J. Jedediah Smith

BIFX 552

Final Project Essay

12/4/202

**Mothur has the ability to run one or a few commands from a terminal window. What would be the advantages and disadvantages of converting `stability.batch` into a Snakemake file with one rule for each step of the SOP?**

Some advantages of turning the batch file into a Snakemake file would be the use of wildcards, compartmentalized chunks, dry runs, and controlled processor use on a server. All of these things lend themselves well to pipeline creation. Wildcards make it easy to process many files at once and change how the output files are named or numbered. The compartmentalized chunks are great for organization. It makes it easier to tell which functions happen where, outsource specific functions to other programs, debug specific errors, change things or move them around, and run specific chunks by themselves. Dry runs are nice because you can test the code for errors without actually changing anything or producing output and taking up processor space. And the control of processors is nice so that when you are working on a big server with hundreds of people, the entire world doesn’t stop just because you are running this singular process. Overall, it seems to lend itself well to crunching a massive amount of data and would be good for when you work at an institution with a big server and hundreds of scientists running different things for their work.

While that certainly sounds like all sunshine and rainbows, there are some pretty significant disadvantages too, mainly surrounding the complexity of setting it up Snakemake and writing Snakefiles. In our class, it was quite a struggle to get Snakemake up and running. Even once we did, the process of creating each of those compartmentalized chunks is rather time-consuming. There is a large amount of redundant information linking the output of one chunk to the input of another. Each redundancy is a chance to make errors. The batch file we use runs ~31 commands. Assuming each of those needs to be its own chunk, that would take way longer to code and debug, potentially resulting in a file probably at least 300 lines long, possibly more. While this might be worth it for a robust pipeline or product that is going to be reused and shared by thousands of scientists on various servers, it doesn’t make sense for a quick analysis that may only be run once or twice for a paper or project. Overall, it seems like the batch file approach is more efficient, though perhaps less thorough, and as a result more suited for smaller projects.