ORCESTRA: a platform for orchestrating and sharing high-throughput pharmacogenomic analyses

Anthony Mammoliti[1,2], Petr Smirnov[1,2], Minoru Nakano[1], Zhaleh Safikhani[1,2], Chantal Ho[1], Gangesh Beri[1], Benjamin Haibe-Kains[1,2,3,4*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[3]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[4]Ontario Institute of Cancer Research, Toronto, Ontario, Canada

*Corresponding author: Benjamin Haibe-Kains, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2C4 Canada

Phone: +1 (416) 581-7628

E-mail: benjamin.haibe.kains@utoronto.ca

Abstract

Reproducibility is essential to Open Science, as there is limited relevance for finding that cannot be reproduced by independent research groups, regardless of its validity. It is therefore crucial for scientists to describe their experiments in sufficient detail so they can be reproduced, challenged, and built upon. However, due to recent advances in the biological and computational sciences, it has become difficult to process, analyze, and share data with the community in a manner that is transparent. This has made reproducing research findings more challenging, with some researchers going as far as suggesting that the biomedical sciences are experiencing a "reproducibility crisis". To overcome these issues, we created a cloud-based platform called ORCESTRA (www.orcestra.ca), which provides a flexible framework for the reproducible processing of multimodal biomedical data. The platform enables processing of genomic and pharmacological profiles of cancer samples through the use of automated processing pipelines that are user-customizable, which are executed through Pachyderm, a data versioning and orchestration tool. ORCESTRA creates an integrated and fully documented data object known as a PharmacoSet (PSet), with a persistent identifier (DOI), that can be used and shared for future analyses using the Bioconductor PharmacoGx package.

Introduction

The demand for large volumes of multimodal biomedical data has grown drastically, partially due to active research in personalized medicine, and further understanding diseases [1–3]. This shift has made reproducing research findings much more challenging because of the need to ensure the use of adequate data handling methods, resulting in the validity and relevance of studies to be questioned [4]. Even though sharing of data immensely helps in reproducing study results [5], current sharing practices are inadequate with respect to the size of data and corresponding infrastructure requirements for transfer and storage [2,6]**.** As computational processing required to process such data is becoming increasingly complex [3], expertise is now needed for building the tools/workflows

for this large-scale handling [1,2]. There have been multiple community efforts in creating standardized workflow languages, such as the Common Workflow Language (CWL) and the Workflow Definition Language (WDL), along with associated workflow management systems such as Snakemake and Nextflow, in order to promote reproducibility [7,8]. However, although Nextflow and Snakemake are advantageous in regard to portability [9,10], a large learning curve is encountered for these programming heavy solutions, in comparison to user-friendly data processing platforms like Galaxy, which provide both storage and compute resources, but have limited features and scalability [11–13]. Therefore, there is a dire need for reproducible and transparent solutions for processing and analyzing large multimodal data that are scalable, while providing full data provenance.

The generation of large volumes of data has led to a sharing paradigm in the research community, where data is more accessible and open for public use. For studies to be reproduced and investigated for integrity and generalization by other researchers, the sharing of raw and processed data is crucial. However, providing access to data is not enough to achieve full reproducibility, as the shared data must be identifiable, transparent, and reusable, as outlined in the Findability, Accessibility, Interoperability, and Reusability (FAIR) data principles [14]. These foundational principles include providing rich metadata that is detail-oriented, including a persistent unique-identifier (Findability), accessing (meta)data with authentication and the unique-identifier using a communications protocol (Accessibility), assigning (meta)data with a commonly understood format/language (Interoperability), and achieving data provenance with an accessible usage-license (Reusability). Therefore, there is a need for a community-agreed standard for sharing multimodal biomedical data, such as the FAIR principles, which when translated into practice would promote efficient handling, sharing, and meaningful interpretation of such data.

Pharmacogenomic data are a prime example of multimodal biomedical data with a long history of sharing in the field of biomarker discovery. Pharmacogenomics involves the use of a

genome-wide association approach to understand the effect of drug compounds on the genome, which includes correlating drug response with molecular profiling, such as gene expression [15,16]. This is important for cancer patients, as treatment administration and success does not follow the notion of "one-size-fits-all", resulting in a reduction of patient survival and adverse effects such as toxicity [17]. Therefore, there is a need for precision medicine, where new therapies/strategies that are tailored towards a patient for improving treatment success/outcome. One step towards achieving this is through preclinical pharmacogenomics [18], such as studying the effect of drug compounds on cancer cell lines [19]. However, the utilization of pharmacogenomic data in a clinical setting has been slim, which may be attributed to non-reproducible findings in the field [20,21]. Facilitating the use and documentation of complex data processing, analysis, and sharing will substantially improve the reproducibility and validity of pharmacogenomic research.

In order to address these issues, we developed ORCESTRA (www.orcestra.ca)*, a* cloud-based platform that provides a transparent, reproducible, and flexible computational framework for processing and sharing large multimodal biomedical data. The ORCESTRA platform integrates genomic and pharmacological profiles of cancer samples through the orchestration of automated processing pipelines to curate a customized and fully documented data object for future analyses.

Complexity and challenges of pharmacogenomic data

The high-dimensionality, complexity, and scale of pharmacogenomic data present unprecedented challenges for researchers in the biomedical field, in regards to their ability to effectively manage, track, and process the data. The nature of heterogeneous and complex data negatively impacts data provenance, through incomplete or no accompaniment of metadata for a dataset, resulting in the uncertainty of a data lineage [22–24]. Because the granularity of metadata is a determinant of the value of a dataset [25], it should provide a rich description of dataset content, following the FAIR

data principles, which includes information about dataset origin, how it was generated, if there were any modifications that were made to it from precedent versions, and what these modifications were [14,26,27]. However, datasets published online, including ones that reside in repositories and from journals are often not accompanied with sufficient metadata [28]. In the field of genomics, issues with metadata often include mislabelling/annotation of data (e.g., incorrect identification numbers), improper data characterization (e.g., mapping files to respective samples and protocols), and inconsistency in the way metadata is presented/communicated (non-uniform structure used across consortiums) [29]. Overall, data quality will suffer and a low level of confidence will arise with newly performing or repeating analyses with a dataset that contains poor or no metadata. Provenance also extends to the scientific workflows/pipelines that are developed to process datasets [2], as sharing relevant source code is often not provided [30] along with relevant documentation about the workflow, such as in graphical user interface (GUI) based systems like Galaxy, affecting the ability to reproduce results [2]. In addition, pharmacogenomic data maintainers/consortiums, such as the Cancer Cell Line Encyclopedia (CCLE) [31], the Genomics of Drug Sensitivity in Cancer (GDSC) and the Genentech Cell Line Screening Initiative (gCSI) [32], often only process the dataset using one pipeline that they believe is the most suitable, without providing any reasoning in the form of documentation as to why the chosen processing pipeline was selected over other competing ones in the field [32,33]. Therefore, only a single-form of the dataset is released, which makes it difficult for other researchers to perform a diverse set of analyses that require the use of different processing pipelines on the dataset. A lack of provenance and utilization of a dataset in a single-form affects transparency, expressing a need for sharing pharmacogenomics data in a reproducible manner.

Current approaches to sharing pharmacogenomics data and pipelines

In the field of pharmacogenomics where genomics data are combined with phenotypic measurements of drug response, data portals have been created for accessing and sharing the

molecular and pharmacological data components of these large datasets, but with limitations in regard to reproducibility [34] (Table 1). The Genomic Data Commons Data Portal (NIH/NCI GDC) hosts raw data for the Cancer Cell Lines Encyclopedia (CCLE) from the Broad Institute, including RNA-seq, WXS, and WGS, allowing users to select and download the data type(s) of interest. Obtaining the data can be done through direct download or their GDC Data Transfer Tool by providing a manifest file that possesses the unique-identifiers (UUID) of each file, which also allow users to locate the files again through the portal, along with their corresponding run, analysis, and experimental metadata. This is advantageous, as all the raw data (public and controlled access), for both datasets are located within one portal and can be accessed in an efficient manner. However, no release notes are provided for any data that is newly uploaded or modified to GDC, which makes it challenging for researchers to keep track of different versions of the dataset and ensure their analyses on it are reproducible. Moreover, the recent addition of new CCLE data (e.g. additional RNA-seq cell lines) [33], is found on the European Nucleotide Archive (ENA), but not on GDC, resulting in data source inconsistency that becomes difficult to manage and follow for users. In addition, current and previous versions of other CCLE data (i.e. annotation, drug-response) are hosted on a Broad Institute portal, but there are no release notes or documentation present with each version, forcing researchers to manually identify changes within each file after every release. gCSI, a dataset generated by Genentech, provides easy access to their drug response and processed molecular data via the compareDrugScreens R package (http://research-pub.gene.com/gCSI-cellline-data/) [32], along with their raw molecular data through the European Genome-Phenome Archive (EGA) and ArrayExpress data portals. GRAY, a dataset generated by Dr. Joe Gray's lab at the Oregon Health and Science University, has had three updates with raw data hosted on NCBI, and drug-response and annotation data hosted on SYNAPSE, DRYAD, and/or the papers supplementary section [35–37]. Because each version of the dataset is associated with a different respective paper, the data is scattered among various hosts, which makes it challenging to keep track of each source, and for each source to ensure that the

data remains readily available, as one failed link would make it difficult for a researcher to reproduce any results. However, for the GRAY dataset, NCBI provides detailed information about the methodology used for the experiments, SYNAPSE provides a wiki and contact source for the dataset and a provenance tracker for each file that is uploaded, and DRYAD stores each publications data as a package organized with subsequent descriptions to keep data organized. Along with CCLE, the Genomics of Drug Sensitivity in Cancer (GDSC) from the Wellcome Trust Sanger Institute also represents good data sharing practices, as they also provide a data downloading tool, as well as a single web page including links and descriptions to both drug-response and raw data, keeping everything in one desired space. Other than a short description of each data file, they do not provide any information about what was specifically updated within each file after an update, making it difficult to track provenance. Therefore, there is a need for a platform that acts as a common point of entry for both molecular and drug sensitivity data, as well as combining all of the positive data sharing practices used by the data portals available.

Key functionalities for an effective orchestration platform

The increasing utilization and demand for big data have resulted in the need for effective data orchestration [38], which is a process that involves organizing, gathering, and coordinating the distribution of data from multiple locations across compute resources with specific processing requirements, in order to meet the needs of an end user. An ideal orchestration platform for handling large-scale heterogeneous data would consist of the following: (1) a defined workflow; (2) a programming model/framework [38], and (3) broad availability of a compute infrastructure. At the workflow level, data from different sources/lineages, including data that is not static, must be effectively managed through the definition of workflow components (tasks) that interact and rely on one another [38]. Moreover, a programming model should be utilized for the workflow components responsible for handling the respective data (static/dynamic), such as a batch processing model (MapReduce)[38]. Lastly, the utilization of a scalable compute environment (i.e.

cloud) would allow for the management and processing of the big data, with requirements such as necessary compute and storage resources, ability to transfer data, and monitoring of executed workflows and respective components/tasks, including tracking data provenance. There are multitudes of publicly-available orchestration platforms and tools that are currently being used for the storage, processing, and sharing of genomic data, each with their own respective advantages and disadvantages, to our knowledge (Table 2). Our platform utilizes Pachyderm (http://www.pachyderm.io/), an open-source orchestration tool for multi-stage language-agnostic data processing pipelines, maintaining complete reproducibility and provenance through the use of Kubernetes, with the following characteristics:

Programming language. Supports creating and deploying language-agnostic pipelines across on-premise or cloud infrastructures.

Large dataset support. Users can upload large datasets to the Pachyderm file system for utilization in their pipelines.

Automatic pipeline triggering. Reproducibility and provenance are highlighted through its features like automatic pipeline triggering when existing data is modified or newly added, which results in the automatic generation of new versions of an output data object. However, because automatic triggering requires the state of each pod within the Kubernetes cluster to be saved, there is a permanent allocation of CPU/RAM for each pod (and therefore each pipeline), which requires a user to create a cluster with a great deal of potentially costly resources.

Reprocessing. A feature that is found in Pachyderm, DNANexus, and Lifebit is the prevention of recomputation for each pipeline trigger, which comes in handy when a pipeline contains processed raw data that does not need to be reprocessed if there is a change in metadata such as an annotation file.

Docker utilization. Each pipeline can be equipped with a Docker image connected to Docker Hub for running various toolkits, which allows for simplistic pipeline updating when there are future updates to any component of the Docker image.

Versioning of data and pipelines with unique identifiers. Each commit is supplied with a unique identifier, which is updated with each new commit (parent-child system), allowing users to track different versions of a pipeline and dataset with ease.

Parallelism support. A pipeline can be parallelized (constant or coefficient strategy) using workers, which is useful for workloads with large computational requirements.

Data versioning system. Pachyderm provides direct GitHub integration for data versioning, which enables users to track changes at the file-level and submit updates to Pachyderm through commits triggered through webhooks on GitHub. In addition, this also provides users with the ability to publicly view, track, and share all updates made to a pipeline or file connected to Pachyderm with ease.

Open access. Pachyderm provides a free and open-source version of the tool, as well as an enterprise version that provides additional features.

Direct mounting of data. The current version of Pachyderm does not allow for direct mounting of data from a cloud storage system (i.e. bucket) to a Pachyderm repository. Data must be transferred to the tools own file system, resulting in essentially an additional copy of the data within a cloud environment. A feature that Databricks and Lifebit offer that is immensely important for decreasing computation time and cost is that they are not copying data into a file system for it to be used by the platform. This is important when large data sizes will be used in an analysis, which allows a user to simply store their data in a bucket/blob storage account, and mount it to the platform of interest, giving the user the ability to also use the data with other platforms or cloud services without having to repeatedly copy it in an inefficient manner.

Cost efficiency. Pachyderm utilizes VM's through a Kubernetes cluster if deployment on a cloud environment, which are costly to keep running indefinitely. Therefore, utilizing Pachyderm on a cloud infrastructure impacts cost efficiency, in comparison to an on-premise HPC. A notable feature that is supported by Lifebit, is cost-efficiency through low-priority instance utilization on a cloud provider, allowing for users to execute large-scale analyses at a reduced cost.

Resource allocation. Pachyderm requires persistent RAM/CPU allocation for each pipeline within the Kubernetes cluster, even after a pipeline is successfully executed, which permits automatic pipeline triggering. Thus, an increased amount of compute resources (VM's scaled up/out) may be required for specific pipelines, which also impacts cost efficiency.

ORCESTRA platform

We have developed a cloud-based paradigm for data sharing and processing in pharmacogenomics based on automation, reproducibility, and transparency. The platform is deployed on Microsoft Azure and encompasses automated Pachyderm pipelines that allow a user to create a custom PharmacoSet (PSet), which is an R data object that stores molecular profile, drug-sensitivity, and experimental metadata for the respective cell lines and drugs, allowing for integrative analysis of the molecular and pharmacological data for biomarker and drug discovery [39]. The platform utilizes datasets from the largest pharmacogenomic consortia, including the Cancer Cell Line Encyclopedia (CCLE: Broad-Novartis) [31,33], the Genomics of Drug Sensitivity in Cancer (GDSC: Wellcome Trust Sanger Institute) [40,41], the Genentech Cell Line Screening Initiative (gCSI) [32,42], the Cancer Therapeutics Response Portal (CTRP: Broad Institute) [43–45], GRAY (Oregon Health and Science University breast cancer screen) [35–37], University Health Network Breast Cancer Screen (UHNBreast) [46–48], and the Institute for Molecular Medicine Finland (FIMM) [49,50]. After our curations, the platform collectively consists of 1,676 cell lines, 928 drug compounds, and 1,549,272 sensitivity experiments across multiple drug sensitivity releases, in the form of PharmacoSets (PSet) (Table 3). PharmacoSets can accommodate all types of molecular profile data, however, ORCESTRA currently supports gene expression (RNA-sequencing, microarray), CNV, mutation, and fusion molecular data. For RNA-seq data, a user is given the ability to select a reference genome of interest, a combination of quantification tools and their respective versions, along with reference transcriptomes from two genome databases (Ensembl, Gencode) to generate custom RNA-seq expression profiles for all of the cell lines in

the dataset. Therefore, each PSet will be generated through a custom orchestrated Pachyderm pipeline path, where each piece of input data, pipeline, and output data option is tracked and given a unique-identifier to ensure the entire process is completely transparent and reproducible. Unlike other platforms, to ensure PSet generation is fully transparent and that provenance is completely defined, each PSet is automatically uploaded to Zenodo and given a public DOI, where the DOI is shared via a personalized persistent web-page that possesses a detailed overview about the data that each DOI associated PSet contains and how it was generated. This includes publication sources, drug-sensitivity information and source, raw data source, exact pipelines parameters used for the processing tools of choice, and URLS to reference genomes and transcriptomes used by the tool(s). This page gets automatically sent to each user via email, providing users with one custom page that hosts all of the information required to understand how the PSet was generated. Therefore, all of the data used in the PSet is shared in a transparent fashion, where researchers can identify the true origins of all data used with confidence and effectively reproduce results.

ORCESTRA structure

In order for the platform to be as transparent as possible, it harnesses an architecture with three distinct layers that not only works independently to process and interpret precedent data, but also have the capability to scale (Figure 1). The first layer contains the web application which was developed using a Node.js API and React front-end with MongoDB as a database. The layer provides the user with an interaction point to the *ORCESTRA* platform, allowing users to search for existing PSets, request a new PSet by entering pipeline parameters, view PSet request status, and register a personal account to save existing PSets of choice.

The second data-processing layer encompasses a Kubernetes cluster on Microsoft Azure that hosts Pachyderm, which utilizes Docker images for PharmacoGx and respective R-packages. All of the RNA-seq raw data for the CCLE, GDSC, gCSI, GRAY, and UHNBreast PSets have

been pre-processed with Kallisto and Salmon Snakemake pipelines using an HPC environment, and subsequently pushed to assigned data repositories on Pachyderm, allowing for specified selection from the web-app (transcriptome and tool version). The Pachyderm PharmacoGx pipeline incorporates the RNA-seq and drug-sensitivity data specified via the web-app, along with the annotation data that is hosted on GitHub, and other molecular profile data (e.g. CNV, mutation) into a PSet (Figure 2). The GitHub hosted files encompass cell line, drug, and tissue annotation data that can be viewed at the file-level for changes and edited which automatically triggers the Pachyderm PharmacoGx pipeline with the new modifications to produce a new PSet. A unique feature of Pachyderm is the prevention of re-processing computed data, such as where an update of RNA-seq annotations will not trigger the re-processing of thousands of drug-sensitivity experiments. Each generated PSet enters the third data-sharing layer where the PSet gets automatically uploaded to an online data-sharing repository known as Zenodo, with a DOI so that the data object can be given a persistent location on the internet that can be uniquely identified. The generated DOI is then associated with a custom meta-data web page that is generated based on the contents of the PSet. Below is a detailed overview of each ORCESTRA layer:

Web Application Layer: Features and Use Cases

ORCESTRA comes with a web application interface allowing users to interact with the data-processing and data-sharing layers. These features are outlined in the following use cases:

Search and obtain information about existing PSets ("Search/Request" and "Individual PSet" views)

Users can search existing PSets in the "Search/Request" view by filtering existing PSets with the "PSet Parameters" panel. Users can filter existing PSets by selecting datasets with associated drug sensitivity releases, genome references, RNA-seq transcriptomes, RNA-seq processing tools with respective versions, which associates with other respective DNA data types (mutation or CNV) and RNA data types (microarray or RNA-seq). Changes in the parameter

selections trigger the web app to submit a query request to a MongoDB database which returns a filtered list of PSets (Figure 3). The PSet table is then re-rendered with an updated list of PSets. This allows users to search through existing PSets to determine if a PSet that satisfies users' parameter selections already exists, preventing recomputation. Information about the datasets and tools used to generate a PSet can be viewed by clicking on a PSet name and navigating to its PSet meta-data web-page. Users can obtain information such as associated publications, links to the raw drug-sensitivity and RNA-seq data as well as a Zenodo DOI. In addition, the individual PSet view provides users with the option to download the PSet of choice directly from the view.

Request a PSet with customized pipeline configuration ("Search/Request" view)

Users can request a customized PSet in the "Search/Request" view by turning the "Request PSet" toggle on. This action reconfigures the dropdown options in the "PSet Parameters" panel to be in request mode, and displays, on the "Summary" panel, two text input fields for entering a custom name for the PSet and a user's email to receive a notification upon PSet pipeline completion, with the accompanied Zenodo DOI and custom ORCESTRA meta-data page link. Pachyderm continuously scans for a new request from the web-app, which will automatically trigger the respective pipelines to build the custom PSet, while storing a unique ORCESTRA ID, Pachyderm pipeline commit ID, and Zenodo DOI into the MongoDB database, which increases the level of data provenance and reproducibility, as each PSet can be identified through three unique-identifiers after creation (Figure 3). The PSet filtering process as described above continues to function as users select the request parameters, which displays existing PSet(s) that satisfy users' parameter selections. Upon selecting all the required parameters, the "Submit Request" button becomes active for users to submit the pipeline request.

View usage metrics ("Home", "Statistics" and "Request Status" views)

ORCESTRA provides several usage metrics for users. These metrics can be accessed through "Home", "Statistics" and "Request Status" views. The "Home" view provides an overview of currently available datasets, tools and references to generate PSets, most downloaded PSets,

and a number of pending or in-process PSet requests. The “Statistics" view provides a visualized PSet popularity ranking, along with a plot of the number of cell lines, drugs, and genes for the canonical PSets, including intersection, which can be accessed by clicking the “View Statistics" button in the “Home” view. The “Request Status” view displays a tabulated list of PSet requests that are either “pending” (the request has been submitted and saved, but has not been processed in Pachyderm), or “in-process” (the request has been submitted and is processed in Pachyderm).

Register for an account and save “favorite” PSets (“User Profile” view)

ORCESTRA offers users the option to register for an account with a valid email address. Registered users are able to select existing PSets in the “Search/Request" view and save them as their “favorites" which can be accessed in the “User Profile” view. However, the web application keeps track of PSet requests submitted by users based on their email addresses even without registration. These PSets are automatically added to a user’s favorite PSets and can be viewed in the “User Profile” view.

ORCESTRA Deployment Options and Costs

In order to cater to a specific set of computing needs and resource limitations for each user, *ORCESTRA* offers flexibility in platform configuration and deployment for each of its three layers. The web application and data-processing layers can be customized and deployed onto different environments such as a remote server, a virtualized server on a cloud computing platform such as Microsoft Azure, Google Cloud and AWS, or an on-premise server. While the platform uses Zenodo as a data-sharing layer by default, it can be configured to use different data sharing repositories if a user wishes to do so. Such flexibility in the platform deployment and configuration enables users to deploy and utilize *ORCESTRA* within the limitations of their computing environment. For example, users may choose to deploy the web application and data-processing layers on a physical server located on-premise. This allows the users to enjoy greater control over the platform configuration and cost management, especially with regards to deploying a

Kubernetes cluster, a core framework used for the data-processing layer. Running the cluster on a cloud computing platform such as Microsoft Azure could incur a large cost (in the order of thousands of US dollars per month). Therefore, deployment of the data-processing layer on-premise may be favorable if an adequate server is available. On the other hand, commercially available cloud computing platforms typically come with a number of useful features to assist users in deploying web applications, databases and computing clusters. Microsoft Azure's App Service plan and database services for example, allow users to easily deploy web applications and configure databases without installing additional tools or software. Pachyderm can also be easily installed and configured using Azure Kubernetes Service (AKS). These build-in services offer users ease and reliability in deploying and maintaining *ORCESTRA*. With this deployment flexibility, users can choose to deploy each platform layer in different computing environments to meet their needs and limitations in computing and monetary costs. Users may choose to take advantage of the ease of web application deployment and maintenance processes offered by a commercially available cloud computing platform while deploying the data-processing layer on-premise to reduce the monetary costs. Another way *ORCESTRA* allows users to reduce computing costs is by having online/offline request processing features. For example, users who choose to deploy the data-processing layer on a cloud computing platform may be restricted to activate the data-processing layer only when pipeline requests need to be processed, and deactivate it once the pipeline is complete, in order to reduce the monetary cost. *ORCESTRA* accommodates such use case by providing Pachyderm status checking, offline request processing, and manual push request features as described below:

Pachyderm status check. Users can easily identify the status of Pachyderm (instances hosting it are either online or offline) with the status indicator on the right-hand corner of the top navigation bar on ORCESTRA. The web application checks the online/offline status each time the web application is reloaded onto a web browser.

Offline request submission. The web app interface allows users to submit pipeline requests even

when the Pachyderm cluster is offline. These requests are saved in the database with “pending” status which can be viewed in the “Request Status” page.

Manual pipeline request push. The web app interface provides an administration feature that allows an administrator to manage pending pipeline requests that are submitted while Pachyderm is offline. When a user is logged in as an administrator, the user can “push” those pending requests to the data-processing layer in the “Request Status” view, upon activation of Pachyderm.

Summary and Future Directions

ORCESTRA is a platform that harnesses a robust framework that ensures the reproducible and transparent processing and sharing of pharmacogenomic data with the scientific community. Through the use of Pachyderm, the platform is able to track provenance of all pharmacological and molecular data across a plethora of cell lines and drug compounds that are to be incorporated into a single unified data object for cancer research. Not only are we hoping to add additional pharmacogenomic datasets to the platform, but we are also interested in expanding our platform to securely process and host private PSets that can be accessed through the existing user profiles. This will allow researchers to effectively collaborate and share data with other teams, along with standardizing the space in which their data is hosted. Lastly, we aim to allow users to select additional tools for the processing of RNA-seq data, such as STAR and HISAT2, along with CNV and mutation data, which will increase the flexibility of the platform.

Acknowledgements

This project is supported by the Canadian Institutes of Health Research (CIHR), under the frame of ERA PerMed.

References

1. Madduri, R. *et al.* Reproducible big data science: A case study in continuous FAIRness. *PLoS One* **14**, e0213013 (2019).
2. Kanwal, S., Khan, F. Z., Lonie, A. & Sinnott, R. O. Investigating reproducibility and tracking provenance - A genomic workflow case study. *BMC Bioinformatics* **18**, 337 (2017).
3. Toga, A. W. & Dinov, I. D. Sharing big biomedical data. *J Big Data* **2**, (2015).
4. Huang, Y. & Gottardo, R. Comparability and reproducibility of biomedical data. *Brief. Bioinform.* **14**, 391–401 (2013).
5. Finak, G. *et al.* DataPackageR: Reproducible data preprocessing, standardization and sharing using R/Bioconductor for collaborative data analysis. *Gates Open Res* **2**, 31 (2018).
6. Suthakar, U., Magnoni, L., Smith, D. R., Khan, A. & Andreeva, J. An efficient strategy for the collection and storage of large volumes of data for computation. *Journal of Big Data* **3**, 21 (2016).
7. Goble, C. *et al.* FAIR Computational Workflows. *Data Intelligence* **2**, 108–121 (2020).
8. Kulkarni, N. *et al.* Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics* **19**, 349 (2018).
9. Federico, A. *et al.* Pipeliner: A Nextflow-Based Framework for the Definition of Sequencing Data Processing Pipelines. *Front. Genet.* **10**, 614 (2019).
10. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
11. Ahmed, A. E. *et al.* Managing genomic variant calling workflows with Swift/T. *PLoS One* **14**, e0211608 (2019).
12. Bourgey, M. *et al.* GenPipes: an open-source framework for distributed and scalable genomic analyses. doi:10.1101/459552.

13. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
14. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
15. Sanoudou, D., Mountzios, G., Arvanitis, D. A. & Pectasides, D. Array-based pharmacogenomics of molecular-targeted therapies in oncology. *Pharmacogenomics J.* **12**, 185–196 (2012).
16. T P, A., M, S. S., Jose, A., Chandran, L. & Zachariah, S. M. Pharmacogenomics: the right drug to the right person. *J. Clin. Med. Res.* **1**, 191–194 (2009).
17. Rodríguez-Antona, C. & Taron, M. Pharmacogenomic biomarkers for personalized cancer treatment. *J. Intern. Med.* **277**, 201–217 (2015).
18. Picco, G. & Garnett, M. J. A Road Map for Precision Cancer Medicine Using Personalized Models. *Cancer discovery* vol. 7 456–458 (2017).
19. Hafner, M., Niepel, M. & Sorger, P. K. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nat. Biotechnol.* **35**, 500–502 (2017).
20. Ventola, C. L. The role of pharmacogenomic biomarkers in predicting and improving drug response: part 2: challenges impeding clinical implementation. *P T* **38**, 624–627 (2013).
21. Ong, F. S. *et al.* Clinical utility of pharmacogenetic biomarkers in cardiovascular therapeutics: a challenge for clinical implementation. *Pharmacogenomics* **13**, 465–475 (2012).
22. Razick, S. *et al.* The eGenVar data management system--cataloguing and sharing sensitive data and metadata for the life sciences. *Database* **2014**, bau027 (2014).
23. Hu, R., Yan, Z., Ding, W. & Yang, L. T. A survey on data provenance in IoT. *World Wide Web J. Biol.* (2019) doi:10.1007/s11280-019-00746-1.
24. Sivarajah, U., Kamal, M. M., Irani, Z. & Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **70**, 263–286 (2017).

25. Corpas, M., Kovalevskaya, N. V., McMurray, A. & Nielsen, F. G. G. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput. Biol.* **14**, e1005873 (2018).
26. Wise, J. *et al.* Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov. Today* **24**, 933–938 (2019).
27. Boeckhout, M., Zielhuis, G. A. & Bredenoord, A. L. The FAIR guiding principles for data stewardship: fair enough? *Eur. J. Hum. Genet.* **26**, 931–936 (2018).
28. Kanduri, C., Domanska, D., Hovig, E. & Sandve, G. K. Genome build information is an essential part of genomic track files. *Genome Biol.* **18**, 175 (2017).
29. Learned, K. *et al.* Barriers to accessing public cancer genomic data. *Sci Data* **6**, 98 (2019).
30. Mangul, S. *et al.* Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 1393 (2019).
31. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
32. Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
33. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
34. Smirnov, P. *et al.* PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res.* **46**, D994–D1002 (2018).
35. Hafner, M. *et al.* Quantification of sensitivity and resistance of breast cancer cell lines to anti-cancer drugs using GR metrics. *Sci Data* **4**, 170166 (2017).
36. Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14**, R110 (2013).
37. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2724–2729 (2012).

38. Barika, M. *et al.* Orchestrating Big Data Analysis Workflows in the Cloud: Research Challenges, Survey, and Future Directions. (2019).

39. Smirnov, P. *et al.* PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).

40. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).

41. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–61 (2013).

42. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).

43. Rees, M. G. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).

44. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **5**, 1210–1223 (2015).

45. Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).

46. Marcotte, R. *et al.* Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell* **164**, 293–309 (2016).

47. Safikhani, Z. *et al.* Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nature Communications* vol. 8 (2017).

48. Mammoliti, A., Smirnov, P., Safikhani, Z., Ba-Alawi, W. & Haibe-Kains, B. Creating reproducible pharmacogenomic analysis pipelines. *Sci Data* **6**, 166 (2019).

49. Mpindi, J. P. *et al.* Consistency in drug response profiling. *Nature* **540**, E5–E6 (2016).

50. Pemovska, T. *et al.* Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov.* **3**, 1416–1429 (2013).

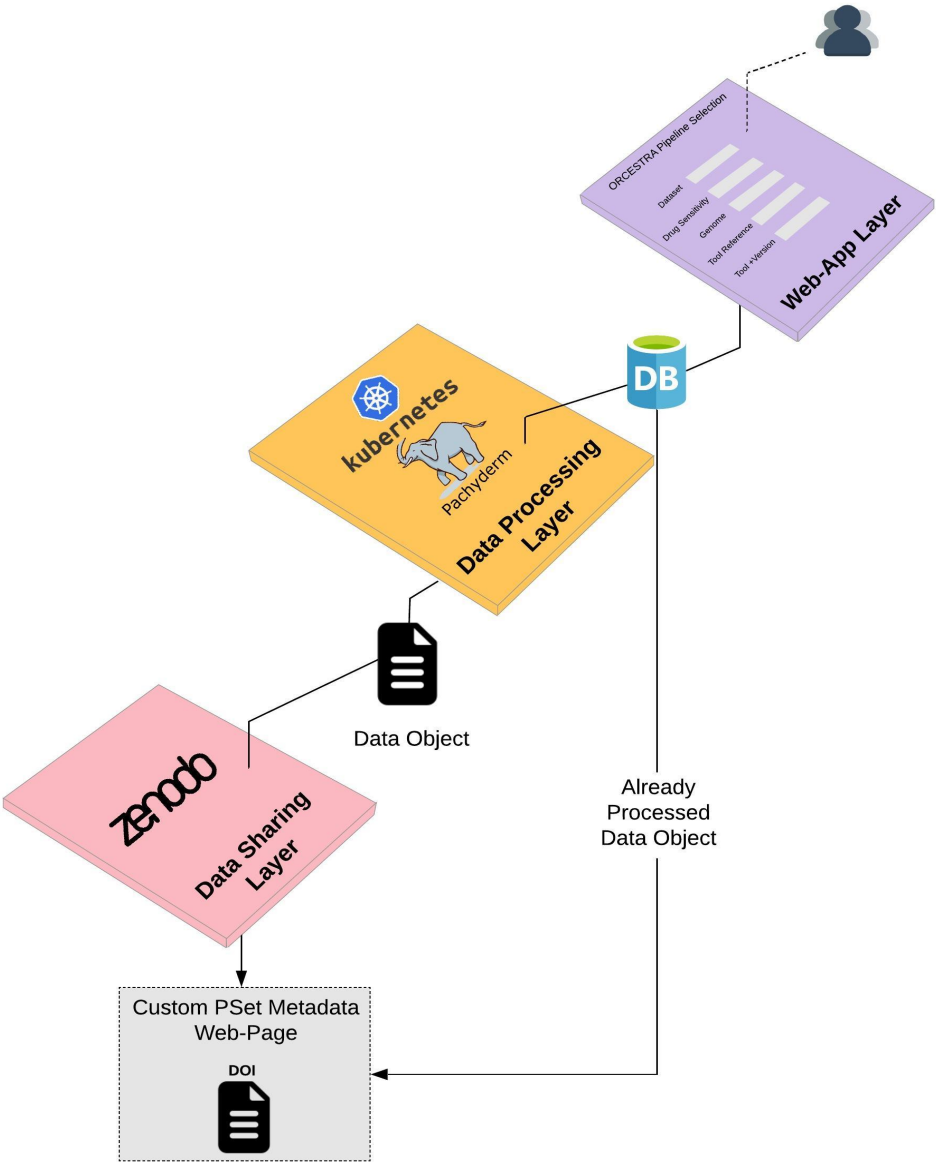

Figure 1. ORCESTRA framework layers for pipeline selection, data object generation, and DOI sharing with a custom metadata webpage.

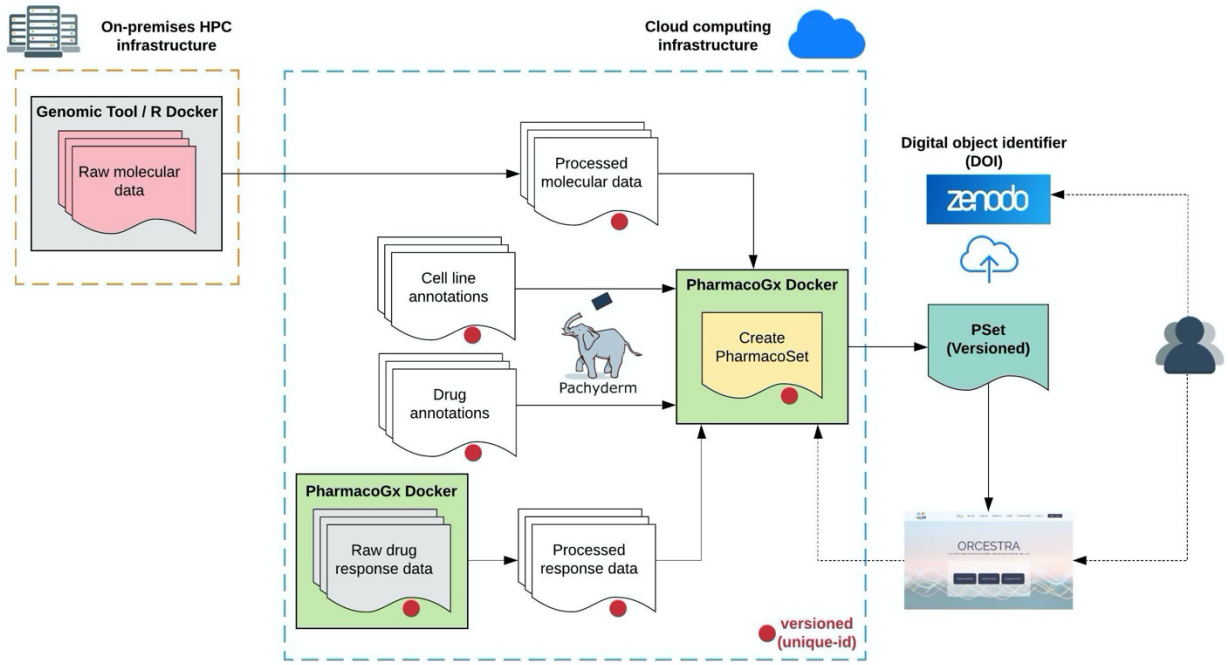

Figure 2. Cloud-based deployment of ORCESTRA data processing layer with automatically versioned data using Pachyderm and sharing of generated PSets through Zenodo via a persistent identifier (DOI).

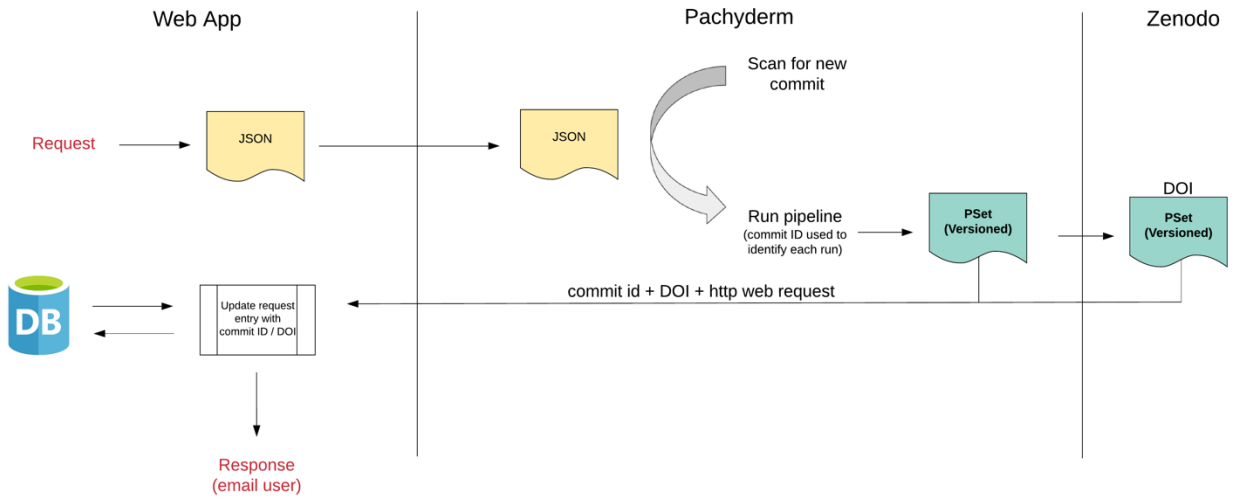

Figure 3. ORCESTRA web-application connectivity with data processing layer through commit ID scanning for user selected pipeline requests, and subsequent PSet DOI tracking with MongoDB queries.

Data Portal	Notable Functionalities	Expected Functionalities
SYNAPSE	Wiki integration Data provenance tracker Dataset discussion board	Track updates to a dataset at the file level
Broad Institute (CCLE)	Easy access of current, previous, and legacy data versions Display of data by data-type (e.g. pharmacological profiling, mRNA expression)	Display of data processing tools, and their versions, with all accompanying files (e.g. transcriptome)
DRYAD	Simplistic navigation Dataset metrics Respective publication download	Direct access to all processing pipelines utilized
NCBI	SRA sequencing information with additional sample metadata	Option of downloading processed data through multiple methods/pipelines
CancerRxGene (GDSC)	Direct integration of cell line and drug metadata Direct links to raw data hosted on other portals Description of changes to each drug sensitivity data release	

Table 1. Common data portals for sharing genomic data with notable and expected functionalities for transparent and reproducible processing, analysis, and interpretation.

Features	ORCESTRA (Pachyderm)	DNAnexus	Databricks	Lifebit
Create language agnostic pipelines in the cloud	✓	✓	✓	✓
Large dataset support (TB in size)	✓	✓	✓	✓
Automatic pipeline triggering with updated data (out-of-the-box)	✓	×	×	×
Prevents recomputation of entire dataset with each new pipeline trigger	✓	✓	×	✓
Docker utilization	✓	✓	✓	✓
Every pipeline run and data sources are versioned with unique identifier	✓	✓	*	*
Parallelism support	✓	✓	✓	✓
Versioning system (e.g. GitHub) for pipelines, input data, and pipelines	✓	✓	✓	✓
Open access (free)	✓	×	×	×
Direct mounting of data (no copying into file system)	×	×	✓	✓
Automatic cost efficiency implementation for instances (low-priority)	×	×	×	✓
No permanent resource allocation for a pipeline (memory/CPU)	×	✓	✓	✓

* indicates partial support of feature

Table 2. Data processing platforms and their respective features for handling genomic data.

Dataset	Drug sensitivity	No. of cell lines	No. of drugs	No. of drug sensitivity experiments
GRAY	2013	73	89	9413
GRAY	2017	74	107	9756
gCSI	2017	747	16	6471
GDSC	2019 (v1-8.0)	1104	303	323032
GDSC	2020 (v1-8.2)	1104	343	377305
GDSC	2019 (v2-8.0)	1104	169	197332
GDSC	2020 (v2-8.2)	1104	190	215780
UHNBreast	2019	84	8	689
FIMM	2015	50	52	2561
CTRPv2	2015	887	544	395263
CCLE	2015	1094	24	11670
Total		1676	928	1549272

Table 3. Datasets (PSets) processed and shared through ORCESTRA with respective drug sensitivity versions, number of cell lines and drugs, and number of drug sensitivity experiments.