

Part 1: Write a script using slurm metadata

Author: Zach Maas

The goal of this exercise is to take the skills that you practiced yesterday and use them to write a script formatted so that it can run using the slurm workload manager. Slurm is software that allows for large compute jobs to be delegated across a cluster of computers, scheduling things so that resources are used optimally. This makes it so that many people can use the same compute system at the same time. Important: If you write a script for slurm and run it on the login node instead of submitting it to slurm, you will slow down the system for everyone else, potentially rendering it unusable. Double check your commands.

Things you need to do:

1. Make a folder on the short read system called workshop-day3
2. Make folders there for data, scripts, and results
3. Write a slurm script to download the fastq file located here:
`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00096/sequence_read/SRR062641.filt.fastq.gz`
4. Submit your slurm script using the sbatch command and check to see that it does what you expect
5. Use the `md5sum` command to make sure that the checksum for the file matches `412d95d53ac858dbc899494f9db42ed2`. A checksum is a way of verifying that a file is the same as the original we wanted to download. Why would we want to use this?
6. Back up your script and results to your home directory

Some useful things to know:

- Your sbatch script needs additional directives added to the top of the file so that the slurm system knows how many compute resources you need:

```
#!/bin/bash
#SBATCH --output=<a path to a directory>/%x_%j.out
#SBATCH --error=<a path to a directory>/%x_%j.err
#SBATCH --mail-user=<an email address to get notifications at>
#SBATCH -p <the name of the slurm partition we're on>
#SBATCH -N <the number of nodes to use, should be 1>
#SBATCH -c <the number of cores to use>
#SBATCH --mem=<the amount of memory to use, formatted 256mb, 8gb, etc>
<the rest of your script>
```

- `wget` is a command that will download a file from a remote location. The `-c` flag stands for (c)ontinue and means that your download can resume if it is interrupted for some reason.