A registry of Australian Genomics bioinformatics pipelines incorporating a structured representation and interactive visualisation

Mailie Gall[1,2,3], Bernard Pope[2,3], John Pearson[1,4] & Natalie Thorne[1,3,5]

1. Australian Genomics Health Alliance, 2. Melbourne Bioinformatics, 3. The University of Melbourne,
4. QIMR Berghofer, 5. Melbourne Genomics Health Alliance

Introduction

The contents, structure and configuration of bioinformatics pipelines are critical to understanding their behaviour, but have traditionally been buried within their source code implementations. This lack of transparency reduces our confidence in pipeline correctness, severely limits our ability to compare and contrast different systems, and is a lost opportunity for providing key provenance information for clinical tests. There are currently no standards or conventions currently adopted by the bioinformatics community for documenting pipeline metadata, and few integrated tools for visualising this information. We aim to address these limitations by building a registry of clinical bioinformatics pipelines used within the Australian Genomics Health Alliance, that takes advantage of emerging data ontologies and standards.

Purpose

- Develop a standard, platform-agnostic, detailed structured document to describe clinical bioinformatics pipelines.
- Develop an interactive, dynamic tool to visualise and explore pipeline descriptions.
- Describe pipelines used by Australian Genomics flagships and compile into an online registry.

Pipeline Registry

- Pipeline descriptions were gathered through inspection of pipeline and tool manuals, command logs and interviews with bioinformaticians.
- 14 pipelines (WES, WGS, panel and trios) from 8 flagships have been fully (7) or partially (7) described (Fig. 1a & Fig.1b).
- The pipeline descriptions are available in an online registry: https://agha.qimrberghofer.edu.au/wiki/PipelineRegistry.

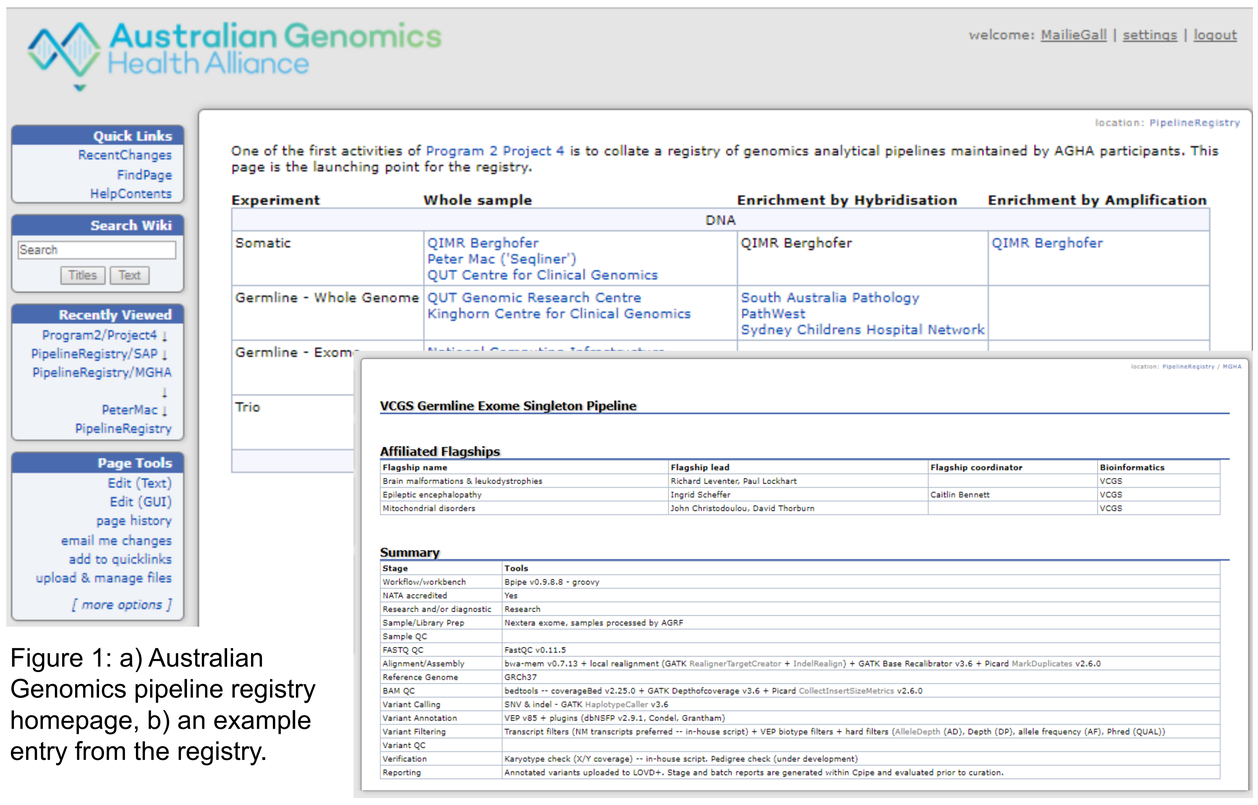

Figure 1: a) Australian Genomics pipeline registry homepage, b) an example entry from the registry.

sub-workflow inputs/outputs

```
cwlVersion: v1.0
class: Workflow
label: read_alignment
doc: 'mapping of paired-end reads to the reference assembly',
     'specs: http://bio-bwa.sourceforge.net/bwa.shtml',
     'identifier: https://identifiers.org/rrid/RRID:SCR_010910'
inputs:
    forward_reads:
        type: File
        doc: 'Forward read input.'
        format: edam:format_1930
    reverse_reads:
        type: File
        format: edam:format_1930
        doc: 'Reverse read input.'
    reference_assembly:
        type: File
        format: edam:data_2340
        doc: 'hg19 human reference genome.'
outputs:
    aligned_merged_bam:
        type: File
        outputSource:
        merge_alignments/aligned_merged_bam
```

pipeline stages

```
steps:
  align_to_ref:
    run: ../tools/align.cwl
    in:
      forward_reads: [forward_reads]
      reverse_reads: [reverse_reads]
      reference: [reference_assembly]
    out:
      [ref_aligned_bam]
  merge_alignments:
    run: ../tools/merge_bam.cwl
    in:
 ref_aligned_bam: align_to_ref/ref_aligned_bam
    out:
      [aligned_merged_bam]
```

Figure 2: An example CWL workflow file.

A standard to describe pipelines using CWL

- The Common Workflow Language (CWL) and Workflow Description Language (WDL) were evaluated for their suitability as descriptive formats/languages to describe pipelines.

- CWL was adopted by this project for the following reasons:
  → declarative and flexible specification with links to ontologies and registries (EDAM ontology, bio.tools registry).
  → follows collaborative open standards with large community involvement in its development.

- A structured document (see Fig. 2) based on CWL specifications includes:
  → workflow, sub-workflow and tool class objects.
  → EDAM ontology references providing links to common terms, definitions and additional metadata for file standards.
  → bio.tools registry ID to enhance tool descriptions.
  → a draft schema describing mandatory/optional data constraints in development.

Future directions

- Publication of the registry and development of an online interface.
- Support Alliance members to contribute their pipeline to the registry.
- Submit CWL Explorer to the community for evaluation/further development.
- Use pipeline description standard when sharing genomic data across the Alliance.

CWL Explorer

We have developed an interactive, visualisation tool for pipeline descriptions.

Specifications:
- Web-based tool based on Cytoscape.js for interactive graph visualisation.
- Input is a bioinformatics pipeline defined in CWL.
- Nodes represent objects (reference files, data objects or stages).
- Edges represent data flow.
- Supports workflows and nested sub-workflows.

Features
- Collapsible nodes to reduce complexity, and guide the users view.
- 'Tooltips' and panels to view metadata, URLs to external data standards.
- Sub-workflows to represent nested and/or repeated workflow stages.
- Potential for workflow editing in future versions of the tool.

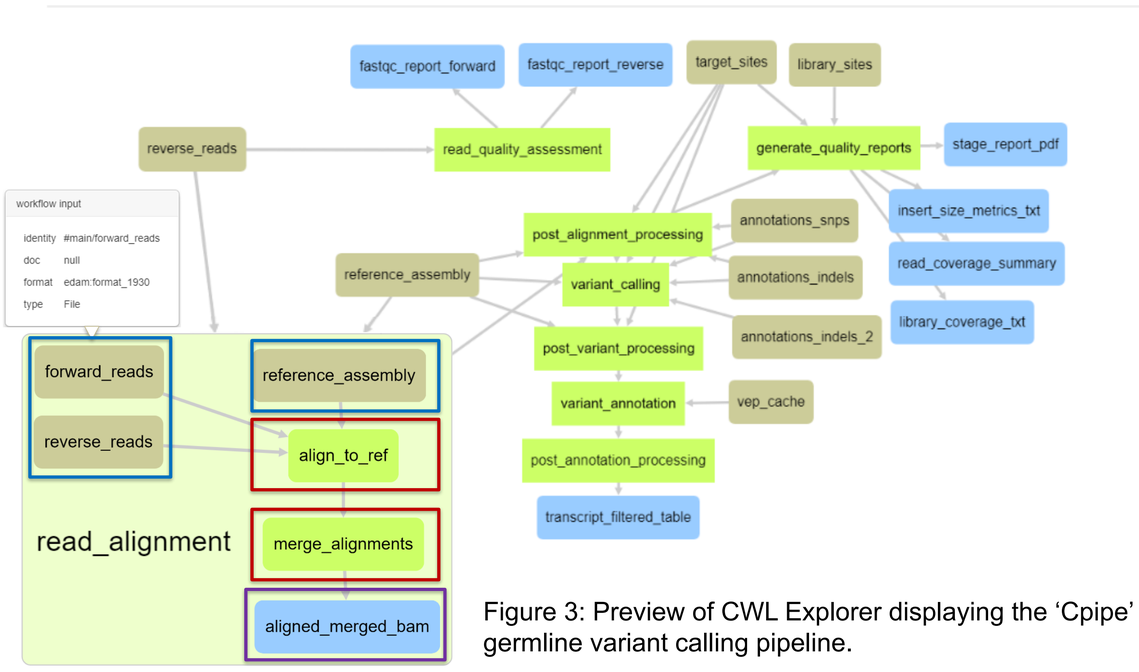

Figure 3: Preview of CWL Explorer displaying the 'Cpipe' germline variant calling pipeline.

Contacts

Project lead: Natalie Thorne (natalie.thorne@melbournegenomics.org.au).
For more project information: Mailie Gall (mailie.gall@unimelb.edu.au).

Acknowledgments

Australian Genomics is funded by NHMRC grant 1113531. P2.p4 working group contributed information and provided feedback on the design of the registry. Michael Crusoe provided advice and expertise on the CWL spec. Information on the pipelines was contributed by bioinformaticians from many organisations (see registry for personnel list).