Deliverable

D9.3 Generic workflow engine

Design, implementation and deployment of a COMPARE computational workflow environment for autonomous sequence data analysis

Version: 1.0

Due: 36

Completed: 36

Authors: Nima Pakseresht and Guy Cochrane, EMBL-EBI

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

Introduction to report

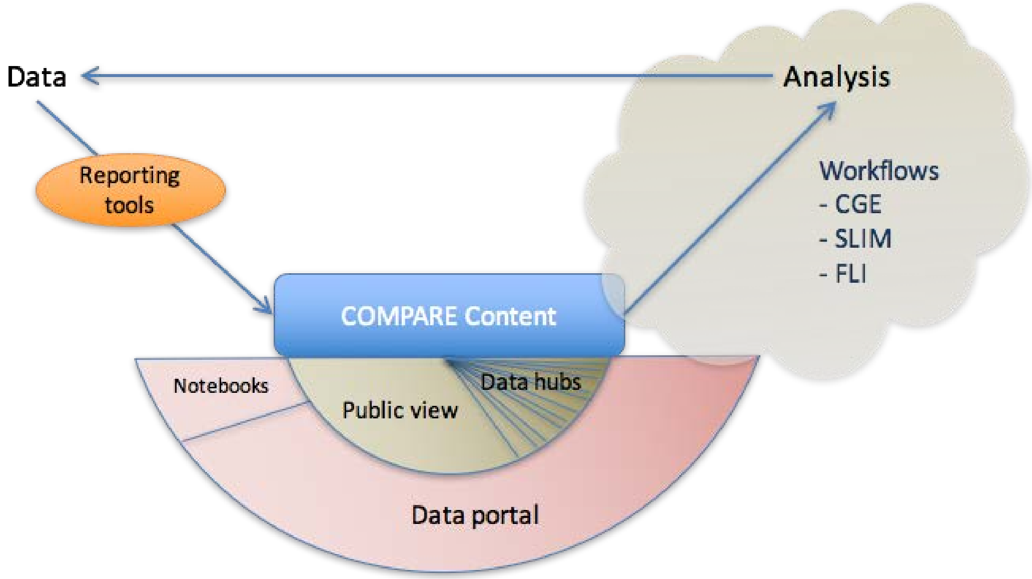
In this report, we detail our work on the design, implementation and production deployment of a new COMPARE Work Package 9 software component. This component is a generic workflow engine that sits within the COMPARE Data Platform and serves to automate and manage systematic computational analysis processes on incoming pathogen sequence data. We first provide context for the development of this component, move to details of the system itself, cover its current usage and, finally, lay out future plans. In order to provide context, we reference the computational analysis workflows that operate on the workflow engine, but we note that it is the workflow engine, rather than the specific computational analyses, that are the subject of this deliverable.

Context for a workflow engine

COMPARE Data Platform

The COMPARE Data Platform supports the sharing and analysis of whole genome pathogen sequencing data. Comprising a number of components, a user workflow typically involves the reporting of raw data into the system through one of the reporting tools, the sharing of these data with collaborating scientists through a Data Hub, the autonomous triggering of computational analysis through one of several cloud-hosted workflows, as appropriate for the input data, and the presentation of the raw data and analysis outputs through search, navigation and visualisation tools within the Data Portal (see figure I).

FIGURE I: THE COMPARE DATA PLATFORM SHOWING MAJOR COMPONENTS.



Requirements of compute in COMPARE

Support for systematic computational analysis in COMPARE is provided based on three key requirements. First, the system must allow the installation of a number of different parallel computational workflows from COMPARE partners with expertise in different areas of analysis. Second, the system must be scalable and appropriate for extensible compute infrastructure. Third, the system must support autonomous computational analysis, from detection of incoming data, through selection of an analysis workflow appropriate for these data, to the reporting of the results of analysis appropriately to be accessible by COMPARE users. We have met the first and second requirements by leveraging cloud compute infrastructure, specifically from the EMBL-EBI

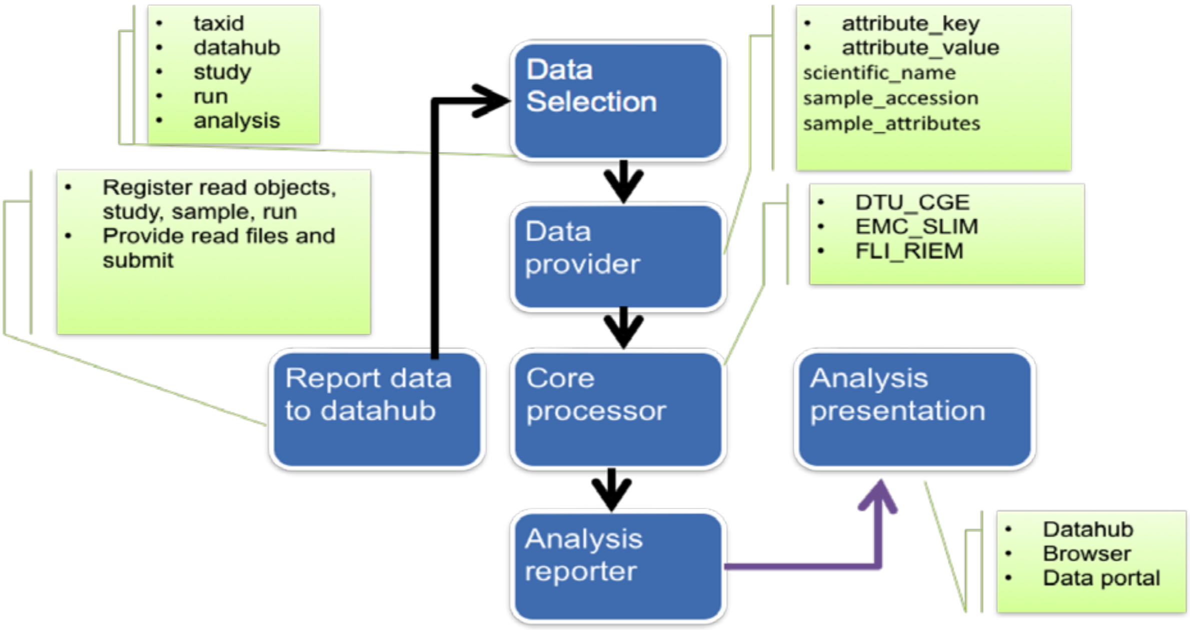
Embassy system, to allow scaling in future and portability to alternative cloud infrastructure. We meet the third requirement with a “workflow engine” that provides autonomous management of systematic compute in the cloud environment; we refer to this workflow engine internally as “SELECTA”.

SELECTA: the workflow engine

System design

The SELECTA system supports a number of processes, as laid out in figure II. Once data have been reported by users into an appropriate Data Hub, the SELECTA system operates rule-based selection of data, captures data from the hub into compute processes, triggers and manages computational actions upon the data and finally reports the results of the analysis back into the Data Hub as a “derived” or “analysis” data file.

FIGURE II: SCHEMATIC OF PROCESSES WITHIN SELECTA, THE COMPARE WORKFLOW ENGINE



Implementation

SELECTA has been implemented as a database-centric system with supporting process code written in Python. The core is a set of database tables that capture data selection rules, analysis parameters, sources of data, summary of metadata and all tracking information for the processes through which SELECTA must run, as detailed in the previous section.

Data selection can currently be configured to operate in two modes. In the first of these, SELECTA is configured to select data sets defined by project accessions. These accessions represent high-level records that group data sets within the system. Project records are created by data providers at the point of submission and can be used as a one-off for a given study submitted at one time, or over an extended period of time for multiple submitted data sets representing, for example, a surveillance programme. The second mode allows an entire Data Hub to be given over to a particular computational workflow; regardless of project accession, if the data provider opts to present a data set in the Data Hub, SELECTA will select the data sets.

In developing SELECTA, we have defined a set of conventions for those wishing to install computational workflows in the system. These include requirements relating to the structure of the software to be used in the workflow, the (open) licensing of software components and the presentation of the software in an open code repository within GitHub. For each workflow, we add a new class to SELECTA that defines the workflow within the system and allows its automation. We provide support to the software developers in preparing and installing their workflows.

SELECTA's analysis reporting process serves to manage the data files that are the output from a given computational workflow. In the case of one of the installed workflows, that of DTU CGE (see further details below), for example, this output is a compressed archive of intermediate files (covering such areas as assembly data and blast search results) and an accompanying summary table (in tab-separated value format). The first step is to capture these output files from the computational workflow and the second to despatch them, via the data reporting API, into the appropriate Data Hub. The outcome of this final autonomous process from SELECTA is the appearance of the analysis results in the Data Hub (and hence the Data Portal), for exploration by users.

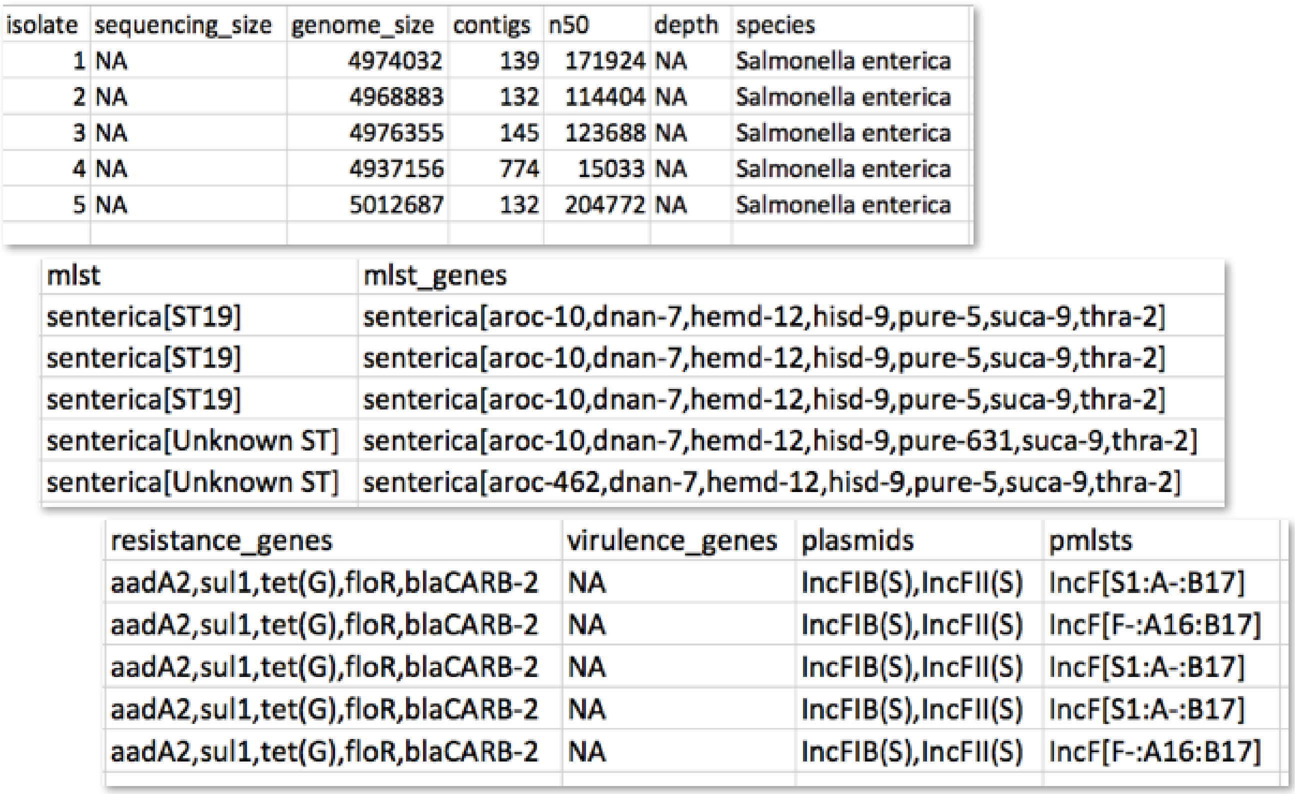
Availability

SELECTA has been in production since February 2017; since this time we have released improved and extended functionality iteratively.

Use

At the time of writing, there are two production computational workflows, each variously configured to operate across relevant Data Hubs on appropriate data sets. The first computational workflow, deployed in February 2017, was the DTU's CGE system for the analysis of bacterial isolate data (https://bitbucket.org/genomicepidemiology/cge-tools-docker; http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157718), with analysis outputs including MLST assignment and virulence and anti-microbial resistance gene and allele calls (see figure III).

FIGURE III: EXAMPLE OF SUMMARY OUTPUT DATA TABLE FROM THE DTU CGE WORKFLOW.



isolate	sequencing_size	genome_size	contigs	n50	depth	species
1	NA	4974032	139	171924	NA	Salmonella enterica
2	NA	4968883	132	114404	NA	Salmonella enterica
3	NA	4976355	145	123688	NA	Salmonella enterica
4	NA	4937156	774	15033	NA	Salmonella enterica
5	NA	5012687	132	204772	NA	Salmonella enterica

mlst	mlst_genes
senterica[ST19]	senterica[aroc-10,dnan-7,hemd-12,hisd-9,pure-5,suca-9,thra-2]
senterica[ST19]	senterica[aroc-10,dnan-7,hemd-12,hisd-9,pure-5,suca-9,thra-2]
senterica[ST19]	senterica[aroc-10,dnan-7,hemd-12,hisd-9,pure-5,suca-9,thra-2]
senterica[Unknown ST]	senterica[aroc-10,dnan-7,hemd-12,hisd-9,pure-631,suca-9,thra-2]
senterica[Unknown ST]	senterica[aroc-462,dnan-7,hemd-12,hisd-9,pure-5,suca-9,thra-2]

resistance_genes	virulence_genes	plasmids	pmlsts
aadA2,sul1,tet(G),floR,blaCARB-2	NA	IncFIB(S),IncFII(S)	IncF[S1:A-:B17]
aadA2,sul1,tet(G),floR,blaCARB-2	NA	IncFIB(S),IncFII(S)	IncF[F-:A16:B17]
aadA2,sul1,tet(G),floR,blaCARB-2	NA	IncFIB(S),IncFII(S)	IncF[S1:A-:B17]
aadA2,sul1,tet(G),floR,blaCARB-2	NA	IncFIB(S),IncFII(S)	IncF[S1:A-:B17]
aadA2,sul1,tet(G),floR,blaCARB-2	NA	IncFIB(S),IncFII(S)	IncF[F-:A16:B17]

The following month, we deployed a second workflow: the SLIM system from ERASMUS MC (https://github.com/EBI-COMMUNITY/slim_emc). This workflow focuses on the analysis of viral mixtures, yielding identification and typing data, again in tabular form.

At the time of reporting, two further computational workflows are in preparation for installation in SELECTA. These are the FLI's RIEMS viral metagenomics pipeline (https://github.com/EBI-COMMUNITY/fli-RIEMS; https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0503-6) and the BACPIPE system emerging from Work Package 3. We plan for them to be added to the SELECTA system in Q1 and Q2 of 2018.

Future plans

There are number of further developments planned for the SELECTA system:

- Extend the underlying system to support computational cluster access within cloud and other compute infrastructures, in order to scale with computational demand.
- Switch to PostgresSQL from MySQL to support future functionality extensions.
- Ensure SELECTA is portable across compute infrastructures; adapt to the Ansible framework and build VMs and clusters using a technology such as Terraform.
- Build an analysis report integration system that allows grouping of results based on selections of studies, taxa and sample records.
- Develop an API that can be used to provide control communications to SELECTA, with a view to (possible) expert user-led triggering of individual analyses.
- Add a third data selection mode: by taxon (tax ID) and Data Hub.
- Improve the SELECTA management interface.
- Add update functionality to allow refreshing of previous analyses in Data Hubs.
- Add iterative analysis capacity – the ability to start a computational analysis based on combinations of pre-existing computational analyses (e.g. in tree-building across isolates).
- Complete work on supporting installation of new computational analysis workflows to SELECTA (RIEMS and BACPIPE) and support new COMPARE workflows that arise.