Homework Week 6

Friederike Duendar and Luce Skrabanek

ANGSD Course 2020

Exercises (14 pts)

1. Write a script that will: (**3pt**)
    - run `BWA` on one of the samples from the Gierlinski dataset
    - run `STAR` on the same sample
        - Remember those three checks after read alignment:
        - Is it a BAM file?
        - Is it sorted?
        - Is it indexed?
2. Subset the aligned reads to select only those that map to chromosome I. (**1pt**)
3. Compare the output from `BWA` and `STAR`, and summarize any results or differences.
    - Which optional `SAM` fields does `STAR` add and what do they represent? (**1pt**)
    - Which optional `SAM` fields does `BWA` add and what do they represent? (**1pt**)
4. Run `bamqc` on your `BAM` files (Note: this is a tool that's not available in spack, but you can use it via `/softlib/apps/EL7/BamQC/bin/bamqc` after logging on to a compute node). You will need to figure out how to run this on your own (hint: `/softlib/apps/EL7/BamQC/bin/bamqc --help`).
    - Describe 3 differences between the `bamqc` results for both the `BWA` and the `STAR` output files. (**3pt**)
5. Explain the difference between **alignment score** and **mapping quality** in SAM/BAM files. How does the interpretation of the mapping quality field differ between `STAR` and `BWA`? (**2pt**)
6. What is the difference between a **multi-mapping read**, and a **split read**? Find a read that has been split in `STAR`. How did `BWA` handle the mapping of that read? (**2pt**)
7. How can you remove the unmapped reads from the `BWA` output? (hint: go back to the notes where FLAG values were explained) (**1pt**)

Project work (5 pts)

If you need a different program than what we have used in the class, you can use `spack find` to see if the tool is already installed and loadable via spack. If your tool is not there, get in touch with scu@med.cornell.edu to ask them to install it for you.
If you have processes that will take a long time, go back to the notes from the first day and try to make use of `sbatch`.

1. Download at least one FASTQ file that you will be working with for your project. Document the following details: (**2pt**)
    - where did you get it from?
    - what publication is it linked to?
    - who generated the data?
    - how was the NA extracted?
    - what library prep was used?
    - what cell type was used?
    - what was the treatment/experimental condition?
    - what sequencing platform was used?
2. Align the FASTQ file with an appropriate aligner (you may have to build a new index). Document: (**3pt**)
    - parameters (and why you chose them)
    - summary of outcome and basic QC

Compile the `.Rmd` file and send both the `.Rmd` and the `HTML` files to **angsd_wmc@zohomail.com** by Saturday night. If you need support, get in touch with Merv on Thursday, 3-4pm.