REVIEW ARTICLE

JOURNAL OF PANCREATOLOGY®

OPEN

Exploiting open source omics data to advance pancreas research

Gayathri Swaminathan, Toshie Saito, Sohail Z. Husain*

Abstract

The "omics" revolution has transformed the biomedical research landscape by equipping scientists with the ability to interrogate complex biological phenomenon and disease processes at an unprecedented level. The volume of "big" data generated by the different omics studies such as genomics, transcriptomics, proteomics, and metabolomics has led to the concurrent development of computational tools to enable in silico analysis and aid data deconvolution. Considering the intensive resources and high costs required to generate and analyze big data, there has been centralized, collaborative efforts to make the data and analysis tools freely available as "Open Source," to benefit the wider research community. Pancreatology research studies have contributed to this "big data rush" and have additionally benefitted from utilizing the open source data as evidenced by the increasing number of new research findings and publications that stem from such data. In this review, we briefly introduce the evolution of open source omics data, data types, the "FAIR" guiding principles for data management and reuse, and centralized platforms that enable free and fair data accessibility, availability, and provide tools for omics data analysis. We illustrate, through the case study of our own experience in mining pancreatitis omics data, the power of repurposing open source data to answer translationally relevant questions in pancreas research.

Keywords: Gene expression, Omics, Open source data, Pancreatitis

The evolution of open source data sharing in biomedical research

The discovery and application of high-dimensional approaches to address biological questions ushered in the "omics" era, which caused an exponential increase in "big" data and a complementary suite of software tools and applications to mine and analyze large datasets. The term "omics" is used to describe studies of the entire complement of a specific biomolecule (eg, DNA, RNA, protein) in cells or the totality of the molecular processes in an organism. Some well-known examples with the suffix omics include genomics, transcriptomics, proteomics, and metabolomics, which refer to large-scale, global profiling of genes, RNA, protein, or metabolites, respectively, at the cellular or organism level. Such data is powerful as they launch wide characterization of biological phenomenon using appropriate in vitro and in vivo experimental models, and these methods have transformed the biomedical research landscape. Omics-based research has thus enabled us to gain a holistic view of physiological and pathological phenomenon at a single cell as well as systems level and spawned ground-breaking discoveries.

The high cost of omics approaches is an impediment for its widespread application in individual research laboratories and has been a major impetus to rally for open access to omics data from its originating source. Thus, the concept of "open source data" emerged, which may be defined as high-quality data that has an open license, is freely accessible, well-structured, machine-readable, well-maintained, and typically made available through government-funded (eg, National Institutes of Health) or private (eg, Chan Zuckerberg Initiative, American Association of Cancer Research) initiatives. To streamline data deposition, data quality, access, interpretation, integration, and enable optimal reuse of large-scale datasets to support knowledge discovery and innovation, the "FAIR" data guidelines (Table 1) were proposed during a meeting of various stakeholders engaged in funding, generating, analyzing, and publishing such datasets. These foundational principles of data management and stewardship help govern not only datasets but also analysis tools and workflows that were used to generate the datasets.[1] FAIR refers to findability, accessibility, interoperability, and reusability of data across different sources and disciplines. Some important aspects of the implementation of FAIR guidelines are the provision of standardized metadata information and its appropriate referencing to make it both human and machine-actionable.

Many journals now require authors to make their omics data publicly available at the time of manuscript publication to enable their reuse by other research groups and to enhance reproducibility. This movement toward open sharing of data and data-mining tools has reshaped scientific inquiry and discovery by facilitating a tremendous opportunity to repurpose the freely available data to either generate novel hypotheses for empirical testing in experimental models and clinical samples or to mine the data with specific questions. At present, the dissemination of open source data extend beyond the traditional omics platforms and encompasses large-scale datasets from spatial atlases (eg, Pancreatlas[2]; http://www.pancreatlas.org) of different tissues in various organisms, thus providing affirmation for the success of the open source data

Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA

GS and TS are co-first authors.

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

** Corresponding author: Sohail Z. Husain, Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Stanford University School of Medicine, 750 Welch Road, Stanford, CA 94305. E-mail: sohail.husain@stanford.edu*

Journal of Pancreatology (2024) 7:1

Received: 9 December 2023; Accepted 3 February 2024.

Published online 9 February 2024

http://dx.doi.org/10.1097/JP9.0000000000000173

Table 1

The FAIR principles governing open source data

Findable	To be findable, data must
	(1) have a unique and persistent identifier
	(2) be registered or indexed in a searchable resource
	(3) have sufficient metadata
Accessible	To be accessible, the data should be
	(1) supported by sufficient metadata and be human and machine-actionable
	(2) deposited in a trusted repository
Interoperable	To be interoperable,
	(1) data must share a common structure
	(2) metadata must use well-defined, formal vocabulary
Reusable	To be reusable, data and metadata should
	(1) have a clear and accessible data usage license
	(2) be associated with detailed provenance
	(3) meet community standards relevant to the domain

sharing concept. However, the precious open data resources are still largely underutilized. This is in part due to lack of knowledge about the available data types, different repositories/resources that host the data, and the technical know-how to access and leverage the data.

How to procure open source omics data and analyze it to address a research question?

Unlocking the treasure trove of data

The workflow (Fig. 1) typically begins with a focused search with pertinent keywords related to the research question of interest, of well-curated omics data repositories that host high-quality, "standardized" data sets. Frequently accessed, popular data repositories such as NCBI GEO, ArrayExpress, Expression Atlas (EMBL-EBI; https://www.ebi.ac.uk/biostudies/arrayexpress, https://www.ebi.ac.uk/gxa), The Cancer Genome Atlas (TCGA; NCI GDC portal; https://www.portal.gdc.cancer.gov) cBioportal[3] (Broad Institute, USA; https://www.cbioportal.org), etc, containing well-annotated, harmonized datasets have tutorials on how to construct an effective query (often using Structured Query Language or SQL). In addition, they have helpful search features to facilitate the retrieval of datasets relevant to the research question. For instance, a search query for a repository that hosts transcriptomics datasets (eg, NCBI GEO) could be tailored to look for changes in expression of a single gene across 2 defined experimental conditions (eg, control vs disease or control vs drug; query gene X and disease) that may link it to a functional phenotype in an organ of interest (eg, pancreas), or to unearth/reveal connected gene expression signatures that may result from a common underlying mechanism triggered by an experimental condition of interest. Other notable examples of open access resources that provide valuable omics data include the Human Protein Atlas (HPA; https://www.proteinatlas.org),[4] PanglaoDB (https://panglaodb.se),[5] and pancreas expression database (PED; https://www.pancreas-expression.org).[6] PanglaoDB provides access to and allows the exploration of several mouse and human single-cell gene expression datasets through an interactive, web-based interface. The HPA allows free access to and exploration of 12 unique sections each focusing on a particular aspect (eg, single cell type section, tissue section) of genome-wide analysis of the human proteome. The HPA program has contributed to thousands of publications since its inception in 2003 underscoring the power and utility of leveraging open source data for addressing research questions. The PED hosts clinical and molecular information from pancreas-specific datasets and incorporates published findings from pancreatic malignancies and benign conditions such as chronic pancreatitis. There are also centralized, open source repositories such as Omics Discovery Index (Omics DI; https://www.omicsdi.org) that host multiple omics datasets (Table 2) and provide user-friendly search and analysis tools to make data findable, accessible, and reusable.[7–9] We will provide some details of our experience in retrieving data from the OmicsDI platform in the latter part of the review.

Sifting through the data trove

Once we have identified a dataset of interest, how can we analyze the data? The impact editor at Nature, Anna Nowogrodzki, listed 11 tips for working with large data sets[10]: Cherish your data, Visualize the information, Show your workflow, Use version control, Record metadata, Automate, Make computing time count, Capture your environment, Don't download the data (downloading and storing large data sets is not practical), Start early, Get help. These tips can be applied to analyze data obtained from a public repository.

Data mining refers to the process of knowledge discovery in data. It requires sifting through an enormous amount of information with the purpose of extracting useful information that can help answer the research question. It often involves cleaning raw, unstructured data, finding patterns or correlations or associations. The analysis may vary in complexity. Information that we need, or desire, may be extracted from a single dataset or from disparate datasets followed by data integration, that require sophisticated computational tools and operational know-how. Analysis tools are also made freely available through an open source platform, including simple tools like GEO2R (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html) hosted by NCBI GEO for comparing gene expression or complex ones such as Seurat[11,12] (https://satijalab.org/seurat/) for handling single cell genomics data. Some sites such as Expression Atlas, PRIDE Restful API, TCGA, and cBioportal integrate both datasets (eg, genomics, transcriptomics, proteomics, clinical metadata) and analysis tools with tutorials to navigate data acquisition and analysis. This enables researchers with little or no familiarity in coding or even the basic use of bioinformatics tools to unpack an extensive omics dataset and procure the desired information. However, more complex data mining often requires collaboration with a computational biologist/bioinformatician.

As an illustration for the above workflow (Fig. 1), we provide details of how we leveraged open source data and analysis tools to complement our impactful studies on the protective role of vitamin A (VA) in asparaginase-associated pancreatitis (AAP).

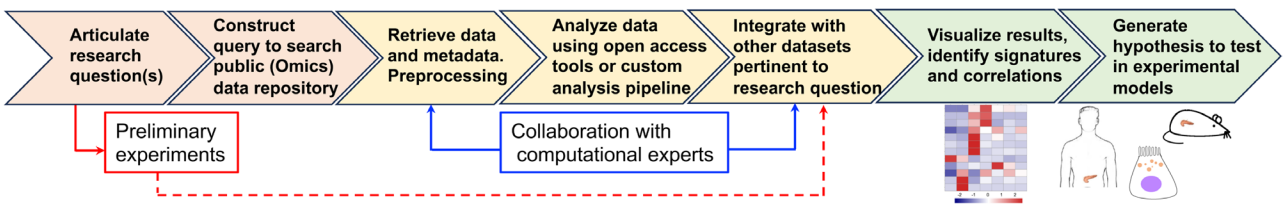


Figure 1. Exploiting open source data to advance pancreas research: a simple workflow.

Table 2

Catalog of databases hosted by OmicsDI

PRIDE (PRoteomics IDEntifications)

-A centralized, standards compliant, public data repository for mass spectrometry proteomics data. The data includes protein and peptide identification and corresponding expression values, post-translational modifications, and supporting mass spectra evidence, both as raw data and peak list files.
https://www.ebi.ac.uk/pride/archive/

PeptideAtlas

-A multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.
http://www.peptideatlas.org/

MassIVE (Mass Spectrometry Interactive Virtual Environment)

-A community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data.
https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp

GPMDB (Global Proteome Machine Database)

-The GPMDB utilizes the information obtained from GPM servers to aid in the difficult process of validating peptide MS/MS spectra as well as protein coverage patterns.
https://www.thegpm.org/

JPOST Repository (Japan ProteOme STandard Repository)

-jPOSTrepo is a data repository for sharing MS raw/processed data.
https://jpostdb.org/

Physiome Model Repository

-The repository is intended to provide a "quantitative description of physiological dynamics and functional behavior of the intact organism." Integration of this database into OmicsDI is in an early testing stage.
http://physiomeproject.org/

EGA (European Genome-Phenome Archive)

-The EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data obtained from biomedical research projects.
https://ega-archive.org/

EVA (European Variation Archive)

-An open-access database of all types of genetic variation data from all species.
https://www.ebi.ac.uk/eva/

ENA (European Nucleotide Archive)

-An open, supported platform for the management, sharing, integration, archiving and dissemination of sequence data.
https://www.ebi.ac.uk/ena/browser/home

LINCS (Library of Integrated Network-based Cellular Signatures)

-The LINCS program is an NIH Common Fund program that includes six Data and Signature Centers: Drug Toxicity Signature Generation Center, HMS LINCS Center, LINCS Center for Transcriptomics, LINCS Proteomic Characterization Center for Signaling and Epigenetics, MEP LINCS Center, and NeuroLINCS Center. The extensive, well-annotated datasets generated by the centers along with relevant experimental information are made openly available with user-friendly search interfaces to access and download the signatures and tools to display the data and perform integrative analysis.
http://lincsportal.ccs.miami.edu/dcic-portal/
https://lincsproject.org

PAXDB (Protein Abundance Database)

-A comprehensive absolute protein abundance database, which contains whole genome protein abundance information across organisms and tissues. Furthermore, it provides information about inter-species variation of protein abundances.
https://pax-db.org/

Cell Collective

-Interactive modeling of biological networks. Integration of this database into OmicsDI is at an early testing stage.
https://cellcollective.org/#

MetaboLights

-MetaboLights is a recommended repository for depositing information derived from Metabolomics experiments. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.
https://www.ebi.ac.uk/metabolights/index

Metabolomics Workbench

-A national and international repository for metabolomics data and metadata that provides analysis tools and access to metabolite standards, protocols, tutorials, training, and more.
https://www.metabolomicsworkbench.org/about/index.php

GNPS (Global Natural Products Social Molecular Networking)

-The GNPS is a platform for providing an overview of the molecular features in mass spectrometry-based metabolomics by comparing fragmentation patterns to identify chemical relationships.
https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp

(*Continued*)

Table2

(Continued)

BioModels

-BioModels is a repository of mathematical models of biological and biomedical systems. It hosts a vast selection of existing literature-based physiologically and pharmaceutically relevant mechanistic models in standard formats.

https://www.ebi.ac.uk/biomodels/

FAIRDOMHub

-The FAIRDOMHub is built upon the FAIRDOM-SEEK software suite, which is an open source web platform for sharing scientific research assets, processes, and outcomes. Integration of this database into OmicsDI is at an early testing stage.

https://fairdomhub.org/

ArrayExpress

-ArrayExpress archives functional genomics data from microarray and sequencing platforms to support reproducible research. Experiments are submitted directly to ArrayExpress or are imported from the NCBI Gene Expression Omnibus (GEO) database.

https://www.ebi.ac.uk/arrayexpress/

dbGaP (Genotypes and Phenotypes)

-The database of dbGaP was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

https://www.ncbi.nlm.nih.gov/gap/

ExpressionAtlas

-The Expression Atlas provides information on gene and protein expression patterns under different biological conditions. Gene expression data can be acquired and re-analyzed to detect genes showing interesting baseline and differential expression patterns.

https://www.ebi.ac.uk/gxa/home

GEO (NCBI Gene Expression Omnibus)

-GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users to query and download curated gene expression profiles with associated experimental information as well as perform differential expression analysis.

https://www.ncbi.nlm.nih.gov/geo/

NODE (The National Omics Data Encyclopedia)

-NODE provides an integrated, compatible, comparable, and scalable multi-omics resource platform that supports flexible data management and effective data release. NODE uses a hierarchical data architecture to support storage of muti-omics data including sequencing data, MS-based proteomics data, MS or NMR-based metabolomics data, and fluorescence imaging data.

https://www.biosino.org/node/

BioStudies

-The BioStudies database allows users to explore datasets from genomic studies, deposited by a range of data providers. Access to datasets must be approved by the specified Data Access Committee (DAC).

https://www.ebi.ac.uk/biostudies/

The path from data acquisition to hypothesis generation: our workflow for mining open source data to study the protective role of VA in AAP

The challenge

Our study began with the purpose of identifying mechanisms underlying drug-induced pancreatitis (DIP). A prototypic example of a drug that can cause pancreatitis is asparaginase. Asparaginase is a cornerstone chemotherapy for the treatment of acute lymphoblastic leukemia (ALL), the most common cancer in the pediatric population.[13] AAP occurs in 2% to 10% of asparaginase users. One-third of affected patients develop severe acute pancreatitis (AP) with pseudocyst formation, pancreatic necrosis, and intensive care unit admissions.[14] AAP leads to discontinuation of asparaginase in over a third of patients, and half of the patients who are rechallenged with asparaginase after AAP experience pancreatitis recurrence. The mechanisms underlying AAP are largely unknown. This unmet need motivated us to adopt a systems approach to delineate the mechanisms toward the goal of characterizing interventions for AAP. We utilized open source transcriptomics data along with clinical data to complement well-conceived, metabolomics studies with clinical samples to generate the hypothesis that VA is a protective factor in AAP. We then tested the hypothesis using experimental ex vivo and in vivo pancreatitis disease models to delineate the underlying mechanisms. Our study is a classic example of how pancreas research can be enriched by the reuse of open source omics data (Fig. 1).

The clue that prompted us to search open source transcriptomics data for VA-regulated gene signatures during AP

We first performed a global metabolomic analysis of plasma samples from a discovery cohort of 50 children and adolescents with ALL enrolled in the Dana-Farber Cancer Institute (DFCI) ALL clinical trial protocols 05-001[15] and 11-001.[16] This analysis revealed remarkable differences in the changes of plasma metabolites between "case" individuals who developed AAP and "control" individuals who did not develop AAP. We discovered that, before and after a single exposure to asparaginase, the plasma levels of the pro-VA carotenoid, beta-cryptoxanthin, and the main circulating VA metabolite, retinol (ROH), were lower in the cases than in controls. The drop in VA levels was not due to malabsorption and was unique to VA, since the plasma metabolites for other vitamins from the screen remained unaffected. These data suggested a link between circulating VA levels and AP. While awaiting clinical samples from a validation cohort, we pivoted to a search for supporting, complementary data from open source omics datasets for our hypothesis and research enquiry into the protective role of VA in AAP.

Our approach to data retrieval and analysis

As a primer to our interrogation of the data repositories, we articulated specific questions that helped us construct a search query for effective data retrieval. The questions were: (1) Can we find publicly available datasets associated with AP? (2) Can we find alterations in VA-related gene expression signature during AP to

compare with our metabolomics data obtained from the discovery cohort? The rationale for the latter question was that VA is a potent regulator of anti-inflammatory and cell-protective effects through its intracellular metabolite retinoic acid (RA) and exerts its effects primarily by modulating gene transcription.[17–20] We searched OmicsDI, a central repository of well-annotated, standardized omics data (Table 2). We narrowed the search by genomics or transcriptomics and found a transcriptomics dataset deposited by Norberg et al[21] (NCBI GEO GSE109227). This dataset provided gene expression changes in the pancreas of mice during cerulein-induced AP (CIP), a well-characterized chemical form of experimental AP. We used GEO2R (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html) to compare the gene expression changes between cerulein-treated AP mice and vehicle-treated control mice. GEO2R is an interactive web tool that allows users to compare 2 or more groups of samples in a GEO series to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance and as a collection of graphic plots to help visualize differentially expressed genes (DEGs) and assess dataset quality (Fig. 2A). The analysis revealed significant alterations in the expression of vitamin A-related gene signatures in the pancreas during cerulein-induced AP. A parallel, independent enquiry of open source transcriptomics data (NCBI GEO GSE94289)[22] by collaborators, Drs Mayur Sarangdhar and Anil G. Jegga (Division of Biomedical Informatics, Cincinnati Children's Medical Center), yielded gene expression data from asparaginase-treated cells. To ascertain whether we can predict candidate therapeutics for AP from the gene expression data that we mined, we performed Connectivity Map (CMap) analysis.[23] CMap is a method to compare transcriptional expression data from a disease or phenotype (using DEGs or "gene signatures") with transcriptional data from in silico treatment with perturbagens, here a library of both U.S. Food & Drug Administration approved and investigational small molecules. We used the DEGs obtained from the CIP mouse model (GSE109227)[21] and Asparaginase-treated cells (GSE94289)[12] against the National Institute of Health (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 Touchstone reference dataset, a large database of small molecules transcriptomics data.[24] The perturbagen analysis on the DEGs identified among the top hits, multiple compounds belonging to the class of Retinoid Acid Receptor agonists (vitamin A and its natural and synthetic analogs) (Fig. 2B). Thus, the exercise in open source data yielded valuable, complementary evidence in support of our hypothesis that VA has a protective role in pancreatitis.[25]

Snapshot of additional studies to demonstrate the value of open source data reuse in enriching pancreas research

The case study of our experience in mining open-access pancreatitis omics data underscores how integrating such data can elevate pancreas research workflows and provide novel insights into pancreas disease mechanisms. Some additional studies encompassing diverse pancreas research themes that highlight how open source data exploration can lead to seminal and translationally relevant discoveries are summarized below.

Delineating pancreas biology and disease-related mechanisms

Utilizing human pancreatic adenocarcinoma clinical and gene expression data from TCGA, in conjunction with their single cell RNAseq (scRNAseq) data, the study by Peng et al[26] reinforced the existence of intra-tumoral heterogeneity, uncovered a connection between tumor-intrinsic transcriptional states and T cell activation and inferred the clinical value of the observed gene expression patterns in malignant ductal cells. A landmark study by Bailey et al[27] integrated genomic and gene expression analysis of human pancreatic ductal adenocarcinoma (PDAC) tissue and compared their data with other open access, PDAC gene expression datasets. This led to the discovery of four subtypes of pancreatic cancer and provided novel insights into PDAC pathophysiology. The data generated by these studies are publicly available, widely used by the pancreas research community and has fostered numerous, high-impact publications. Gopalan et al[28] mined multiple, publicly available pancreas (normal and chronic pancreatitis) and PDAC scRNAseq datasets (including the scRNAseq data published by Peng et al[26]) and identified the existence of a sub-population of non-malignant, healthy acinar cells referred to as acinar "edge" cells with a distinct transcriptional signature from "typical" acinar cells in the healthy pancreas. Interestingly, the acinar edge cells showed enrichment of transcriptomic signatures of pancreatic progenitors as well as PDAC and increased with age. This study highlighted the transcriptional heterogeneity among non-neoplastic acinar cells and provided a potential mechanism for increased PDAC risk with age.

Identifying pancreas disease signatures with translational potential

Yu et al[29] identified a novel diagnostic (d-) signature based on extracellular vesicle long RNA (exLR) profiling of patient plasma and integrating the exLR data with PDAC-related RNA expression data from TCGA. The combinatorial exLR signature derived from the use of both experimental and open source data was specific for PDAC and was superior to the marker CA19-9 in distinguishing PDAC from CP. The d-signature was also able to distinguish resectable PDAC from controls. A study led by Kinker et al[30] leveraged publicly available scRNAseq and TCGA data from human PDAC in conjunction with complementary experimental data to explore the spatial organization and cellular interactions in the PDAC immune environment. This work enabled a better understanding of the functional and clinical relevance of PDAC-associated tertiary lymphoid structures (TLS) and their potential application in guiding patient selection in immunotherapy trials.

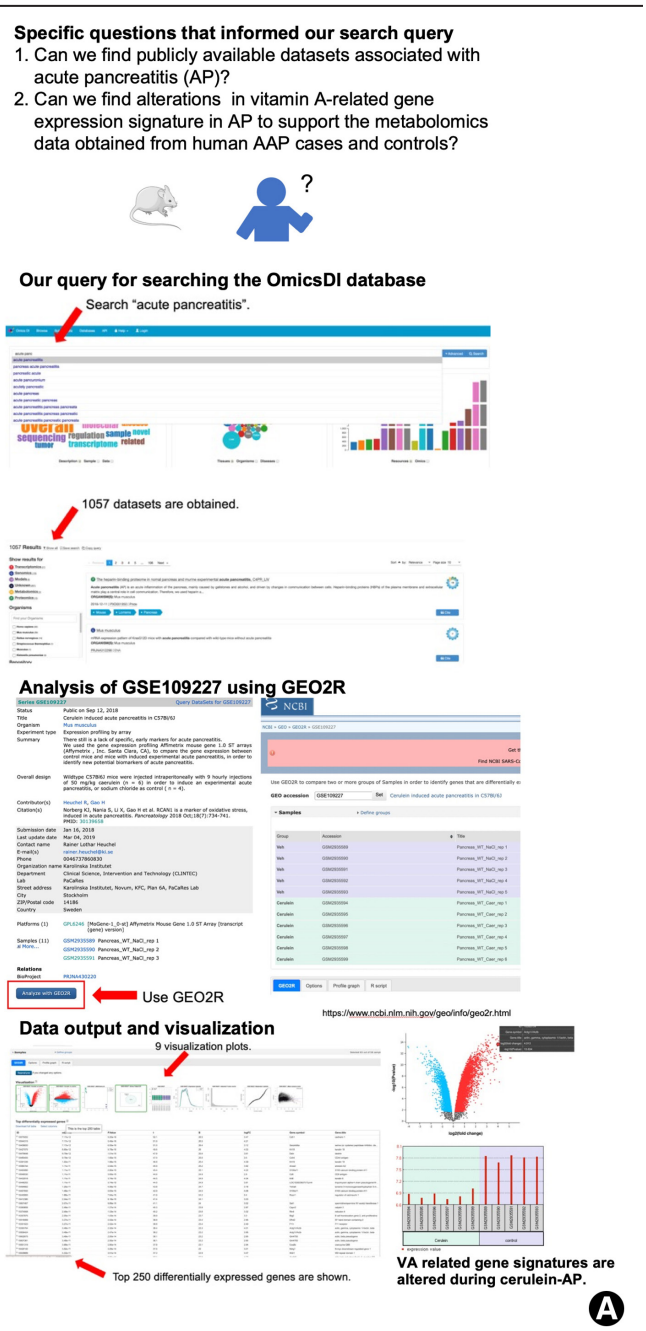
Thus, a burgeoning number of high-quality publications from the pancreas research community reveal the transformative potential of integrating data obtained from open source platforms to spearhead novel discoveries, enhance research workflows and reproducibility.

Limitations

The access and reuse of open source data are not without bottlenecks and limitations. While there has been a tremendous push to embrace the "FAIR" principles to provide "standardized" data that includes metadata and experimental details, some information necessary for data processing may be missing. This necessitates going back to the publication (where available) associated with the dataset or writing to the data depositor to obtain details about the study design, raw data, analytical and QA methods used for pre-processing the data prior to depositing.

Conclusion

The proliferation of omics-based approaches to address biomedical research questions and to accelerate scientific discovery has led a parallel evolution of open source data repositories and resources that enable free access and reuse of the myriad datasets. The FAIR principles encapsulate the salient features of good data management and stewardship with an emphasis on data standardization to support rigorous data repurposing.



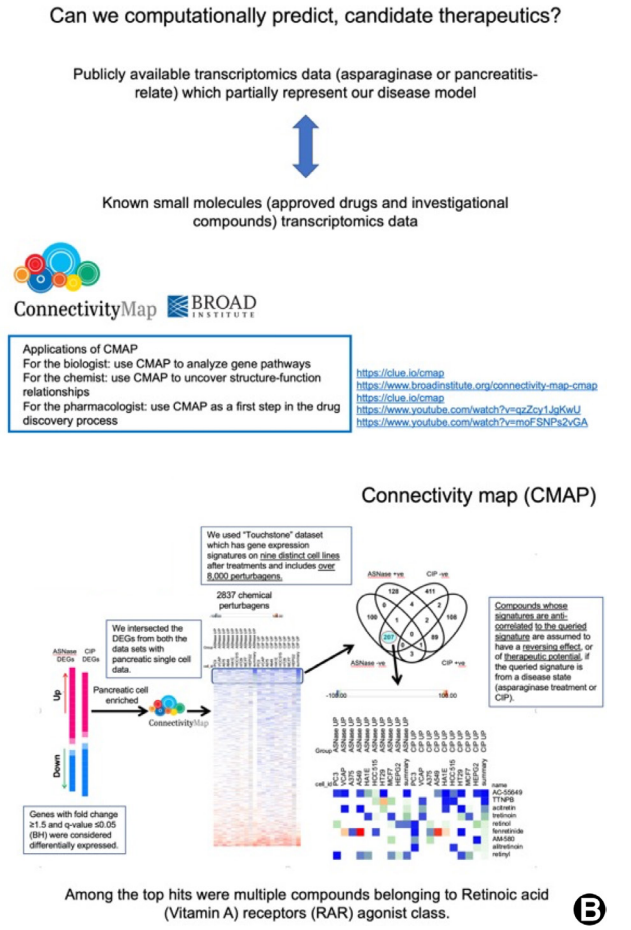


Figure 2. Open source data and analysis tools provided compelling evidence that led us to the hypothesis that VA has a protective role in pancreatitis. (A) Workflow that we adopted to interrogate OmicsDI data repository and data mining using the GEOR analysis tool. (B) Connectivity map analysis using transcriptomics data from cerulein-AP and asparaginase-treated cells predicted that VA (retinoids) can reverse gene signatures associated with AAP. AAP = asparaginase-associated pancreatitis, AP = acute pancreatitis, VA = vitamin A (*Continued*)

Numerous public data repositories are now available to access data for reuse in pancreas research to generate novel hypotheses and/or to complement findings from experimental models and clinical samples. Some of the data repositories host open source, user-friendly analytical tools that enable researchers with little or no computational experience to access and analyze the data to accelerate scientific discovery. However, the analysis of high-dimensional, complex datasets as well as harmonization of disparate omics datasets, requires an in-depth understanding about data science and the ability to utilize computational tools such as R-studio and Python to decode the data to answer specific research questions. This often requires close collaboration with computational biology/bioinformatics experts. There is a burgeoning need for the establishment and provision of a robust, training infrastructure (eg, workshops, courses) by universities, institutes, and professional societies to support the learning needs of biomedical researchers interested in repurposing omics and other high-dimensional data to enrich their research. Such initiatives may help educate researchers on how to navigate the open source data portals for data acquisition and analysis. This in turn, may maximize utilization of the treasure trove of open source data.

Acknowledgments

None.

Financial support

This work was supported by the Stanford Diabetes Research Center (no. P30DK116074) and mentored by SPARK Translational Research Program, Stanford University.

Conflicts of interest

The authors declare no conflicts of interest.

Ethics approval

Our review did not involve any clinical or animal experiments and was analyzed only using published open-source studies, therefore the need for ethical approval was waived according to the regulation of the Institutional Review Board.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

[2] Saunders DC, Messmer J, Kusmartseva I, et al. Pancreatlas: applying an adaptable framework to map the human pancreas in health and disease. Patterns (N Y). 2020;1:100120.

[3] Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–404.

[4] Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28:1248–1250.

[5] Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford). 2019;2019:baz046.

[6] Marzec J, Dayem Ullah AZ, Pirrò S, et al. The Pancreatic Expression Database: 2018 update. Nucleic Acids Res. 2018;46:D1107–D1110.

[7] Perez-Riverol Y, Bai M, da Veiga Leprevost F, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017;35:406–409.

[8] Perez-Riverol Y, Zorin A, Dass G, et al. Quantifying the impact of public omics data. Nat Commun. 2019;10:3512.

[9] Dass G, Vu M-T, Xu P, et al. The omics discovery REST interface. Nucleic Acids Res. 2020;48:W380–W384.

[10] Nowogrodzki A. Eleven tips for working with large data sets. Nature. 2020;577:439–440.

[11] Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–420.

[12] Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024 Feb;42(2):293–304. doi:10.1038/s41587-023-01767-y.

[13] Batool T, Makky EA, Jalal M, et al. A comprehensive review on L-asparaginase and its applications. Appl Biochem Biotechnol. 2016;178:900–923.

[14] Wolthers BO, Frandsen TL, Baruchel A, et al. Asparaginase-associated pancreatitis in childhood acute lymphoblastic leukaemia: an observational Ponte di Legno Toxicity Working Group study. Lancet Oncol. 2017;18:1238–1248.

[15] Place AE, Stevenson KE, Vrooman LM, et al. Intravenous pegylated asparaginase versus intramuscular native Escherichia coli L-asparaginase in newly diagnosed childhood acute lymphoblastic leukaemia (DFCI 05-001): a randomised, open-label phase 3 trial. Lancet Oncol. 2015;16:1677–1690.

[16] Vrooman LM, Blonquist TM, Stevenson KE, et al. Efficacy and toxicity of pegaspargase and calaspargase pegol in childhood acute lymphoblastic leukemia: results of DFCI 11-001. J Clin Oncol. 2021;39:3496–3505.

[17] Al Tanoury Z, Piskunov A, Rochette-Egly C. Vitamin A and retinoid signaling: genomic and nongenomic effects. J Lipid Res. 2013;54:1761–1775.

[18] Balmer JE, Blomhoff R. Gene expression regulation by retinoic acid. J Lipid Res. 2002;43:1773–1808.

[19] Blaner WS. Vitamin A signaling and homeostasis in obesity, diabetes, and metabolic disorders. Pharmacol Ther. 2019;197:153–178.

[20] Ghyselinck NB, Duester G. Retinoic acid signaling pathways. Development. 2019;146:dev167502.

[21] Norberg KJ, Nania S, Li X, et al. RCAN1 is a marker of oxidative stress, induced in acute pancreatitis. Pancreatology. 2018;18:734–741.

[22] Takahashi H, Inoue J, Sakaguchi K, et al. Autophagy is required for cell survival under L-asparaginase-induced metabolic stress in acute lymphoblastic leukemia cells. Oncogene. 2017;36:4267–4276.

[23] Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–1935.

[24] Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437–1452.e17.

[25] Tsai CY, Saito T, Sarangdhar M, et al. A systems approach points to a therapeutic role for retinoids in asparaginase-associated pancreatitis. Sci Transl Med. 2023;15:eabn2110.

[26] Peng J, Sun B-F, Chen C-Y, et al. Author Correction: single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019;29:777.

[27] Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531:47–52.

[28] Gopalan V, Singh A, Rashidi Mehrabadi F, et al. A transcriptionally distinct subpopulation of healthy acinar cells exhibit features of pancreatic progenitors and PDAC. Cancer Res. 2021;81:3958–3970.

[29] Yu S, Li Y, Liao Z, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut. 2020;69:540–550.

[30] Kinker GS, Vitiello GAF, Diniz AB, et al. Mature tertiary lymphoid structures are key niches of tumour-specific immune responses in pancreatic ductal adenocarcinomas. Gut. 2023;72:1927–1941.

How to cite this article: Swaminathan G, Saito T, Husain SZ. Exploiting open source omics data to advance pancreas research. J Pancreatol 2024;7:21–27. doi: 10.1097/JP9.0000000000000173