Developing robust, high-volume reproducible sequencing bioinformatics pipelines for cancer research

Morgan Taschuk[1], Michael Laszloffy[1], Peter Ruzanov[1], Brian O'Connor[1], Denis Yuen[1], Timothy Beck[1], Francis Ouellette[1,2]

1 Ontario Institute for Cancer Research, 661 University Ave, Suite 510, Toronto, Ontario, M5G 0A3, Canada
2 Department of Cell and Systems Biology, University of Toronto, 27 King's College Circle, Toronto, Ontario, M5S 3G4, Canada

Introduction

In projects like the International Cancer Genome Consortium, thousands of matched tumour/normal samples are sequenced by centres distributed throughout the world. Sequence analysis constantly evolves as new sequencing technologies, algorithms, and software are introduced. The volume and scope of analysis requires rigorous documentation of analysis methods and the need to share methods and pipelines among multiple centres. **Recent concerns surrounding the reliability of preclinical findings emphasize the need for agility, robustness and reproducibility in large-scale sequence analysis.**

Technology

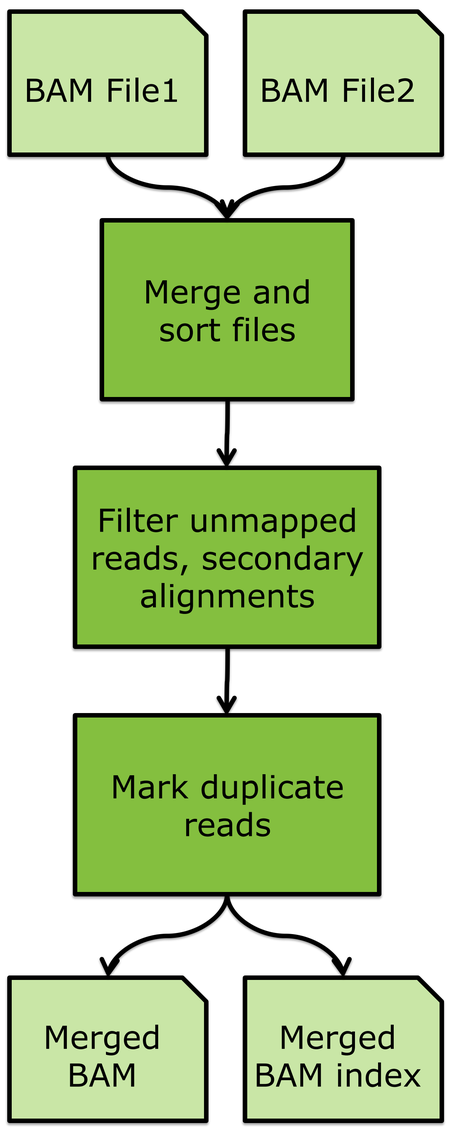
We use the SeqWare infrastructure (http://seqware.io) to create pipelines to analyze different types of cancer data, preserve metadata, and evaluate and incorporate new algorithms and technologies to remain on the forefront of cancer research. However, technology alone is insufficient to ensure that pipelines are agile, robust and reproducible.

Pipelines

A pipeline is an ordered series of analysis steps that are chained together so that the output of one step becomes the input for the next.

In order to make it worthwhile to turn a set of commands into a pipeline, we have a number of assumptions:

1. **Volume** – the pipeline will regularly be required to process a lot of samples
2. **Variety** – the input data will vary in size, type, or analysis techniques required
3. **Well-defined** – the tools used for a particular task are standard (does not include parameters)



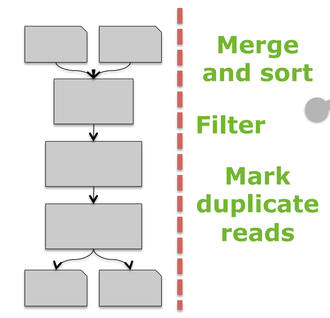
As an example, we provide one of our workhorse pipelines: BAM Merge Filter Collapse. This pipeline accepts one or more BAM inputs, merges them and performs pre-processing in order to facilitate downstream analysis.

Best Practices

We have developed a set of best practices for developing high-throughput sequence analysis pipelines, focusing on three target characteristics: agility, reproducibility and robustness.

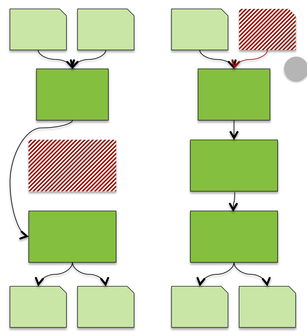
Agility

Quickly adapt to changing demands



Separate command from hierarchy
- Separate the analysis commands from the ordering of the jobs. This allows you to change the commands without revising the job hierarchy and vice versa.

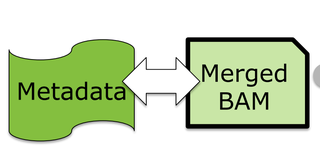
Use standard input and output formats



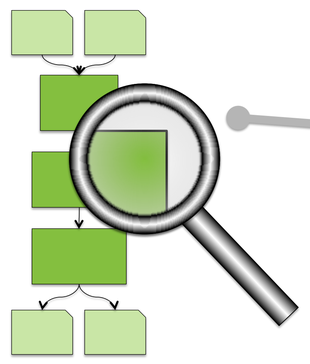
Set reasonable defaults but allow flexibility
- Functions well using the default settings
- Allows a knowledgeable user to customize aspects of the pipeline for their own use.

Reproducibility

Replicate the results in-house or in other locations



Record analysis metadata/provenance
- Use standard formats if available (e.g. MINSEQE).
- Record software tools, versions and parameters *for specific runs*
- Associate metadata with the results files



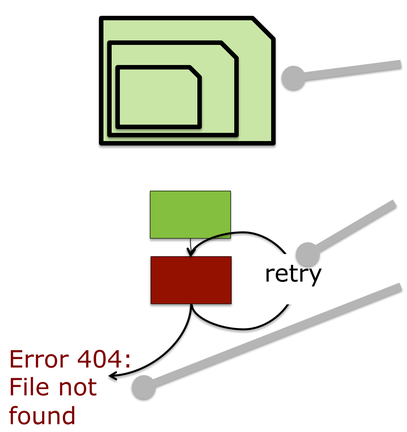
Package with test data
- Distribute the pipeline with data that has known results

Test, test, test

Time spent testing before release saves support time later
- Automated regression tests
- Test on multiple data sizes and types

Robustness

Prevent failures from affecting production.



Design for scale
- Keep memory requirements in mind as data scale up

Robust
- Produces reasonable results on valid input data

Fault-tolerant
- Handles transient errors without failure.

Exit statuses and error messages
- Fails loudly with good error messaging and non-zero exit status on non-transient errors

Availability

Our analysis pipelines as well as tools that support our best practices are available on GitHub (https://github.com/pipedev). These advances allow us to analyze high volumes of data in a reliable and reproducible fashion while keeping up with the latest developments in the field of cancer sequence analysis

Ontario

Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario.