Automating A Workflow For High Throughput Genomic Analysis Of Wildlife Pathogens In Wyoming

Yasin M. Ahmed, Silba S. Dowell, Chris MacGlover, Tera N.Swaby, Liudmila S. Mainzer

Abstract

Automated Workflow Management system makes it possible to orchestrate multistep, complex, time-consuming processes in a well-organized, parallelized, reproducible fashion. In our current study, we developed an automated genome analysis workflow to identify bacterial isolates from infected wildlife samples using Nextflow platform [1]. For that purpose, individual bioinformatics programs were channeled together in a single pipeline deployed on the Teton HPC cluster at the University of Wyoming [3-9].

Background

- Whole genome sequencing technologies are becoming robust and inexpensive. Yet the cost of computational analysis and the human effort in deploying and maintaining the code is still very significant.
- Our objective was to develop a data analysis pipeline that can process very large datasets in a rapid, efficient, standardized manner using the high performance Nextflow platform.
- This enables the discovery of the microbial groups linked to wildlife diseases researched at the Wyoming State Vet Lab [2].

Methods

- Nextflow is an automated workflow management platform that allows the incorporation of multiple programs written in different computer languages. The advantages include data-level parallelism (running several processes at once) and ease of code maintenance [1].
- We defined the parameters used by the processes in a .config file, whereas the scripts themselves were written as .nf files.
- We used the DSL2 syntax for programming the input channels and emit methods to organize several programs into a streamlined workflow.
- Workflow was deployed across several 16-core nodes, each with 128 GB RAM. Total runtime was 1hr and 42min 56s.

Figure 1. Samples were collected from diseased wildlife: Bighorn sheep, wild turkey. Bacteria were isolated and hybrid sequencing was performed (Oxford nanopore and Illumina) to serve as input to the standardized automated bioinformatics workflow.

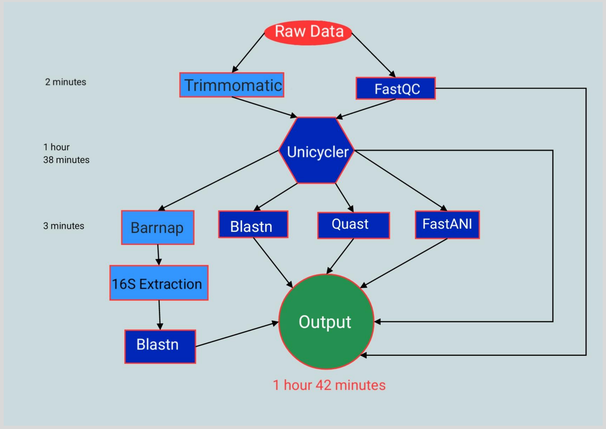


Figure 2. The workflow included standard bioinformatics tools used for whole genome analyses. Intermediary steps = light blue. Final output steps = dark blue.

```
process Unicycler_Hybrid_Assembly {
        label 'unicycler'
        publishDir "${params.output}/unicycler_results"
        input:
                path Input_Short_Read_1
                path Input_Short_Read_2
                path Input_Long_Read
                val Threads
                path Output

        output:
                path("output/assembly.fasta"), emit: contigs

        script:
                """

                module load miniconda3
                conda activate /pfs/tc1/project/arcc-students/bio2

                unicycler -1 ${Input_Short_Read_1} -2 ${Input_Short_Read_2} -l ${Input_Long_Read} -t ${Threads} -o ${Output}

                """
}
```

Figure 4. Example syntax layout of the Unicycler process in Nextflow.

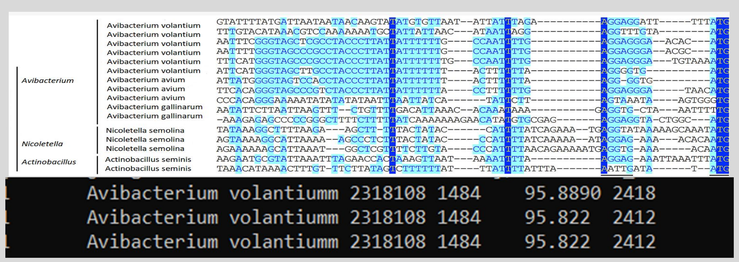
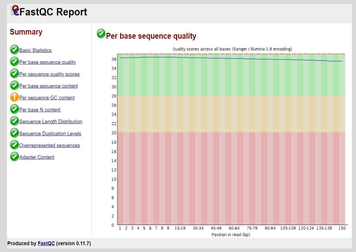


Figure 5. Identification of the pathogens by Blast search against the NCBI database of reference genome [11].

a)



b)

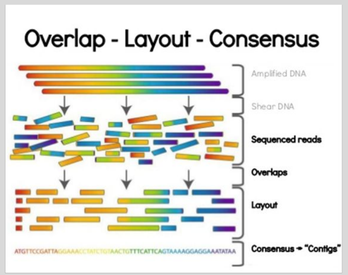


Figure 3. a) Quality of input reads assessed by FastQC. b) De-novo hybrid assembly of whole bacterial genome from short and long reads by Unicycler [10].

```
N E X T F L O W  ~  version 21.10.6
Launching `full.nf` [fervent_bartik] - revision: 7e78e06e20
executor >  slurm (11)
[90/ecee78] process > fastqcRun (1)              [100%] 2 of 2 ✔
[cc/ab8856] process > TrimSequences_PE           [100%] 1 of 1 ✔
[fe/a778e8] process > TrimSequences_SE           [100%] 1 of 1 ✔
[a2/d7e5a6] process > Unicycler_Hybrid_Assembly  [100%] 1 of 1 ✔
[f2/22b3be] process > quastRun                   [100%] 1 of 1 ✔
[91/c5d021] process > blast1                     [100%] 1 of 1 ✔
[19/96acdc] process > fastani                    [100%] 1 of 1 ✔
[7b/1a7ec7] process > barrnap                    [100%] 1 of 1 ✔
[ed/97e3bc] process > foo                        [100%] 1 of 1 ✔
[5e/ae3232] process > blast2                     [100%] 1 of 1 ✔
Completed at: 30-Jul-2022 20:44:05
Duration    : 1h 41m 56s
CPU hours   : 27.7
Succeeded   : 11
```

Figure 6. Parallel execution of multiple programs run by Nextflow.

Discussion

Our nextflow automation is advantageous:

- A single Nextflow command deploys 9 bioinformatics programs along with their numerous dependencies.
- User does not have to change every single script every time to run different input data files.
- User can run this analysis without having significant programming knowledge.
- Several programs run in parallel, saving time.
- The process is highly reproducible.
- Passing of intermediary data between programs is automatic. This is human-error proof.
- The outputs are delivered in a very well-organized way, easy to find and interpret.
- The pipeline can be used currently for any bacterial pathogen identification.

Future Directions

- Simple modifications of the current pipeline will enable the program to perform additional array of analyses. It is maintainable.
- We will apply the pipeline for large amount of dataset of wildlife and livestock pathogens in Wyoming.
- The workflow is easy to adapt to other pathogens, such as viruses (e.g. monkey pox, COVID, etc.)

Conclusion

Our Nextflow based pipeline is efficiently capable of pangenomic analysis of bighorn sheep pathogens. In the future, it will be used to perform other types of analyses such as metagenomics to understand the affects of microbial communities in livestock and wildlife diseases.

Acknowledgments

- Many thanks to the Advanced Research Computing Center (ARCC) for supporting this work and providing the computing resources.
- We are grateful to the UW School of Computing and Prof. Bryan Shader for supporting the initial POC of writing this workflow in Bash.
- Special thanks to Simon Alexander for assisting with the initial setup and training on the Teton cluster at ARCC.

References

[1] P. Di Tommaso, et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017) doi:10.1038/nbt.3820
[2] B. Harris, J. Hicks, M. Prarat, S. Sanchez, and B. Crossley, "Next-generation sequencing capacity and capabilities within the National Animal Health Laboratory Network," *Journal of Veterinary Diagnostic Investigation*, vol. 33, no. 2, pp. 248–252, 2020.
[3] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[4] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
[5] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017.
[6] Barrnap (RRID:SCR_015995)
[7] Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/
[8] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086 First published online: February 19, 2013
[9] Jain, C., Rodriguez-R, L.M., Phillippy, A.M. *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018). https://doi.org/10.1038/s41467-018-07641-9.
[10] http://www.chromnet.net
[11] https://www.bitesizebio.com