

CWL

Translating workflows into Nextflow with Janis

Common Workflow Language (CWL)

“Common Workflow Language (CWL) is an open standard for describing how to run command line tools and connect them to create workflows.”

Designed as a workflow definition standard:

- Describes a tool / workflow - does not run it
- Execution handled by any engine which adopts the standard: cwltool, Arvados, Cromwell, Toil
- Ethos is that tools / workflows can be reused in multiple pipelines

CWL focuses on reproducibility, portability, and interoperability

Chose to support CWL in janis-translate as CWL is widespread.

Description



COMMON
WORKFLOW
LANGUAGE

Execution



Features

CWL tends to be verbose

- Restricted & specific grammar
- Must be easily read into memory so execution engines can run CWL descriptions

CommandLineTool only designed to execute single command

- Discouraged from doing this:
`samtools view -BS file.sam | samtools sort -o file.bam`
- Can be done by manipulating CommandLineTool arguments, but...
- Best practise is to create two CommandLineTools in series for this situation
- WDL / Nextflow embrace a free-form script section where you can write arbitrary shell code
- IMO this enables more rapid progression from 1+ shell scripts to a formalised workflow

```
cwlVersion: v1.2
class: CommandLineTool
baseCommand: ["salmon", "index"]
requirements:
  - class: ResourceRequirement
    coresMin: 4

inputs:
  transcriptome:
    type: File
    inputBinding:
      position: 2
      prefix: "-t"

arguments:
  - position: 1
    prefix: "--threads"
    valueFrom: "${runtime.cores}"
  - position: 3
    prefix: "-i"
    valueFrom: "salmon_index"

outputs:
  output_index:
    type: File
    outputBinding:
      glob: "salmon_index"
```



```
process INDEX {
  input:
    path transcriptome

  output:
    path 'salmon_index'

  script:
    """
    salmon index --threads $task.cpus -t $transcriptome -i salmon_index
    """
}
```



Users

Career bioinformaticians

Bit of a learning curve, but generally solid option once familiar.

Multiple high-profile research ventures use CWL

- NCI Cancer Genomics Cloud
- Seven Bridges
- Human Cell Atlas
- European Bioinformatics Institute
- Etc

Shines in collaborative efforts

- Researchers can pool CommandLineTool / Workflow descriptions
- Run on multiple different compute environments



Accessing Tools / Workflows

Lots of CWL is open-source!

Github / Gitlab repos best bet.

- UNLOCK <https://gitlab.com/m-unlock/cwl>
- KidsFirst <https://github.com/kids-first>
- bio-cwl-tools <https://github.com/common-workflow-library/bio-cwl-tools>

For Today

Will translate 2 CWL CommandLineTools to Nextflow

- https://github.com/genome/analysis-workflows/blob/master/definitions/tools/samtools_flagstat.cwl
- https://github.com/genome/analysis-workflows/blob/master/definitions/tools/gatk_haplotype_caller.cwl

Will translate 1 CWL Workflow to Nextflow

- https://github.com/genome/analysis-workflows/blob/master/definitions/subworkflows/align_sort_markdup.cwl

Let's Begin