns and files removed for clarity)

```
49M 29 Nov 10:46 aln.filtered.sam
59M 28 Nov 16:28 aln.sam
5.0M  2 Jul 15:04 ecoli_genome.fa
18M 28 Nov 15:53 ecoli_left_R1.fq
18M 28 Nov 15:53 ecoli_right_R2.fq
```

The file sizes are in the left-most column. Check out the relative size of the two read files (18M each) and the alignment SAM files (59M and 49M). The output file is much larger than the input. This has implications for storage when the files are really large (many GB) and there are lots of them. The disk space gets used really quickly. Consider also the redundancy we have - that `aln.filtered.sam` is the one we're interested in, not the `aln.sam` so it is taking up unnecesary disk space. It's easy to see that when you are doing a real experiment with lots of samples and hundreds of GB file size, you're going to eat up disk space. Also larger files take longer to process, so you're going to have a long wait. This has implications too when you get to later stages in the analysis

In this chapter we're going to look at a technique for reducing those housekeeping overheads and speeding things up.

5.1 BAM Files

BAM files are a binary compressed version of SAM files. They contain identical information in a more computer friendly way. This means that people can't read it, but it is rare in practice that you'll directly read much of a SAM file with your own eyes. Let's look at the command to do that

```
samtools view -S -b aln.filtered.sam > aln.filtered.bam
```

Again we're using `samtools view` and our options are `-S` which means SAM format input and the new one is `-b` means BAM format output. Our input file is `aln.filtered` and we're sending the output to `aln.filtered.bam`.

If we check the files with `ls -alh` now we get

```
9.2M 29 Nov 14:05 aln.filtered.bam
49M 29 Nov 10:46 aln.filtered.sam
59M 28 Nov 16:28 aln.sam
5.0M  2 Jul 15:04 ecoli_genome.fa
18M 28 Nov 15:53 ecoli_left_R1.fq
18M 28 Nov 15:53 ecoli_right_R2.fq
```

The BAM file is about a fifth of the size of the SAM file. So we can save space in this way. We have another trick up our sleeve though. We can connect together command lines, so that we don't have to create intermediate files - this reduces the number of files we have to save. We can do this by using something called pipes.

5.2 Connecting Program Input and Output With Pipes

Most command line programs print their results straight out without sending it to a file. This seems strange, but it adds a lot of flexibility. If also set up our programs to read in this output then we can connect them together. We can do this with pipes. The usual way to do this is to use the `|` operator. Let's look at a common example.

Here we'll use the command `ls` and `shuf` to see how this works. We know `ls` will 'list' our directory contents, `shuf` shuffles lines of text sent to it. If we use `|` in between we can connect the output of one to the other. Try running `ls` a couple of times to verify you get the same output both times and then try this a few times

```
ls | shuf
```

you should get different output everytime. The important thing to note is that `shuf` is doing its job on the data sent from `ls`, which sends consistent data every

time. We don't have to create an intermediate file for `shuf` to work from. The `|` character joing two commands is the key.

We can apply this to our `minimap2` and `samtools` commands.

5.3 From reads to filtered alignments in one step

So let's try reducing the original alignment pipeline to one step with pipes. We'll work in the BAM file bit later.

Simply take away the output file names (except the last one!) and replace with pipes. It looks like this

```
minimap2 -ax sr ecoli_genome.fa ecoli_left_R1.fq ecoli_right_R2.fq | samtools view -S -h -q 25 -f
```

when you do `ls -alh` you should see the new `aln.filtered.from_pipes.sam` file, its size is identical to the file we generated when we created the intermediate `aln.sam` file, but this time we didnt need to, saving that disk space.

5.3.1 From reads to filtered alignments in a BAM file in one step

Let's modify the command to give us BAM not SAM, saving a further step. We already know that `samtools view` can output BAM instead of SAM, so lets add that option (`-b`) in to the `samtools` part.

```
minimap2 -ax sr ecoli_genome.fa ecoli_left_R1.fq ecoli_right_R2.fq | samtools view -S -h -b -q 25
```

If you check the files with `ls -alh` now you should see that you have the new `aln.filtered.from_pipes.bam` file with no extra intermediate file and the smallest possible output file. Congratulations, you know now the fastest and most optimal way to make alignments and filter them.

5.4 Sorting BAM files

In practice a BAM file of alignments needs to be ordered with the alignments at the start of the first chromosome at the start of the file and the alignments on the end of the last chromosome at the end of the file. This is for computational reasons we don't need to worry about, but it does mean we need to do another sorting step to make our files useful downstream.

Because all the alignments need to be present before we can start we can't use the pipe technique above. So we use an input and output file. The command is `samtools sort` and looks like this.

```
samtools sort aln.filtered.from_pipes.bam -o aln.filtered.from_pipes.sorted.bam
```

Doing `ls -alh` shows a new sorted BAM `aln.filtered.from_pipes.sorted.bam` that contains the same information but is actually a little smaller due to being sorted. We can safely delete the unsorted version of the BAM file.

5.4.1 Automatically deleting the unsorted BAM

If the sorting goes fine, we have two BAM files with essentially the same information and don't need the unsorted file. We can of course remove this with `rm aln.filtered.from_pipes`. A neat space saving trick is to combine the `rm` step with the successful completion of the sort. We can do this by joining the commands with `&&`.

That looks like this

```
samtools sort aln.filtered.from_pipes.bam -o aln.filtered.from_pipes.sorted.bam && rm a
```

The `&&` doesn't connect the data between the two commands, it just doesn't let the second one start until the first one finishes successfully (computers have an internal concept of whether a command finished properly).

This means if the `samtools sort` goes wrong the `rm` part will not run and the input file won't be deleted so you won't have to remake it. This is especially useful later when we wrap all this into an automatic script.

5.5 Indexing the sorted BAM

Many downstream applications need the BAM file to have an index, so they can quickly jump to a particular part of the reference chromosome. This is a tiny file and we usually don't need to worry about it. To generate it use `samtools index`

```
samtools index aln.filtered.from_pipes.sorted.bam
```

Using `ls -lah` we can see a tiny file called `aln.filtered.from_pipes.sorted.bam.bai`, this is the index.

5.6 Further Reading

For a primer on some more aspects of `samtools` see this tutorial

Chapter 6

Automating The Process

We now know everything we need to do an alignment of reads against a reference in an efficient way. What's next is to consider that this process needs to be done for every set of reads you might generate. That's a lot of typing of the same thing over and over, which can get tedious. In this section we'll look at how we can automate the process to make it less repetitive using a script.

6.1 Shell scripts

Scripts that contain commands we usually run in the Terminal are called shell scripts. They're generally just the command we want to do one after another and saved in a file. We can then run that file as if it were a command and all the commands we put in the file are

Shell scripts must be a simple text file, so you can't create them in programs like Word, you'll need a special text editor. On most systems we have one called `nano` built into the Terminal.

6.1.1 Using `nano` to create a shell script

To open a file in `nano` type `nano` and the name of the file, if the file doesn't exist it will be created.

```
nano my_script.sh
```

Will create a file and open it. To save and exit type press `Ctrl` then `X` (thats what `^X` means in the help at the bottom. You can enter your script in here. Remember its not a word processor, its a Terminal text editor, so you have to use the mouse to move round and cutting and pasting is a bit clunky.

6.2 Creating a script that automates our alignment pipeline.

Let's enter our script into `nano`. We'll do it as we did in the earlier chapters, but we'll change file names to make it clear which files are coming from the script.

First, create a script called `do_aln.sh`

`nano do_aln.sh`

Once `nano` opens, add the following into it

```
minimap2 -ax sr ecoli_genome.fa ecoli_left_R1.fq ecoli_right_R2.fq | samtools view -S -
samtools sort aln.script.bam -o aln.script.sorted.bam && rm aln.script.bam
samtools index aln.script.sorted.bam
```

That's all the steps we want to do. Use `Ctrl-X` to save the changes to the file.

6.3 Running the script

To run the script we use the `sh` command and the script name. Try

`sh do_aln.sh`

You should see progress from the script as it does each step in turn. When it's done you can `ls -alh` to see the new sorted BAM file from the script.

Congratulations! You just automated an entire analysis pipeline!

6.4 Running on different input files

So our script is great but the input filenames will be the same every time we run it meaning we'd need to go through the whole file and change them which is error prone. Also the output files are the same each time, meaning we could accidentally overwrite any previous work in there, which is frustrating. We can overcome this with a couple of simple changes in our script that make use of variables.

Variables are place holders for values that the script will replace when it runs. Consider these two commands

```
MY_MESSAGE="Hello, world!"
echo $MY_MESSAGE
```

Recall that `echo` just prints whatever follows it. Try running this, you get `Hello, world!` which shows that the process created a variable called `MY_MESSAGE` and stored the message in it. When used by a command the `$`

showed the command that it should use the message stored in the variable and printed `Hello, world!`. We can use this technique in our scripts. Note the command `MY_MESSAGE="Hello, world!"` must not have spaces around the equals sign.

Now we can expand our script to take advantage. Look at this script.

```
LEFT_READS="ecoli_left_R1.fq"
RIGHT_READS="ecoli_right_R2.fq"
REFERENCE_SEQUENCE="ecoli_genome.fa"
SAMPLE_NAME="ecoli"

minimap2 -ax sr $REFERENCE_SEQUENCE $LEFT_READS $RIGHT_READS | samtools view -S -h -b -q 25 -f 3
samtools sort $SAMPLE_NAME.bam -o $SAMPLE_NAME.sorted.bam && rm $SAMPLE_NAME.bam
samtools index $SAMPLE_NAME.sorted.bam
```

Right at the top we create a variable for each of our read files (`LEFT_READS` and `RIGHT_READS`), our reference files (`REFERENCE_SEQUENCE`) and a unique sample name (`ecoli`). These variables get used whenever we need them, saving us from typing the information over and over. The practical upshot of this being that we only need to change the script in one place every time we reuse it for a different sample and set of reads.

Now try this out. Save the new script in a file called `do_aln_variables.sh` and run it as before with `sh do_aln_variables.sh`. When it's run you should see an output called `ecoli.sorted.bam`.