Bespin

An open source system to run reproducible workflows on cloud infrastructure

DAN LEEHR, JOHN BRADLEY, HILMAR LAPP, DAVID CORCORAN
Enable Researchers to Run Workflows

Bioinformatics Workflow
- reusable
- reproducible
- efficient

Domain Researcher

Cloud Infrastructure
Bespin is Workflows and Services

CWL Bioinformatics Workflows + Services Run Workflows
Workflows are Versioned

CWL Workflows Versioned in github releases

Docker images built with version specific tags

hints:
- class: DockerRequirement
dockerPull: 'dukekobcb/trim-galore:0.4.4'
Methods
DNA-Seq data was processed using the TrimGalore toolkit\(^1\) (v0.4.4), which employs Cutadapt sequencing adapters from the 3’ end of the reads. Reads were aligned to the b37 version (v0.7.12) Alignment processing and variant calling were performed using the GATK\(^4\) toolkit Workflow\(^5\) (July 2017).

References

CWL CommandLineTool

```yaml
- class: SoftwareRequirement
  packages:
    bwa:
      version: [ "0.7.12" ]
  s:citation: https://dx.doi.org/10.1093/bioinformatics/btp324
```
Services Enable Running Workflows

- Website
  - Select Workflow
  - Configure Job
  - Run Job

- Jobs API

- Job Control Backbone
  - Creates VMs
  - Allocates Storage
  - Stages Data
  - Runs Workflows

Website → Jobs API → Job Control Backbone
Output and a Reproducibility Kit

**Input**
- Input Data Set
- Workflow Parameters

**Bespin Job**

**Output**
- Output Data
- Methods document

**Reproducibility Kit**
- Workflow
- Complete Job Order
- Instructions to run manually
- Logs created during job
Full Automation may produce dubious results

Input Data

Output Data

? batch effects
? systemic bias
? irregularities

Bespin Job
Experts Verify Results

- Input Data
- Verified Data
- Output Data

Bioinformatics Expert

Bespin Job
Source Code

github.com/Duke-GCB/bespin
github.com/Duke-GCB/bespin-cwl

Images

tryingtofly.deviantart.com/art/Star-wars-fan-art-sketch-559414158
openclipart.org