Ecological Genomics Tutorials: Population & Landscape Genomics 3

September 18, 2023

Learning Objectives for 09/18/23

1. Introduce lab notebooks
2. Visualize sequence alignment files (*.sam)
3. Process the mapping file *sam to binary (*.bam), sort, and remove duplicate reads
4. Calculate mapping statistics to assess quality of the result

1. Lab Notebooks: Take notes on your workflow today so you can remember what you did for the future!

An important step is taking good notes on your workflow so you can remember what you did down the road (your "future self") and share your process with others (reproducible science!). It's also really important once you start making detailed decisions that will affect the analysis outcome of your data, so you can recreate the results and explore the effect of different assumptions/decisions.

We want you to keep such a notebook for each module in the course (kind of like you would for each experiment, or each thesis chapter you will work on). We've provided you with a notebook template based on the markdown (md) language.

Notebook template in markdown (md) format

You should save a copy of this template to your github repo and rename it `PopulationGenomics_Notebook.md` in your main repo folder. You can then open it up and edit this file directly within RStudio, taking notes using either the `Source` interface if you know the markdown language (or want to learn) or you can use RStudio's built-in markdown GUI editor under the `Visual` tab. This works very similar to Word.

Here's a cheatsheet for markdown language if you want to write using `Source` code:

Markdown cheatsheet

When you're done your entries, save and quit RStudio, then commit your changes and push to Github. You can check out your notebook on your repo's site, and Github will automatically render the markdown code into a nice format!

2. Visualize the mapping: By now, you should all have Sequence AlignMent (SAM) files for the inds in your populations!

- Let's take a peek at one of the non-binary (sam) alignment files

`/netfiles/ecogen/PopulationGenomics/fastq/red_spruce/cleanreads/bam/`

- First, try looking at a SAM file using `head` and `tail`. Pick one of your files (just one) to play with below:

```
tail -n 2 YOURFILENAME.sam
```

A SAM file is a tab delimited text file that stores information about the alignment of reads in a FASTQ file to a reference genome or transcriptome. For each read in a FASTQ file, there's a line in the SAM file that includes

- the read, aka. query, name,
- a FLAG (number with information about mapping success and orientation and whether the read is the left or right read),
- the reference sequence name to which the read mapped
- the leftmost position in the reference where the read mapped
- the mapping quality (Phred-scaled)
- a CIGAR string that gives alignment information (how many bases Match (M), where there's an Insertion (I) or Deletion (D))
- an '=', mate position, inferred insert size (columns 7,8,9),
- the query sequence and Phred-scaled quality from the FASTQ file (columns 10 and 11),
- then Lots of good information in TAGS at the end, if the read mapped, including whether it is a unique read (XT:A:U), the number of best hits (X0:i:1), the number of suboptimal hits (X1:i:0).

The left (R1) and right (R2) reads alternate through the file. SAM files usually have a header section with general information where each line starts with the '@' symbol. SAM and BAM files contain the same information; SAM is human readable and BAM is in binary code and therefore has a smaller file size.

Find the official Sequence AlignMent file documentation can be found here or more officially.

- Here's a SAM FLAG decoder by the Broad Institute.
- You can look up TAGs specific to `bwa mem` mapping here

3. Process our mapping files using samtools and sambamba

- We can use the program sambamba for manipulating alignment (sam/bam) files. sambamba. Sambamba is closely related to its progenitor program `samtools` which is written by the same scientist who develop `bwa`, Heng Li. `sambamba` has been re-coded to increase efficiency (speed).
- There are several steps we need to do:
    - convert sam alignment file to (binary) bam format
    - sort the bam file by its read coordinates
    - mark and remove PCR duplicate reads
    - index the sorted, duplicate removed alignment for quick lookup
- Here's a script that we can customize for the above jobs: `/netfiles/ecogen/PopulationGenomics/scripts/process_bams.sh`
- Copy that script into your `~/myscripts` folder and open it in `vim` to edit
- When you're ready, enter a screen session using `tmux` then execute your script `bash process_bams.sh`. If you get an error that the script doesn't exist, then either cd into the `~/myscripts` directory before running your script, or incorporate the path into the file name when you give your bash command.
- Detach from the screen using `<CTRL>` + `b` then `d`. You can always reattach by `tmux attach-session`

4. Calculate mapping stats: How can we get a summary of how well our reads mapped to the reference?

- We can use the program samtools Written by Heng Li, the same person who wrote bwa. It is a powerful tool for manipulating sam/bam files.
- The samtools command `flagstat` gets us some basic info on how well the mapping worked
- We can also estimate depth of coverage (avg. number of reads/site) using the samtools command `depth`
- We'll use both of these commands in loops to assess the mapping stats on each sample in our population.
- We'll also use the `awk` tool to help format the output.
- Here's a script to get us started: `/netfiles/ecogen/PopulationGenomics/scripts/bam_stats.sh`

If there's time while that's running, we can take a look at one of our alignment files (sam or bam) using an integrated viewed in samtools called `tview`. To use it, simply call the program and command, followed by the sam/bam file you want to view and the path to the reference genome. For example:

```
samtools tview /netfiles/ecogen/PopulationGenomics/fastq/red_spruce/
```