WebMeV: A Platform for Intuitive Genomic Data Analysis

Yaoyu E. Wang, Ph.D

Quantitative Biomedical Research Center
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Key Contributors

John Quackenbush  
Brian Lawney  
Anastasia Serebryakova  
Derrick DeConti
Overall Goal of WebMeV

To help assure that analytical access to large public data is democratized so that scientists and physicians can test hypotheses by directly interacting with the data in a way that is not limited by their available computational resources and in a system that helps ensure their research is reproducible.
Genomic Data Consumer Spectrum

**Bioinformaticists/Data Scientists**
- Start with raw data (i.e. fastq)
- Process raw data by privately tuned pipelines
- Perform secondary data analysis on self processed data
- Construct secondary analysis pipeline from software packages
- Let data drive scientific hypothesis generation

**Translational Scientists**
- Start with a specific hypothesis derived from observation
- Select samples/patients of interest for the hypothesis
- Find processed data to perform secondary analysis
- Use readily available tools
- Interpret results in the context of initial hypothesis
Interactive Data Visualization for Transcriptomic Data and Analysis
Results for Public Domain Data

http://mev.tm4.org/

©2020 HSPH-QBRC
Cohort Selection and Set Manipulation Tool

View Details
- View cohort details
- View aggregate statistics
- View value distribution

Actions:
- Filter data to analyze for selected cohort
- Search by self define facets
- Build composite phenotypes
- Build cohort sets
Current Production WebMeV Architecture

Front-end application server driving RServe on GCP for most of the analysis
Current Aims for WebMeV

Specific Aim 1: Maintain and expend WebMeV functionalities and further optimize interactive data visualization system for large high-dimensional genomic data

Specific Aim 2: Integrate into WebMeV a FASTQ file transfer and processing system allowing users to start their analysis from raw sequence data

Specific Aim 3: Integrate methods for systems biology into WebMeV centered on gene network inference and analysis and develop new interactive network-based displays for gene networks
Year 1: Fortify backend analysis engine

• Current WebMeV architecture lacks a flexible backend for large scale computation

• It does not have a convenient mechanism to process privately generated sequence

• It lacks the flexibility to easily incorporate novel analysis pipeline for visualization on WebMeV front end
CNAP: Cloud-based analysis engine optimized for rapid pipeline customization, data upload, and analysis reproducibility

Key features

1) Interface with consumer cloud storage services and parallelize transfer of large data files;

2) Utilize Workflow Description Language (WDL) and the Cromwell engine for managing distributed and scalable analysis pipelines

3) Enforce reproducibility by default, requiring git integration and containerization
CNAP: Cloud-based analysis engine optimized for rapid pipeline customization, data upload, and analysis reproducibility

Key features

1) Interface with consumer cloud storage services and parallelize transfer of large data files;

2) Utilize Workflow Description Language (WDL) and the Cromwell engine for managing distributed and scalable analysis pipelines;

3) Enforce reproducibility by default, requiring git integration and containerization.
CNAP: Cloud-based analysis engine optimized for rapid pipeline customization, data upload, and analysis reproducibility

Key features

1) Interface with consumer cloud storage services and parallelize transfer of large data files;

2) Utilize Workflow Description Language (WDL) and the Cromwell engine for managing distributed and scalable analysis pipelines

3) Enforce reproducibility by default, requiring git integration and containerization
CNAP: Cloud-based analysis engine optimized for rapid pipeline customization, data upload, and analysis reproducibility

Key features

1) Interface with consumer cloud storage services and parallelize transfer of large data files;

2) Utilize Workflow Description Language (WDL) and the Cromwell engine for managing distributed and scalable analysis pipelines

3) Enforce reproducibility by default, requiring git integration and containerization
CNAP: Cloud-based analysis engine optimized for rapid pipeline customization, data upload, and analysis reproducibility

Key features

1) Interface with consumer cloud storage services and parallelize transfer of large data files;

2) Utilize Workflow Description Language (WDL) and the Cromwell engine for managing distributed and scalable analysis pipelines

3) Enforce reproducibility by default, requiring git integration and containerization

4) Easy to deploy on GCP

1) Consumer cloud storage allows parallelized transfer

2) Use WDL and Cromwell for defining and managing pipeline

3) Enforce analysis reproducibility by requiring git and docker containerization
CNAP is available on Github and Docker Hub

CNAP

The Cloud Native Analysis Platform is a web-application platform for reproducible, cloud-based data analysis. CNAP enforces reproducibility in both the application and the computing environment through required git integration and containerization. It integrates with consumer storage providers such as Dropbox and Google Drive for convenient and rapid data transfer.

- To learn more about how CNAP enforces reproducible analyses see here.
- For details on setting up the CNAP platform on the Google Cloud platform, see the documentation here.
- For information on creating and integrating CNAP workflows, see the documentation here.

Architecture

CNAP Workflow

- (a) The CNAP application.
- (b) handles the ingestion of workflows.
- (c) The transfer of data between consumer cloud storage and cloud platform storage.
- (d) Issues workflow requests to a remote Cromwell server.
- (e) Serves web interface to clients.
- (f) The Cromwell server manages the execution of workflows using containerized pipeline components.

Screenshots

https://github.com/qbrc-cnap/cnap
https://hub.docker.com/orgs/hsphqbrc/
Aim 3: Integrate systems biology methods into WebMeV centered on gene network inference

Regulatory Network Methods

PANDA: Integrate multi-omic data for network inference

LIONESS: estimate individual sample network
CNAP allows WebMeV to more easily integrate with ITCR tools.
Thank you