**Chameleon Issue and BWA-mem Metrics Quality**

**Last week's task:**

1. **Collecting BWA-mem system metrics**

**Next week's task we will collect system metrics of:**

1. **Samtool**
2. **Picard SortSam**
3. **Picard MergeSamFiles**
4. **Picard MarkDuplicates**
5. Overview

In this report, we have compiled BWA-mem metrics for 10 files. The collection of metrics starts with 10 files for each tool because there is an issue–until an unknown date–in Chameleon which causes us to not be able to access CHI@TACC, a site with high availability, so the resources used are quite limited. The quality of the data obtained is not very good because there are several missing values in Slurm accounting, especially in the memory usage field for jobs that are executed in less than 30 seconds.

1. Chameleon Issue

Currently, there is an issue in Chameleon. This bug causes us to not be able to activate the instance on the CHI@TACC site. When an instance is attempted to activate, it will continue to be in spawning state for more than 30 minutes–usually less than 7 minutes–and end with an error. Apparently, the error occurs only on the CHI@TACC site, because the instance can be activated on the CHI@UC site. However, the availability of CHI@UC is lower than that of CHI@TACC. As an illustration, it's pretty easy to get 10 instances on the CHI@TACC site. Meanwhile, to get more than 3 instances at the same time on CHI@TACC is very difficult.

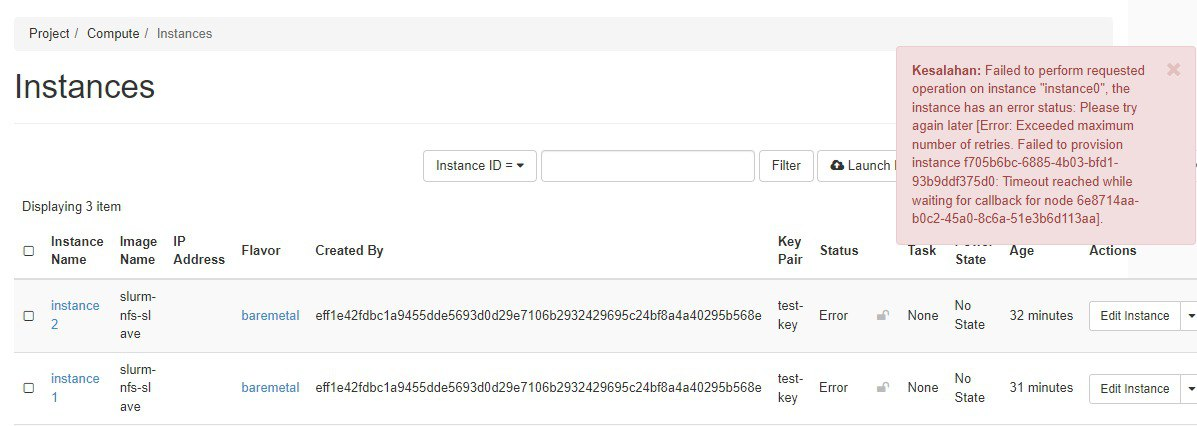


Figure 1. Error when activating CHI@TACC instances

1. Collect BWA-mem

Due to the limitation of instances in an uncertain time, the BWA-mem collection process is carried out with 10 files first. These files were taken from the thyroid papillary carcinoma study (GSE63511). The input files are processed a little first so that we can immediately work on other tools and enter the analysis stage. So while we're doing the tools analysis, we can execute the tools for some of the other input files.

| File (fastq) | Size (MB) |
| --- | --- |
| SRR2059426 | 845 |
| SRR2059427 | 1006 |
| SRR2059428 | 1311 |
| SRR2059429 | 605 |
| SRR2059430 | 1155 |
| SRR2059431 | 687 |
| SRR2059432 | 1083 |
| SRR2059433 | 1435 |
| SRR2059434 | 1351 |
| SRR1658393 | 653 |

Table 1. Input size of BWA-mem

1. Data quality

After successfully collecting BWA-mem metrics for 10 input files, we get quite a lot of missing values. These data are collected using the Slurm accounting feature. Of the 937 jobs with the status of "completed", the memory usage of 147 jobs is not recorded which is indicated with a value of 0. The missing data is quite concentrated on the execution time of under 30 seconds (as shown in figure 2). More specifically, all data with an elapsed time value below 30 has a memory usage value of 0 (missing value).

To ensure that these data are missing values, BWA-mem is run with the parameter values associated with the missing values and monitored using the top command. For example, BWA-mem with threads of 16, memory allocation is 16GB, and CPU allocation is 6 using the file SRR1658393.fastq is run and found that the memory usage is 8GB on top, but 0MB on Slurm accounting. And the numbers are consistent when the execution is repeated. When the number of threads is reduced–which causes the execution process to take more than 30 seconds–the memory usage can be captured by Slurm accounting, which is 6GB.

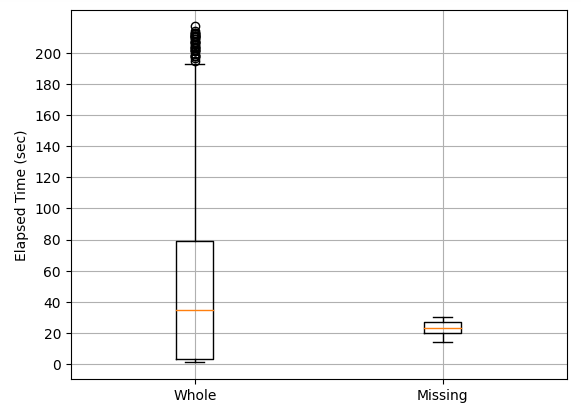


Figure 2. Box plot of missing values. The missing memory values are quite focused on the elapsed time of less than 30 seconds compared to the entire completed job data which varies up to 200 seconds.