Homework Week 4

Friederike Dündar and Luce Skrabanek

ANGSD Course 2020

Questions (6pts)

1. Which base call is more likely to be incorrect – one with a Phred score of `#` or one with a Phred score of `;`? **(1pt)**
2. Explain at least 2 reasons for base calling uncertainties (i.e. what factors could explain lower than expected/desired sequencing scores) and how they can be avoided/alleviated. **(2pts)**
3. What is the baseline uncertainty that Illumina attaches to its base calls? In other words, how likely is it that a base call is wrong even if it got the highest possible Phred score of 41? How many bases can you therefore expect to be wrong in a file with 1 million 50bp-long reads? Does this concern you? (Justify your answer) **(3pts)**

Exercises (with questions) (9pts)

1. Download more `FASTQ` files from the Gierlinski data set so that you have all the technical replicates for 3 WT and 3 SNF2 samples (= 6x7 `FASTQ` files). Place each set of 7 technical replicates into one sensibly named folder respectively. (**1pt**)
2. Write a for-loop that will run `FastQC` on **all** (6x7) of the `FASTQ` files that you previously downloaded from the Gierlinski dataset. Select one sample for which you write an additional for-loop that will:
    - run `TrimGalore`
    - run `FastQC` on the trimmed datasets. **(2pts)**
3. Describe one detail of the QC results that changes after `TrimGalore` and one result that stays the same and explain why. **(2pts)**
4. Combine the initial `FastQC` results for all 6x7 FASTQ files into one document using `MultiQC`. You can load the tool using `spack load -r py-multiqc`. Export one image of either of the results where the SNF2 samples are highlighted in a different color than the WT samples and add it to this report. (**2pts**)
5. Based on the QC, would you be justified in combining any of the `FASTQ` files given that they are technical replicates? **(1pt)**

6. Even if the answer to the previous question is “no”, what command(s) would you use to combine the several `FASTQ` files into one? **(1pt)**
7. *Bonus point*: If you had to determine the version of the Sanger quality score encoding used in a given `FASTQ` file without the help of `FastQC`, what would you do?

Project work (4pts)

1. Expand your project ideas. Come up with (at least) one **specific hypothesis** that you want to test.
2. Specify the **data** you will need.
    - Locate potential datasets and describe them (when/where were they generated, what sequencing platform was used, etc).
    - Think about possible biases or technical problems that you might run into if you were to use these data. (remember the lecture about experimental design!)