**MCB 525, Fall 2016**

**2.2: Reference-guided alignment**

In this section, we will create a 2bRAD reference from the *Aiptasia* reference genome and map high-quality reads to the 2bRAD reference.

**Objectives:**

* Gain familiarity with fasta and sam file formats
* Perform reference-guided alignment

**Protocol:**

1. First, we will need to obtain the *Aiptasia* reference genome. The genome scaffolds are freely available on the web at <http://aiptasia.reefgenomics.org/download/>. From your home directory, make a new folder to contain the reference:

$ mkdir reference

$ cd reference/

$ wget <http://aiptasia.reefgenomics.org/download/aiptasia_genome.scaffolds.fa.gz>

#annotation info is also available for download at this site, but we want the genome scaffolds file.

1. Unzip the genome file and take a look. This time let’s use the less command instead of the head command – it’s a little more wieldy when individual lines of text are very long.

Use the space bar to scroll through the file, and CTRL+Z to exit.

$ less aiptasia\_genome\_final.fa

1. Notice the format of this fasta file. Each sequence in a fasta file is associated with at least two lines. The first line is the sequence header, denoted by a ‘>’, and the next lines (until the next header line) give the sequence:

>scaffold1size1833262

GATGGGGTTATAGTCAATGGTCTGATAATGTAGTGTATTTGATAATTGGCGATGGGCTTCCTGTGTGTACATGTCAACAGGCCACAAAACAAGAGAGCCACCTTTGTCAGCCGGTTTAATGACGGTGTCTTGGAGTCTAGAGAGATGGTGTAAAGCCGTTCTTTCTTCCCGCGAGATGTTGGTTTTGTGTGTGTTG

1. How many fasta sequences are in this file? To count the number of sequence headers in the file:

$ grep –c ‘>’ aiptasia\_genome\_final.fa

1. The *Aiptasia* genome reference contains many ‘contigs’, or long stretches of assembled sequence. We want to map reads from our 2bRAD assembly to the reference, but we expect these reads to match only a specific fraction of the genome. To reduce mapping time and generate a reference that contains only expected 2bRAD fragments, we will use the BcgIExtract.pl script (also available through links on the Meyer Lab website). To generate a 2bRAD reference:

$ SGE\_Batch –c ‘../scripts/BcgIExtract.pl aiptasia\_genome\_final.fa BcgI\_ref.fa’ –r em1

$ grep –c ‘>’ BcgI\_ref.fa

$ head BcgI\_ref.fa

#How many sequences are in the BcgI reference file? How long is each sequence?

#if everything looks good, go ahead and ‘index’ the reference with samtools faidx. Samtools faidx will create several additional files we will need later on…

$ samtools faidx BcgI\_ref.fa

1. Check the output of the BcgIExtract.pl script:
2. Now, we are ready to map high quality reads to the BcgI reference. There are many different aligners we could use (e.g. bwa, bowtie, etc.), but SHRiMP seems to work particularly well with 2bRAD data. Run the gmapper tool of the SHRiMP package using the following settings:

$ cd ../CC7/

$ SGE\_Batch –c ‘gmapper --qv-offset 33 -Q --strata -o 3 -N 1 CC7\_cleanreads.fastq ../reference/BcgI\_ref.fa >CC7.sam’ –r my\_sge\_run\_name

#you can monitor the status of this job by typing qstat

Now you should have a file of the aligned reads output from gmapper. The output is in 'sam' format, a common output format for reference-guided alignment. More on sam format can be found at <https://samtools.github.io/hts-specs/SAMv1.pdf>.

1. To find out the percentage of reads that mapped, check out the run output files. Gmapper prints it towards the bottom:

$ cat my\_sge\_run\_name/my\_sge\_run\_name.o\*

$ cat my\_sge\_run\_name/my\_sge\_run\_name.e\*

#cat will read and print one or more files to standard output.

1. Repeat the mapping (Step 2.2g and h) with the reads files for your unknown sample.

*How do the percentages of mapped reads compare in the two samples?*

*Probably you will find that a larger percentage of reads from CC7 mapped to the reference compared to your unknown sample. Why might this be?*

1. Finally, we will need to filter the alignments to remove weak alignments, and any alignments that may map to multiple reference sequences, possibly indicating that these regions derive from multi-copy regions or repetitive sequences of DNA. To do this, we will use another script from the Meyer lab’s useful arsenal of 2bRAD sequence analysis utilities:

$ ./SamFilter.pl CC7.sam 30 32 CC7\_flt.sam CC7\_counts.tab

#these options will exclude alignments <32 bp and alignments with <30 matching bp.

#counts.tab is a file containing the number of reads mapped per reference sequence. Save this for later!