Short Read Analysis Best Practices

By Mary Allen and Robin Dowell

Goal of any analysis:

Discover some cool results WHILE maintaining integrity and reproducibility.

1. Download your data -- from a sequencing facility or a repository
    a. Lock its permissions so it can never be edited!!!
    b. Keep the raw data somewhere, if you generated it then it has to be backed up!!!
    c. NEVER TOUCH RAW DATA!!!
    d. Note where you got it from (facility, machine, etc)
        i. including the pub if not your data (citations!)
    e. Make a meta data table (how was it generated? Cells? Perturbation? etc.)
2. Run your analysis
    a. Set up your storage system optimally
        i. /scratch/ vs /Users/ -- fast vs backed up
        ii. Make a directory on /scratch for each of your projects (/scratch/Shares/labname/ or scratch/Users/username/)
        iii. Make an input and output directory in your project directory
            1. Rsync your raw data to your input directory on scratch
                a. scratch is not backed up!!!!!
            2. Keep a README or NOTES file with the path of the raw data and when you copied it over. If you do any massaging of your data -- record it.
        iv. Make a scripts directory
            1. Use version control -- ex: github
                a. Github walk through
            2. Be sure your scripts are backed up (this is free with github)
    b. All software you run should be in a script (not on command line!)
        i. Make a README file that tells everything you would put in your lab notebook, track as you go
            1. Where does the raw data live?
            2. Which scripts did you run on it (and why)
            3. What files did you make and where are they
            4. What versions of software, genomes, annotations, etc where used?
    c. Keep the living room clean!
        i. In pipelines, there will often be intermediate files (output of program A used as input for program B). Actively manage these intermediates.
            1. When you get intermediate files you want to backup, rsync them to somewhere backed up -- otherwise, delete as soon as you don't need them.

2. Delete stuff on /scratch periodically (Data on scratch costs more and clogs up the system).

d. Always, Always, sanity check your results
    i. QUALITY, QUANTITY (NUMBER OF READS), VISUALIZE
        1. Programs do not always fail gracefully!!! So did it even work?
        2. Did all the data get used?
        3. Is the output as expected? Did you get interpretable results?
    ii. Don't believe people when they say their program does X (check!)

3. Publish ***all*** your data (the ultimate goal!?!?)
    a. Upload the raw data, meta-data and the final processed files to NIH GEO
    b. You must report all manipulations of data (manual or through analysis tools)
    c. All versions of all programs used must be noted and the paper associated with the program should be CITED in the methods section. Example: "We used Tfit (Azofeifa 2017) v 1.1 to identify eRNAs."
        i. For reproducibility, you MUST note all the flags/options you used if not standard (i.e. defaults are assumed)
            1. You can do that in your scripts in github and provide the "source" code for version information (must provide github link in methods).
            2. Or you can use Jupyter notebooks to document analysis, plotting, manipulations and versions. Provide the Jupyter notebook in the methods.
            3. Or you can simply note all the flags in the methods section -- works well if there are few of them, but gets a bit unwieldy if you used a lot of software and unique options.