Agriculture and Agri-Food Canada | Agriculture et Agroalimentaire Canada

Assessing and Improving In-house Genome and Transcriptome Assembly Solutions

Jeffrey Cullis[1], Iyad Kandalaft, Zaky Adam, Christopher T. Lewis[2]

Eastern Cereal Oilseed Research Centre (ECORC), Agriculture and Agri-Food Canada (AAFC), Ottawa, Ontario, Canada.

1. Jeff.Cullis@agr.gc.ca [Presenting Author], 2. Christopher.Lewis@agr.gc.ca

Introduction: The Assembly Pipeline

At AAFC, we have developed systems and tools to perform computationally intensive analysis such as genome and transcriptome assembly in an automated, reproducible fashion for researchers working on many target organisms within the organization. The bioinformatics landscape can change quickly, and incorporation of new tools and best practices is necessary in order to be sure of providing outputs of the highest possible quality.

Key Objective: Automated Releases

A central design goal for this pipeline was to automate the creation of a consistent release structure for final output files. This includes use of a consistent naming scheme, generation of metadata, and automated internal publication of results for local researchers (Figure 2).

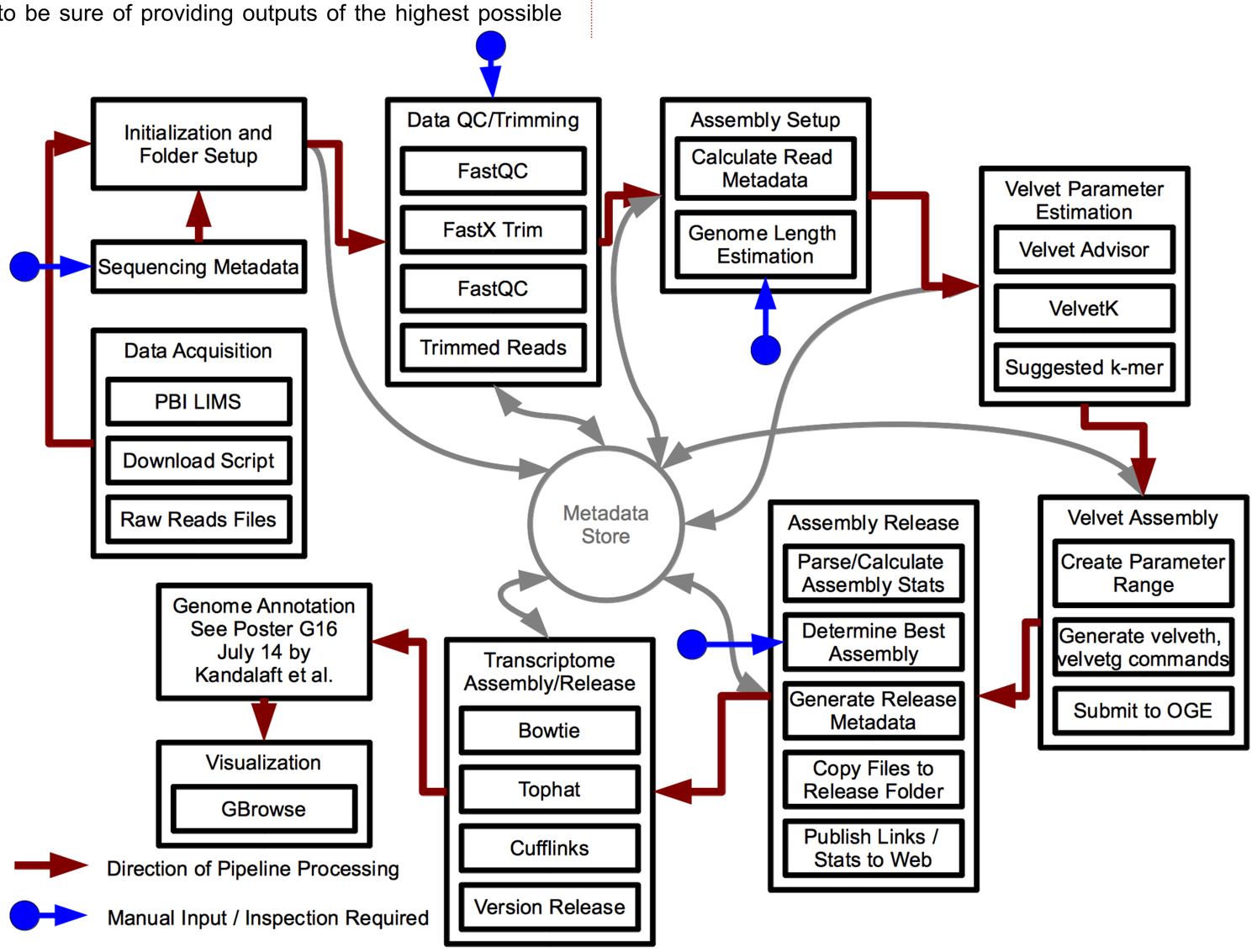

Figure 1. Overview of the current genome and transcriptome assembly pipeline used at AAFC.

Features of our current pipeline (Figure 1) include optimization for our specific OGE cluster computing environment, and full automation of steps from raw data acquisition to assembly quality assessment and comparison, to release generation. The pipeline has been used to assemble multiple genomes and transcriptomes for over 50 different organisms, primarily from fungal, nematode, and bacterial samples.

The included metadata (Figure 3) provides information about sequencing parameters used, quality control and assembly quality metrics, and a detailed description of the processing steps applied to the data. Focus on this aspect of the pipeline has paid benefits in the ability to quickly discern provenance for the many files that different researchers are working with, and to ensure reproducibility of results.

Kingdom	Species	Strain	Release	Type	PE	MP-3kb	MP-8kb	K-mer	Link	Genome Browser	Total Length	Estimated Genome Length	Number of Contigs	Min Contig	Median Contig	Max Contig	N50	Reads Used
Fungi	Tilletia caries	DAOM 238032	R01	Genome	yes	no	no	73	Release	View	31,591,973	20,000,000	8,613	145	430	101,122	16,144	27,037,431/30,406,426
Fungi	Tilletia caries	DAOM 238032	R03	Genome	yes	yes	yes	45	Release		35,261,422	20,000,000	16,301	89	142	2,812,152	855,893	362,018,494/688,480,430
Fungi	Tilletia controversa	DAOM 236426	R01	Genome	yes	no	no	65	Release	View	30,158,085	20,000,000	9,085	129	407	115,074	14,232	16980504/19405294
Fungi	Tilletia controversa	DAOM 236426	R02	Genome	yes	yes	yes	43	Release		33,264,752	20,000,000	18,761	85	142	2,218,965	1,226,554	392,318,737/576,418,916

Figure 2. Links to release folders, with assembly stats, as shown on the internal Wiki.

Future Work: Galaxy Integration

The use of a release strategy has yielded many benefits. However, we have also identified a number of areas where further improvements could be made to the pipeline.

1. Improve the modularity of components for gathering and storing metadata.
2. Improve the modularity of tools, such that new tools for genome assembly (SPAdes, for instance) and new tools for read correction can be more easily integrated.
3. Allow for biologists to run customized versions of the pipeline on their own data.
4. Improve the interface for setup and execution of jobs on the compute cluster.

Our experiences using Galaxy workflows to date have shown that it provides all of the functionality necessary to meet the four criteria listed above. Integration with Galaxy would greatly simplify many tasks now handled by the pipeline, including gathering of metadata, modifications to tools, end-user support, and parallelization. However, in order to make use of Galaxy, we will need to take the time to create wrappers, as well as tool dependencies for those tools not already present within Galaxy.

```
---
- release:
    date: 10/02/2013
    species: Tilletia caries
    strain: DAOM 238032
    version: R01V1
- genome_assembly:
    N50: '16144'
    estimated_genome_length: '20000000'
    ...
    total_length: '31591973'
- samples:
  - read_data:
      R1:
        raw:
          file_path: ...
          num_reads: '15203213'
          read_length: '101'
      R2:
      ...
    sequencing_metadata:
      Barcode: ATTCCT
      ...
      Total_Numreads_in_Lane: '210850661'
    ...
- pipeline:
  - command: /opt/bio/FastQC/fastqc -o ...
    description: FastQC
    qsub_cmd: /opt/gridengine/bin/lx26-amd64/qsub  ...
    run_dir: ... /AssemblyPipeline
    version: FastQC v0.10.1
  - command: /opt/bio/FastQC/fastqc -o ...
    description: FastQC
    qsub_cmd: /opt/gridengine/bin/lx26-amd64/qsub ...
    run_dir: ... /AssemblyPipeline
    version: FastQC v0.10.1
  - command: /opt/bio/velvet/velveth_127 …
    ...
```

Figure 3. Sample release YAML metadata file (abbreviated)

References

1.Patel, R.K. et al. “NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data.” PLoS ONE (2012), Vol. 7, No. 2.

2.Zerbino Daniel, et al. “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.” Genome Research (2008): 18:821-829.

3.Kim, Daehwan et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.” Genome Biology (2013), 14:R36

4.Trapnell, Cole et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.” Nature Biotechnology (2010) doi: 10.1038/nbt.1621

5.Goecks, Jeremy, et al. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol 11.8 (2010): R86.

Acknowledgements

Funding for this work was provided by the Canadian Safety and Security Program (CSSP).

Canada