Applied Genomics 2018

Homework 01 - Linux and SLURM

Required files are located in /scratch/courses/AppliedGenomics2018/Homework01

- cdg21c1_R1.fastq - reads from *Saccharomyces Cerevisiae* resequencing
- Scer3.fa - reference genome sequence in FastA format
- Scer3.gff - Gene annotation file.

Submit your answers to the following questions in a document. You **MUST** provide the **CODE**, **ANSWERS**, and the **EXPLANATIONS** for each question. You are allowed to submit Word and PDF documents.

Linux - answers to the following questions have to be answered using bash commands. (40pts)

1. Create a directory called **Homework01** in your scratch directory and copy the three files there.
2. How many sequences are provided in the Scer3.fa file and what are their names?
3. How many genes are annotated in the Scer3.gff file? *Hint, the third column contains information about the type of feature that is being described in the row*
4. How many different features are provided in the Scer3.gff file?

SLURM (30pts)

Create and execute an sbatch script that will run FASTQC on cdg21c1_R1.fastq. Submit this as a separate file with your homework.

FASTQC (30pts)

Inspect the results of FASTQC and answer the following questions:

1. What is the GC content (in %)?
2. What's the number of sequences flagged as poor quality?
3. What's the total number of sequences?
4. What percentage of the sequences will remain after deduplication?
5. Are any sequences overrepresented?
6. After which base pair does the mean quality score drop below 20?