PANOPLY—A cloud-based platform for automated and reproducible proteogenomic data analysis

D. R. Mani[1], **Myranda Maynard**[1], Karsten Krug[1], Ramani Kothadia[1], Karen Christianson[1], David Heiman[2], Karl R. Clauser[1], Gad Getz[2] and Steven A. Carr[1]

[1]Proteomics Platform, [2]Cancer Genome Computational Analysis, Broad Institute of MIT and Harvard, Cambridge, MA

Overview

PANOPLY is a cloud-based **p**latform for **a**utomated a**n**d repr**o**ducible **p**roteogenomic data ana**ly**sis, enabling the use of state-of-the-art statistical and machine learning algorithms to transform multi-omic data from cancer samples into biologically meaningful and interpretable results. Salient features of PANOPLY include:

- Comprehensive collection of algorithms from CPTAC landmark proteogenomic studies[1-3] and more;
- Is easy to use;
- Integrates Genomic, Proteomic, and PTM data analysis; and
- Automates flexible and reproducible workflows.

It has been applied to routine proteogenomic analysis of a range of CPTAC datasets including breast cancer (BRCA), uterine cancer (UCEC)[3], lung adenocarcinoma (LUAD)[2], lung squamous cell carcinoma (LSCC), glioblastoma (GBM), pancreatic ductal adenocarcinoma (PDAC) and pediatric brain tumors (PBT).

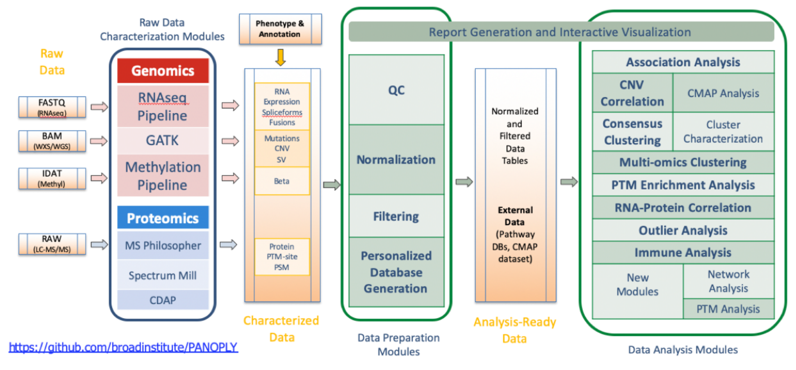
Introduction

Recent technological advances in NGS and MS-based proteomics have enabled the rapidly advancing field of **proteogenomics**—the integrative analysis of genomic, transcriptomic, proteomic, and post-translational modification (PTM) data. Many landmark studies[1-3] by the Clinical Proteomic Tumor Analysis Consortium [CPTAC, proteomics.cancer.gov] have highlighted its impact in promoting deeper insights in cancer biology and in potential drug target identification. PANOPLY is a collection of state-of-the-art algorithms for proteogenomic and multi-omic data analysis, packaged in a simple and easy to use interface with the goal of producing biologically meaningful and interpretable results.

Features & Functionality

PANOPLY leverages **Terra**, [app.terra.bio] to include proteogenomic workflows. It is a Google cloud-based platform developed at the Broad Institute for extreme-scale genome analysis and data sharing. It is designed to be:

- **Flexible–**
  - Easily combine and customize new pipelines using docker images and Workflow Description Language (WDL) specifications.
- **Automated—**
  - Preprogrammed executions.
  - Reuse previous computations (job avoidance) to improve scalability and reduce costs.
- **Reproducible—**
  - Export and share entire pipelines and associated data, with version-control and associated digital object identifiers (DOI).
- **Scalable and Secure—**
  - Inherently scalable cloud-based architecture.
  - Appropriate access control to enforce data privacy requirements.



PANOPLY Architecture Overview

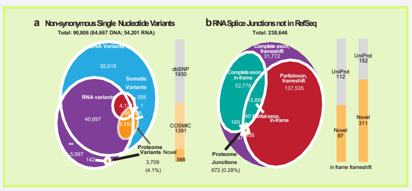
PANOPLY implements a wide array of algorithms applicable to all cancer types. In addition disease-specific customizations can be easily added.

PANOPLY Application to CPTAC Data

[Illustrations show TCGA Breast Cancer analysis]

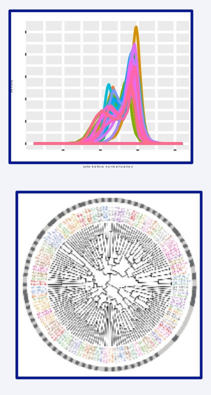
Samples and Data

- 108 TCGA samples profiled with 37 iTRAQ 4-plexes
- 105 tumors –
  - 25 basal-like
  - 29 luminal A
  - 33 luminal B
  - 18 HER2-enriched
  - 3 normal breast
  - 77 tumors used in study after QC
- Global proteome and phosphoproteome
- Genome–WXS, CNV, RNAseq
- **Cost –** Approx. $10 for compute and storage (including MS search, and proteogenomic analysis)

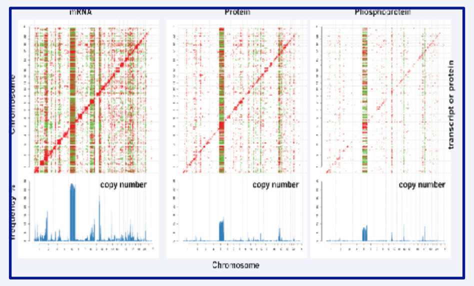


Patient-specific database search of 105 CPTAC breast cancer samples identified protein coding single amino acid variants (SAAVs) not present in RefSeq that were detected by DNA exome sequencing, RNA-seq, and LC-MS/MS (**a**)

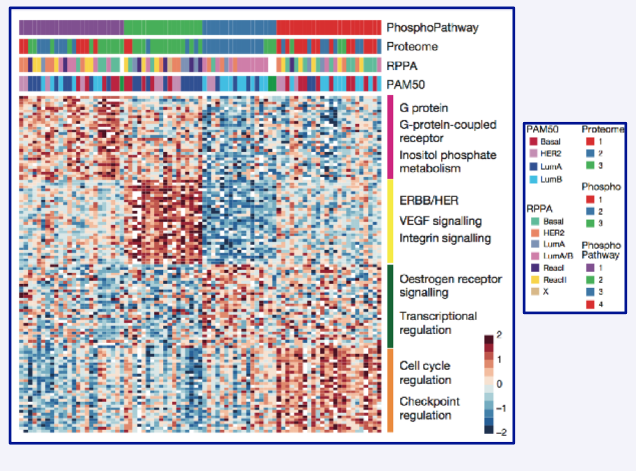
and novel splice junctions identified by LC-MS/MS (**b**).



Sample and normalization quality control (QC) provides several metrics to evaluate sample quality including profile plots of ratio distributions, sample co-clustering and correlation based on RNA, CNA and proteome data, and tumor purity assessment using ESTIMATE scores.



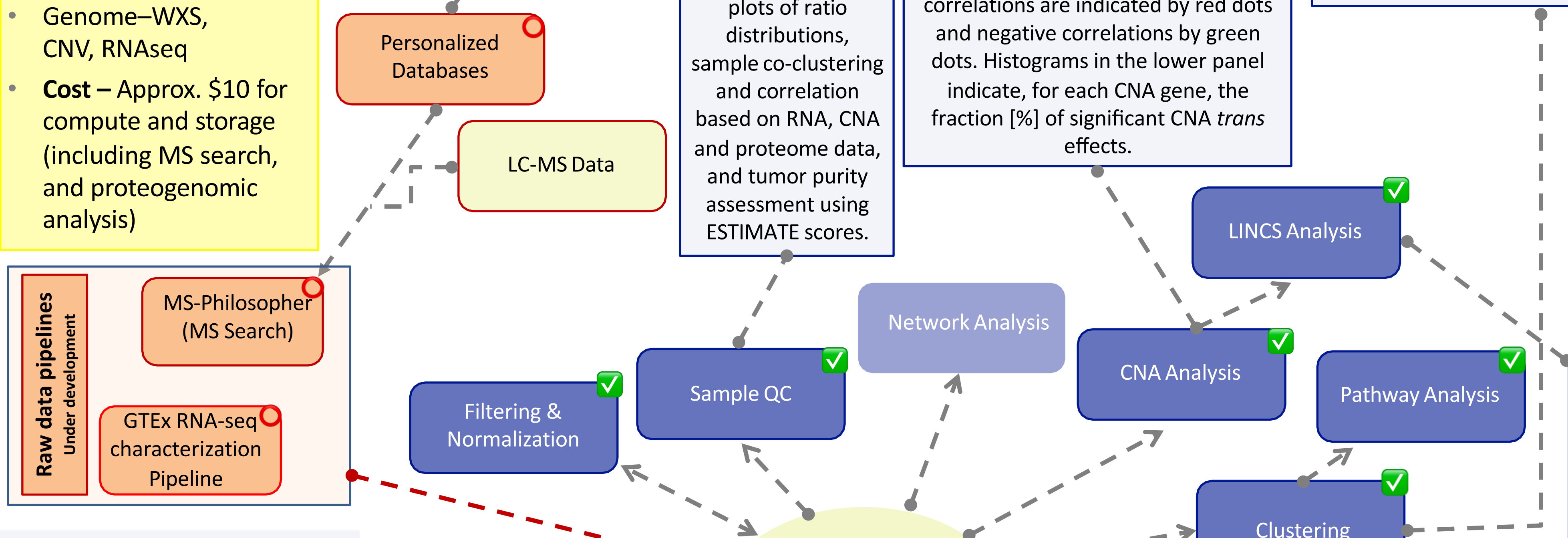
Global **correlation of copy number alterations** to RNA, protein expression and phosphoprotein levels highlighting CNA *cis* and *trans* effects for 77 breast cancer samples. CNAs are plotted on the x-axis; chromosome locations of mRNAs, proteins, and phosphoproteins are plotted on the y-axis. Significant (FDR<0.05) positive correlations are indicated by red dots and negative correlations by green dots. Histograms in the lower panel indicate, for each CNA gene, the fraction [%] of significant CNA *trans* effects.

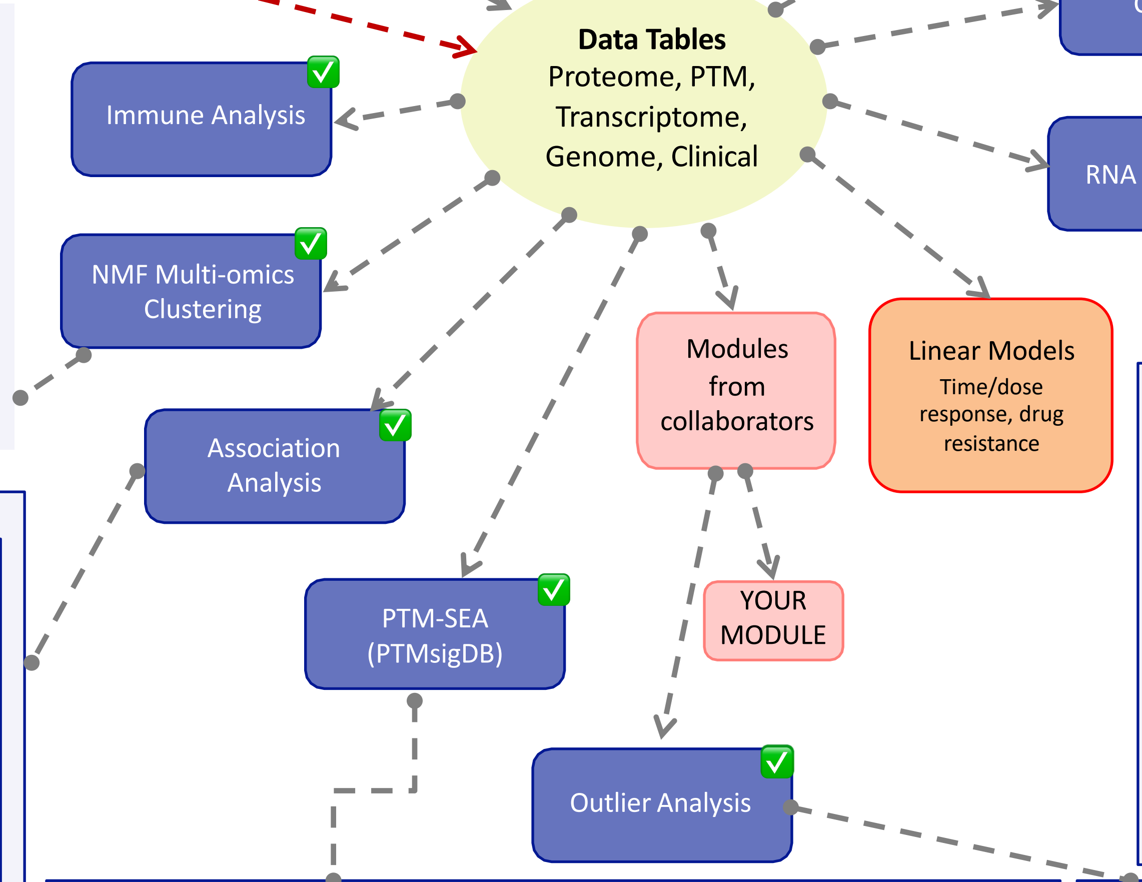


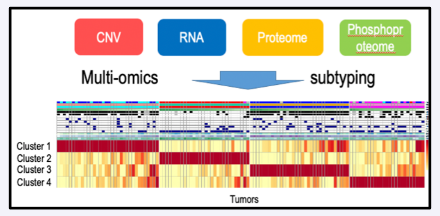
K-means **consensus clustering performed on pathways** derived from single sample GSEA analysis of phosphopeptide data identifies four distinct clusters. One of these clusters (cluster # 1) is unique to the phosphoproteome.

✅ Available
Underdevelopment

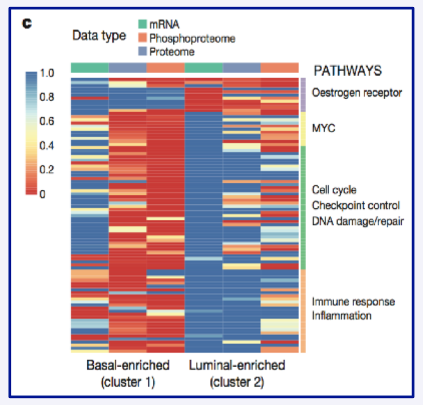
Some modules are under development. Some plots are manually generated using results from PANOPLY.



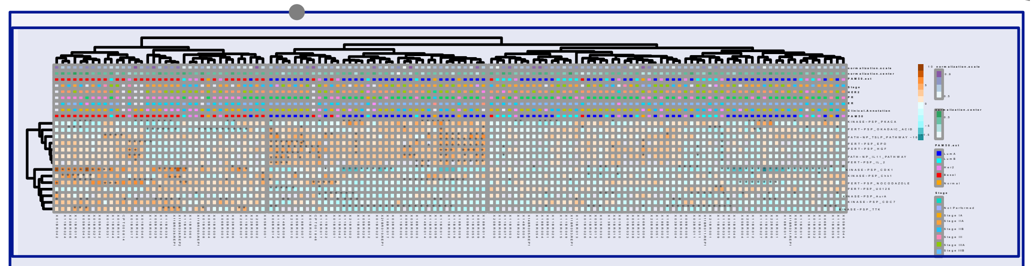

Data Tables
Proteome, PTM, Transcriptome, Genome, Clinical
Clustering
Immune Analysis
RNA Correlation
NMF Multi-omics Clustering
Modules from collaborators
Linear Models
Time/dose response, drug resistance
Association Analysis
PTM-SEA (PTMsigDB)
YOUR MODULE
Outlier Analysis



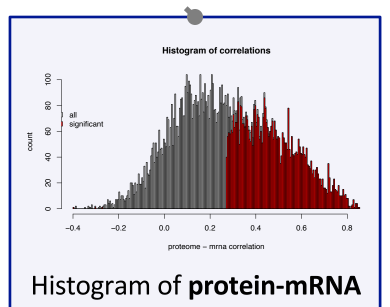
Multi-omics clustering using non-negative matrix (NMF) factorization enables automated discovery and biological characterization of cancer **subtypes**.



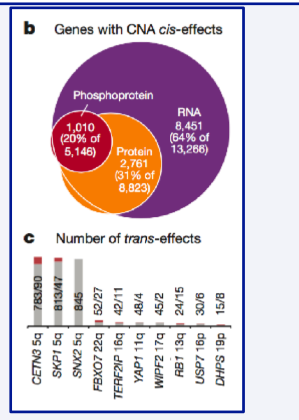
Association analysis followed by gene set enrichment analysis highlights sets of pathways significantly differential between two of the proteome clusters (derived via consensus clustering of proteome data). Association analysis also performs **marker selection and ranking** based on variable importance in a variety of classifiers.



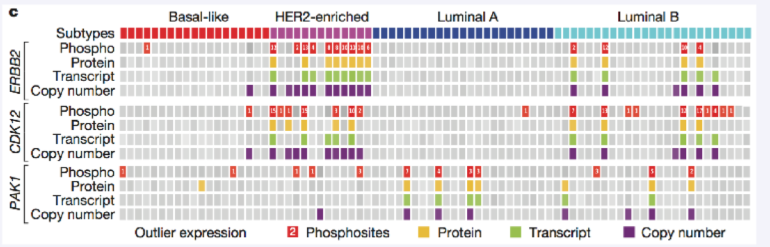
A database of **post-translational modification signatures** (PTMsigDB) is curated and single-sample **PTM signature enrichment analysis** (PTM-SEA) is used to determine enrichment of site-level phosphoproteome data[4]. Kinase substrates for CDK1, CHK1 are activated in basal tumors, while the TSLP and IL11 pathways are activated in a subset of luminal A tumors. Several drug perturbation signatures are also enriched in basal and luminal tumors.



Histogram of **protein-mRNA correlation** for 77 CPTAC breast cancer tumors. Statistically significant (FDR < 0.05) correlations are shown in red. Median correlation = 0.38.

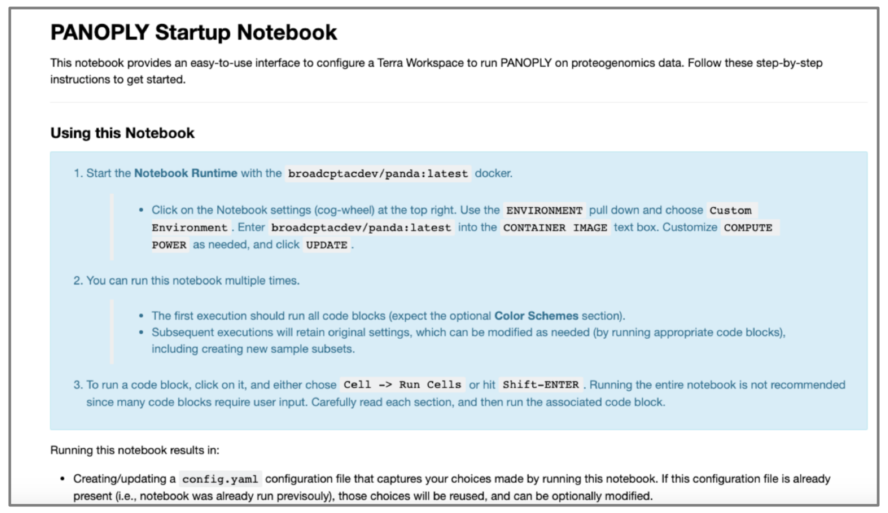


TOP: Overlap of ***cis*-effects** observed at RNA, protein, and phosphoprotein levels (FDR < 0.05). BOT: ***Trans*-effect** regulatory candidates identified among those with significant protein cis-effects using **LINCS CMap analysis**. Bars indicate total numbers of significant CNA–protein *trans*- effects (grey; FDR < 0.05) and overlap with regulated genes in LINCS knockdown profiles (red; 4 cell lines; moderated t-test FDR < 0.1)



Proteogenomic **outlier expression analysis** identifies samples with kinase outliers, and results for ERBB2, CDK12, and PAK1 are displayed. Samples with outlier phosphosite (red), protein (yellow), RNA (green) and copy number (purple) expression are shown.

PANOPLY User Interface



PANOPLY Startup Notebook

This notebook provides an easy-to-use interface to configure a Terra Workspace to run PANOPLY on proteogenomics data. Follow these step-by-step instructions to get started.

Using this Notebook

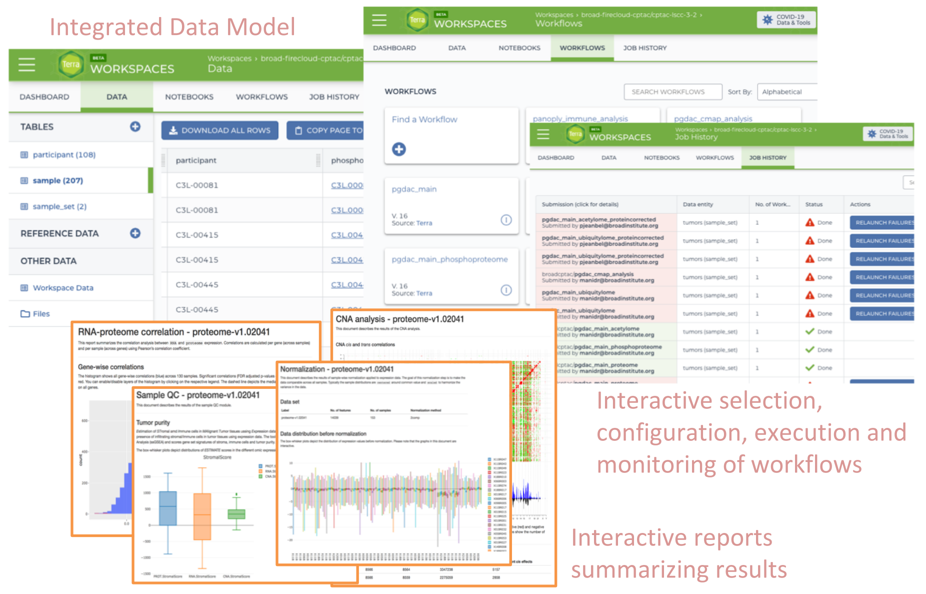
1. Start the **Notebook Runtime** with the `broadcptacdev/panda:latest` docker.
   - Click on the Notebook settings (cog-wheel) at the top right. Use the `ENVIRONMENT` pull down and choose `Custom Environment`. Enter `broadcptacdev/panda:latest` into the `CONTAINER IMAGE` text box. Customize `COMPUTE POWER` as needed, and click `UPDATE`.
2. You can run this notebook multiple times.
   - The first execution should run all code blocks (expect the optional **Color Schemes** section).
   - Subsequent executions will retain original settings, which can be modified as needed (by running appropriate code blocks), including creating new sample subsets.
3. To run a code block, click on it, and either chose `Cell -> Run Cells` or hit `Shift-ENTER`. Running the entire notebook is not recommended since many code blocks require user input. Carefully read each section, and then run the associated code block.

Running this notebook results in:

- Creating/updating a `config.yaml` configuration file that captures your choices made by running this notebook. If this configuration file is already present (i.e., notebook was already run previsouly), those choices will be reused, and can be optionally modified.

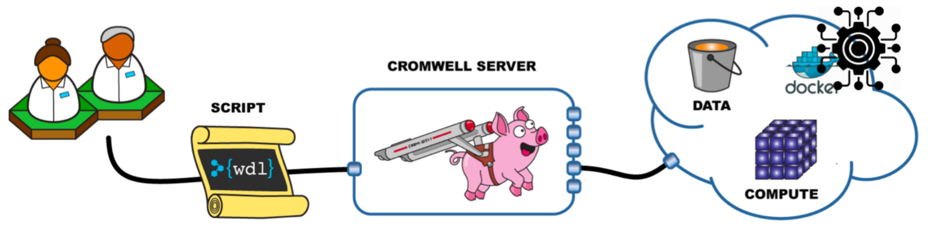
PANOPLY Startup Notebook

Provides simple step-by-step instructions for uploading data, identifying data types, specifying parameters, and setting up the PANOPLY workspace.



Add your module to PANOPLY

- **Docker--** Encapsulate your code files in a docker.
- **Docker Hub--** Build and push it to a docker registry of your choice.
- **WDL--** Write a WDL such that it -
  - Sources the right docker container image.
  - Calls the algorithm code, passing matched arguments.
  - Integrates with other modules via inputs/outputs.
- **Terra Workflow--** Save this WDL as a workflow on Terra.
- Plug in this method into PANOPLY workflows.



Summary

PANOPLY implements a high throughput cloud-based platform for transforming genomic, proteomic, phosphoproteomic and other post-translational modification (PTM) data into biologically meaningful information easily accessible to scientists worldwide. The flexible, robust, reproducible and automated platform has been applied to a range of CPTAC and other cancer datasets.

Availability

Code and documentation at: https://github.com/broadinstitute/PANOPLY

References

1. Mertins, P., Mani, D. R., Ruggles, K. V., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*.
2. Gillette, M. A., Satpathy, S., Cao, S., et al. (2020) Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell.*
3. Dou, Y., Kawaler, E., et.al. (2020). Proteogenomic Characterization of Endometrial Carcinoma. *Cell*.
4. Krug K., et. al. (2018). A curated resource for phosphosite-specific signature analysis. MCP.

Funding

This work was supported by grants U24CA210979 and U24CA210986 from the National Cancer Institute as part of Clinical Proteomic Tumor Analysis Consortium (CPTAC, http://proteomics.cancer.gov) initiative.