Pancreatlas: a comprehensive and interactive image atlas of the human pancreas in health and disease

Diane C. Saunders[1#], James Messmer[1#], Irina Kusmartseva[2], Maria L. Beery[2], Mingder Yang[2], Mark A. Atkinson[2], Alvin C. Powers[1,3,4], Jean-Philippe Cartailler[5*], and Marcela Brissova[1,6*]

[1] Division of Diabetes, Endocrinology, and Metabolism, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

[2] Department of Pathology, Immunology, and Laboratory Medicine, College of Medicine, University of Florida, Gainesville, FL

[3] VA Tennessee Valley Healthcare System, Nashville, TN

[4] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN

[5] Creative Data Solutions Shared Resource, Center for Stem Cell Biology, Vanderbilt University, Nashville, TN

[6] Lead Contact

* Corresponding authors

[#] D.C.S. and J.M. contributed equally to this work

Abstract

Deep phenotyping of human tissues generates complex spatial information from a variety of experimental modalities, yet images are typically condensed into small, static figures for publication and contain substantial data and metadata that never become widely available to the scientific community. Because image files are usually large, data is not easily shared or transferable even amongst collaborating researchers. While comprehensive image maps are available for some organs such as the brain, most resources are limited in their ability to accommodate multiplexed imaging with any degree of user interactivity. In recognition of this unmet need, we developed an online resource called Pancreatlas™ that integrates information technology infrastructure with enterprise imaging storage and visualization solutions. Through our custom interface, users can access curated, easy-to-navigate web pages, drill down to individual images, and deeply interact with them – all online, without lengthy downloads or software installation. Images are annotated with structured metadata, enabling users to dynamically build image datasets with biological and clinical relevance. The first version of Pancreatlas (v1.1) contains over 700 unique images acquired as whole-slide scans, confocal images, and imaging mass spectrometry, and is available at http://www.pancreatlas.org. While the overall system – entitled Flexible Framework for Integrating and Navigating Data (FFIND) – was deployed as a human pancreas-specific biological imaging resource, it can be configured to meet a myriad of imaging or other modular data management needs.

Background and Significance

Rapid advances in microscopy, live cell imaging, and multiplexing technologies are generating a wealth of rich and increasingly complex images, creating an enormous challenge to organize, process, and share this data in a way that facilitates meaningful scientific advancements. Traditionally, scientists observe image data as static images in publications or, when available online, in formats limited to only three color channels with little to no interactivity. Furthermore, comprehensive datasets encompass images generated by multiple imaging platforms and modalities, which are often acquired in proprietary formats and require individualized, proprietary image browsers that are difficult to integrate into a web environment. Therefore, there is an unmet need for more sophisticated image management and dissemination systems with capabilities to integrate data across different imaging platforms.

Although the demand for better image management solutions is apparent, imaging software is inherently difficult and expensive to develop. Moreover, those who stand to benefit most from imaging data and its interpretation are usually not software developers, but those investigating complex biology and disease. While the clinical imaging community has benefited from software and database solutions driven by advances in patient care, the basic science community still relies heavily on non-enterprise level software that is locally developed and chronically under-funded. How, then, can basic scientists make strides to share and/or access imaging data in a way that is "biologist-friendly," scalable, and that leverages existing technology?

Our research team tackled this challenge from the perspective of its requirement for an online resource to document human pancreatic architecture over the human lifespan. No reference datasets are available for human pancreas development, unlike other organ systems[1–5], and this knowledge gap is quite limiting for those working toward an understanding of human diabetes and pancreas diseases such as pancreatitis and pancreatic cancer. We assembled a multidisciplinary team of bio/informatics specialists, software developers, and biologists to build an "atlas" of the human pancreas. We chose to integrate existing tools and workflows wherever possible, layering multiple systems to meet project-specific needs and systematically cataloging the process. We prioritized the ability to handle images with more than 30 channels, allowing display of individual cell markers in user-specified combinations while still preserving spatial relationships within the context of the entire tissue section. The resulting platform (Flexible Framework for Integrating and Navigating Data, or FFIND) is designed to be used for any type of content, with our aptly named Pancreatlas™ being one implementation. Ultimately, we believe that FFIND is a model for integrating multimodal datasets in an accessible, easily digestible manner that will empower and accelerate biomedical research. Pancreatlas is available at http://www.pancreatlas.org.

Objective

The Pancreatlas platform was developed to organize and share full-resolution images, associated metadata, and de-identified clinical traits of organ donors from which the imaged

pancreatic tissue was derived. While the scope and nature of the data is specific to the interests of our research group, the goal was to create a database model that could be easily scaled, adapted, and applied to other data types or tissues of interest, providing a tool for basic science researchers to access reference datasets and advance understanding of complex diseases. The key features of the platform include:

1. Compatibility with commonly used image formats and flexibility to new formats
2. Association of images with technical, experimental, and clinical metadata
3. Inclusion of an interactive, multi-channel, and multi-resolution web-based image viewer
4. Integration of existing web technologies; engineering only new interfaces
5. Ability to run on institutional- or cloud-based infrastructure for portability and scalability

Materials and Methods

The technology platform FFIND (of which Pancreatlas is one implementation) was designed with three tiers: data, logic, and presentation (**Figure 1**). Tier 1, the data tier, is responsible for the ingest, storage, and retrieval of information originally stored in databases and filesystems. Specifically, this tier is composed of a mixture of servers, local and network-attached storage, cloud-based services, and storage components of Glencoe Software's OMERO Plus platform. Tier 2, the logic tier, coordinates all data communication between tiers 1 and 3 by processing commands, making logic decisions, and performing calculations. In our case, this includes the OMERO Plus platform components and application program interface (API), as well as a custom Python-based API, with both serving as interfaces to tier 3, the presentation tier. The presentation tier is the top-most level of the application and houses the user interface, whose main function is to translate user input via the logic tier in order to retrieve the requested data from the data tier. In our case, this is comprised of a custom Javascript/React-based web application, as well as the Glencoe Software PathViewer web application.

IT infrastructure

Our IT infrastructure includes (1) a large virtualized server (CentOS7) to host the OMERO Plus application; (2) several virtualized servers for custom application development and hosting (RedHat Enterprise 7); (3) on-site imaging repositories (Aperio eSlide Manager, Leica Biosystems; direct file high-performance storage for non-managed imaging data); (4) a cloud-hosted Laboratory Information Management System (LIMS) that contains tissue inventory and pancreas donor metadata; and (5) various web services used to connect and monitor overall architecture. All of our virtual machines have access to scalable resources, including memory, processing capacity, and storage. The bulk of our resources are co-located on the same 10Gb ethernet redundant network, and IT management is provided by several institutional support groups at Vanderbilt University and Vanderbilt University Medical Center.

Image management & viewing software

After assessing several commercial imaging applications, we selected Glencoe Software's OMERO Plus system for its outstanding capabilities in image management, viewing, and

selective sharing of data. The OMERO Plus platform was licensed, installed, and configured in collaboration with the Glencoe Software team, and the software's API was then leveraged to build a custom web interface.

OMERO Plus, which is based on open-source OMERO[6], is an enterprise-level system that leverages Bio-Formats[7] to access and integrate 150+ image formats from any modern imaging modality, such as light and electron microscopy and digital pathology. OMERO provides controlled access to imaging data and metadata, allowing it to function as both a project data management tool and an image data publication system[8]. It also includes PathViewer, a proprietary web browser application that allows users to visualize, annotate, and share images in a highly interactive fashion. PathViewer is the only enterprise-level software solution we identified that natively supports virtually a limitless number of image channels; all other products were restricted to 3-4 color channels. This feature of PathViewer was paramount for us, as analyses are increasingly moving toward multiplexed systems like imaging mass cytometry (IMC)[9] and co-detection by indexing (CODEX)[10]. The ease of toggling on and off individual channels – as well as being able to organize them within user-defined and similarly toggleable panels – made PathViewer the natural choice for the images we wanted to highlight in Pancreatlas. Beyond optimal image compatibility and viewing, OMERO Plus enables "annotation" or attachment of metadata attributes to images or sets of images, which can be utilized within a web framework to build datasets that have biological and clinical relevance (*e.g.*, those organized by donor age, tissue region, markers, etc.).

Custom web application

To maximize functionality for "non-technical" users like biologists, we utilized Python- and Javascript-based web frameworks (Django and React, respectively) to create Pancreatlas' custom web application. These frameworks, which are systems providing generic functionality that can be selectively changed by user-defined code, offer ready-made solutions to common web development idioms and design patterns, simplifying our ability to create intuitive interfaces for organizing, filtering, and viewing high-dimensional data. To achieve optimal usability, the web application seamlessly routes visitors from an image preview directly to PathViewer, bypassing the standard OMERO Plus organizational interface.

This user interface (UI), built using the React framework, defines various "components" for specific UI elements and allowed us to create reusable, modular code. Essential components of the application include the grid (default) view for displaying individual images, preview cards for displaying detailed information about a single image, and a filter panel that allows users to eliminate non-relevant images from view in real time. Other viewing modes within the application, such as the matrix view, are also composed of a hierarchy of components (*e.g.*, the number display of images matching a query, the table displaying each filter pair). In addition to the React framework, Pancreatlas uses several other existing Javascript libraries, each of which provides a set of focused functions our custom code can "call" for specific logic. These libraries include Reactstrap (which ports the popular Bootstrap framework to React), the tinycolor2 package (used to manipulate background colors), and the axios package (used to communicate with APIs).

Data imports

OMERO Plus provides a generic import workflow, which we tailored to simultaneously load our imaging data and metadata. This customized workflow allows data curators – those individuals with domain knowledge and expertise on the imaging data, including acquisition and biology – to work in a simple spreadsheet environment that lists images with their associated URL locations. Working by individual image or in sets of images, curators can annotate with additional details (metadata) based on predefined, controlled vocabularies, and can select clinical donor attributes obtained from our cloud-based LIMS.

Once curators (scientists) submit a data spreadsheet, the import mechanism within OMERO Plus validates parameters by ensuring metadata completeness and image accessibility, then proceeds to process images remotely and store newly created pyramidal image data within OMERO's filesystem. This process also populates several objects within OMERO's relational database and associates all of the imported annotations as object descriptors, key-value pairs, or tags (**Figure 2**). Importantly, this data import workflow is semi-automated, reproducible, and amenable to strong quality control steps.

Testing and user feedback

Testing of complex applications like Pancreatlas requires several workflows, including automated (Gitlab Continuous Integration) and functional testing, as well as local (Splunk) server monitoring and remote service monitoring (Uptime Robot).

To ensure maximal usability and address potential problems prior to launch, we shared pre-release versions of the Pancreatlas web application with small focus groups. This "beta testing" occurred at several regional and national conferences where we invited potential users to navigate a test environment and solicited written feedback. We found user-focused testing to be extremely valuable and worth the time and effort required to assemble focus groups. Interactions between database-savvy developers and somewhat tech-illiterate end users frequently uncovered points of confusion that required nominal effort to solve (from the development perspective) but exponentially enhanced usability (from the user perspective). These iterative conversations including both web developers and biologists will remain critical for the evolution of Pancreatlas going forward.

Results

Pancreatlas provides modular image groups with associated metadata, seamlessly delivering contextual information, enabling data filtering, and launching individual images in a state-of-the-art interactive image viewer (**Figure 3**). Image groups, termed *collections*, are built from a study topic or disease phenotype, giving users curated points of entry to the large quantity of images. In the default view of each collection, users can select and filter images by attributes of interest using various UI widgets, with dynamic retrieval of applicable image previews. The user can then enlarge each image to reveal relevant metadata (donor age, gender, markers visualized, etc.) before they launch the full image viewer. In parallel, a "matrix view" allows selection of two attributes (age, gender, pancreas region, etc.) displayed in rows and columns, with each

intersection populated by available images that meet the respective attribute values. These different viewing options, along with other key features of the web application, are summarized in **Table 1**.

Table 1. Key features of FFIND* and Pancreatlas

Feature	Description	Advantage/Significance	Figure(s)
Data organization			
Collections	Datasets; the *Collections* page lists all available datasets, and each collection also has a dedicated page describing relevant background and details	Provides digestible overview of data breadth; gives project-specific context	3A
Multiple viewing modes	• Grid view (default): data displayed as thumbnails with minimal descriptors (for easy browsing) • Temporal view: data grouped by curated age ranges (for data that is temporal in nature) • Matrix view: data grouped by user-selected attributes (*e.g.*, age and sex), displayed in a manner that easily identifies whether data exists for specific attribute combinations (*e.g.*, 6 years and male, 4 years and female, etc.)	Supports variable needs of user base; highlights key differences between data collections	2, 3B
Data filtering	Provides a flexible way to refine the data list to a subset of interest, allowing users to check boxes from lists of attribute options	Familiar interface (typical of eCommerce); intuitive to end users	2, 3B
Data presentation			
Thumbnails	Image associated with each datum (in the case of Pancreatlas, an image) that highlights key features	Provides visual identifier for each datum	3B, 4C
Preview mode	Enlarges data thumbnail and lists additional metadata (tissue, experimental attributes)	Offers basic attributes for a piece of data prior to launching an interactive browser	4D
Interactive browser (*e.g.*, PathViewer)	In the case of Pancreatlas, a web-based (HTML5) tool by OMERO Plus that boasts channel interactivity (toggle on/off, change colors, change range), supports 30+ channels for multiplexed imaging, and offers multi-resolution zoom (pre-processed pyramidal data)	Allows users to view images in a web browser, from anywhere, without downloading any data; multi-resolution zoom enables appreciation of cellular scale within whole organ context	3C-D
General browsing			
Bookmarking	"Save" buttons on data cards/previews let the user add data to a bookmarked collection which persists over multiple sessions	Enables quick reference; users can build and share custom lists via unique URL	4D
Nomenclature page	Lists all metadata terms, defining their biological relevance where applicable and providing diagrams to aid data interpretation	Encourages fieldwide adoption of metadata standards	4E

** Authors are working on making FFIND open-source; please visit the Git repository (http://gitlab.com/powersbrissova/FFIND) for updates.*

Image Collections

Virtualization provides the ability to increase resource allocations such as storage, CPU, and memory CPU as our needs evolve. Hence, its use within FFIND will allow us to organically scale Pancreatlas without complicated migrations due to hardware replacements, which will not impact our system. Summarized in **Table 2**, several collections have been made available in

version 1.1 of Pancreatlas. Each collection is meant to provide a context-specific list of images and their annotations. For example, this could be a set of images provided as supplementary data in a publication, or a set of images from multiple publications and/or unpublished work carefully selected to describe or illustrate biological processes or disease states. Other possibilities for collections include sets of images grouped by imaging modality (*e.g.*, images acquired using a new multiplexed platform), or various images from a single tissue sample that were analyzed by multiple collaborators.

One limitation of traditional scientific publications, even those available online, is the constrained space to present primary data. For example, in a recent study of cystic fibrosis-related diabetes (CFRD) by Hart and colleagues[11], 12 images were published in the main paper with another 7 in the supplement – however, the reported analyses utilized more than 80 images. This full image set has been made available via Pancreatlas as a disease-specific collection (**Table 2**, *CFRD*), and it offers the added advantage for users to navigate around large tissue areas, zoom in on regions of interest, and interact with the data in a way that is not possible with traditional publication formats. Another project that highlights the value of flexible spatial resolution (*i.e.*, viewing cells at high magnification but also retaining the large-scale tissue architecture) is the investigation of processes governing human pancreatic development. In an effort to gain insight into possible triggers of type 1 diabetes, a forthcoming study from our group closely examines islet composition and architecture from birth to ten years of age. Pancreatlas currently houses 281 images from this study (**Table 2**, *HANDEL-P*), many of which are whole-slide scans measuring up to 900 megapixels, or 30,000 pixels in both dimensions. Access to such high-resolution data is critical to appreciating the spatiotemporal context of pancreas and islet development; images provide detailed information of small islet structures (100-200 μm diameter) within the landscape of entire pancreatic cross-section (4-5 cm^2). As highlighted in **Table 1**, the dynamic interactivity facilitated by PathViewer is critical to understanding development on a whole-organ scale.

Table 2. Image collections available in Pancreatlas v1.1

Name, description, & URL	# of images	Image type(s)	Reference
Cystic Fibrosis-Related Diabetes (CFRD) Altered pancreatic architecture in the context of cystic fibrosis http://www.pancreatlas.org/datasets/410/overview	**86 total** (31 ND, 10 CF, 44 CFRD)	Whole-slide scans (H&E, IHC, FL IHC)	Hart et al. 2018[11]
Neonatal Development & Early Life Pancreas (HANDEL-P) Human endocrine pancreas architecture from birth to childhood http://www.pancreatlas.org/datasets/531/overview	**281 total** (274 ND, 7 AAB)	Whole-slide scans (FL IHC), confocal (FL IHC), CODEX	Manuscript in preparation
Human Pancreas Analysis Program (HPAP) Molecular profiling of the islet at various stages of T1D pathogenesis http://pancreatlas.org/datasets/508/overview	**303 total** (129 ND, 77 T1D, 28 T2D, 69 AAB)	Whole-slide scans (H&E), IMC	Wang et al. 2019[12], Kaestner et al. 2019[13]
Network for Pancreatic Organ Donors with Diabetes (nPOD) case #6362 Phenotyping of a recent-onset T1D donor http://pancreatlas.org/datasets/525/overview	**39 total** (all T1D)	Whole-slide scans (H&E, IHC), confocal (FL IHC)	Jackson et al. 2017[14], Canzano et al. 2018[15], Beery et al. 2019[16]

AAB, autoantibody positive; CODEX, co-detection by indexing; CF, cystic fibrosis; FL, fluorescence; H&E, hematoxylin & eosin; IHC, immunohistochemistry; IMC, imaging mass cytometry; ND, non-diabetic; T1D, type 1 diabetes; T2D, type 2 diabetes.

Discussion

We created the technology platform FFIND as a framework for modular data flow that can be easily tailored and applied to meet existing needs of the broad basic science community. FFIND offers solutions to researchers in the form of both data organization and presentation; data are grouped into collections that provide project-specific contextual details, and multiple viewing options cater to the needs of both the casual user and the subject expert. Data can be refined using a highly recognizable filtering paradigm, allowing users to choose attributes of interest with results returned in real time. Filtering occurs in a context where visual thumbnails and basic characteristics give a holistic sense of data before detailed examination, and preview panes highlight key features of each datum (**Figures 2** and **4**). FFIND's tiered architecture enables integration of existing data browsers with a customized web application that addresses project- and field-specific demands. Moreover, the modular data structure and filtering interfaces of the web application have been engineered to be reusable for datasets that will be added in the future and to easily designate sub-collections or browse across multiple collections at once. While our first iteration of FFIND, Pancreatlas, currently houses imaging data, its functionality can be easily extended to support a variety of data types. We hope that documenting our process from a primitive prototype to enterprise-level infrastructure will further the conversation amongst biologists (and funders) about the need for dedicated space and resources that make data more widely accessible to the research community.

The development of FFIND was stimulated by our group's desire to effectively share imaging data. Image databases are notoriously challenging and laborious to construct due to the large file sizes and the need to assimilate multiple imaging modalities and data formats. These obstacles not only prevent investigators from releasing full datasets upon publication, but they also limit accessibility of organ- or disease-specific reference datasets. Whereas innumerable solutions have been developed for sequencing datasets (*Nucleic Acids Research* publishes an annual Database Issue[17] and maintains an online database list that currently tops 1,800 entries), platforms for imaging data lag noticeably behind. The structural imaging community has addressed this shortage by building repositories with support from the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI); the Electron Microscopy Data Bank (EMDB)[18] and Electron Microscopy Public Image Archive (EMPIAR)[19] provide 3D reconstructions and raw 2D data, respectively, for protein structures obtained through cryo-electron microscopy. Excitingly, the first public 'added-value' bioimage databases have recently appeared: The Image Data Resource[20] and the Systems Science of Biological Dynamics Database[21] take submissions for reference datasets and provide substantial annotation and linkage to external resources. Pancreatlas joins these public resources as a bioimage database publishing reference datasets related to the biology and pathology of the pancreas. In the longer term, a common repository for all bioimage datasets related to published studies will be required – a point that was recently highlighted by Ellenberg and colleagues[6] in an editorial that called for development of a broad 'BioImage Archive' that could feed into more field- or application-specific databases like Pancreatlas.

The design and implementation of Pancreatlas required coordinated efforts in the IT environment, interaction with software companies that offer image management solutions, and

the dedicated effort of developers to select existing frameworks and libraries that best fit the unique nature of scientific imaging data. In building a user interface, we considered key biological features of the pancreas and structured our workflow to accommodate incorporation of images from multiple storage locations and in varied formats. The enterprise component of Pancreatlas infrastructure, OMERO Plus, was key to providing end users with interactivity, navigation, and annotation capabilities in a web environment. OMERO Plus' platform-agnostic browser, PathViewer, is particularly suited for both macro- (tissue-holistic) and micro- (cellular resolution) image viewing, and the ability to hide and view individual channels or groups of channels also make it an ideal environment to share multiplexed imaging data that is laborious to store, transfer, and display in static environments. Moreover, the latest release (PathViewer 3) accommodates side-by-side viewing of multiple images, which we are currently working on integrating into our platform so that pre-curated sets of images can be loaded as a "single click" from our front-end site. This will be crucial to illustrating age- or disease-specific phenotypes; for example, users could have the option to view a particular image with a matched "control", or to view the same marker combination at predefined age intervals to illustrate developmental changes. With an emphasis on adopting multiplexed imaging technologies, our group concluded that PathViewer is the most flexible, feature-rich browser available for image viewing and sharing in a meaningful biological context.

We are continuing to improve the FFIND technology platform that supports Pancreatlas, including IT infrastructure improvements, web application improvements (user interface consistency, device-agnostic usability, performance improvements, and error detection), as well as further integration with our laboratory information management system. While we have currently tailored our front-end site and user interface to suit the needs of researchers interested in the human pancreas, FFIND's structure and principles can be applied to any research area or data type.

We envision Pancreatlas to be an ideal repository for previously published image datasets, which have been vetted by peer review and often already contain relevant metadata fields. These datasets are sometimes uploaded as supplemental resources, but they are often hosted on the journal's own server and, thus, subject to file size restrictions. Researchers can elect to upload their raw data to an online open access repository like figshare or Mendeley Data; however, the structure of these databases is necessarily generic and requires further investment to produce a public-facing web resource. With Pancreatlas, image sets can be easily associated with a publication DOI, effectively driving traffic to specific manuscripts and/or journals and providing users the ability to interact with data without requiring journals to invest additional time or resources. We hope ongoing conversations with leadership in the scientific publishing community could result in the permission for authors to submit a Pancreatlas DOI with a forthcoming manuscript. One could envision that such a partnership would drastically increase the impact of a given manuscript and also promote cross-disciplinary discoveries.

Conclusion

Rapidly evolving technologies have far exceeded the limits of traditional data dissemination through printed and electronic publications. Imaging data presents a host of additional challenges: most image viewers available for website hosting are limited in color channel capability, user interactivity, and access to metadata, and they often do not support proprietary file formats. We built FFIND to specifically address image data management challenges, but its principles of interactivity, data organization, and sharing can be applied broadly across the spectrum of biomedical research topics. The creation of FFIND highlights the value of collaboration between bio/informatics specialists, software developers, and biologists to tackle important, interdisciplinary research questions.

FFIND and Pancreatlas are centered on the principles of automation, scalability, customization, and ease of use. By housing data in collections, we provide curated points of entry to vast amounts of data; our filtering menus allow users to view and refine data based on multiple variables of interest. Additionally, we have integrated metadata annotations to offer information about human samples and to encourage adoption of standardized, field-specific nomenclature. By connecting images from several independent programs and efforts, Pancreatlas has the potential to accelerate understanding of the human pancreas leading to transformative changes in diabetes care.

Remaining challenges for Pancreatlas, and for biomedical resource databases more generally, include identifying methods for long-term sustainability and funding. Even successful, established databases require sustained funding to maintain and expand infrastructure, curate data, and provide users with tools necessary for the ever-evolving analysis paradigms for large datasets[22]. Not only can it be difficult to convince funders to invest in resource building initially, but researchers must also demonstrate value to the broader research community in order to justify sustained funding. This can be a tricky but not impenetrable problem, as evidenced by usage statistics and thoughtful commentary provided by those who oversee existing resources[23,24]. With Pancreatlas, we already track usage by collecting voluntary information from users (*e.g.*, type of institution with which the user is associated) and will explore support mechanisms for similar long-lived resources like The Allen Brain Atlas[1–5,25,26].

Building high-value public resources for basic scientists requires continued conversation and collaboration among researchers, scientific journals, and biomedical research funders[6,27]. We view both FFIND and Pancreatlas as living platforms that will evolve with continued collaboration with and input from the broader community, and we look forward to partnering with other scientists to advocate for the value of data sharing through accessible and thoughtfully designed database solutions.

"Bigger Picture"

Basic scientists need cost-effective yet fully-featured database solutions that facilitate large dataset sharing in a structured and easily digestible manner. Flexible Framework for Integrating and Navigating Data (FFIND) is a data-agnostic, scalable web platform that integrates image and non-image data types (e.g., transcriptomics, flow cytometry) into a user interface tailored to project-specific goals. We used FFIND to build Pancreatlas, an online imaging resource containing datasets from publicly and privately funded consortia. In the future we plan to include 3D reconstructions of the pancreas, clinical imaging from pancreatic cancer specimens, and single cell transcriptomic data that links to histology from the same organ. We hope that Pancreatlas can simultaneously serve researchers interested in pancreas development, diabetes, pancreatitis, and pancreatic cancer by incorporating images from seminal publications and connecting to existing programs that link imagery with clinical data.

Author Contributions

Conceptualization, all authors; Methodology, D.C.S., J.M., M.B., and J.P.C.; Software, J.M. and J.P.C.; Resources, I.K., M.L.B., M.Y., and M.A.; Data Curation, D.C.S. and M.L.B.; Writing – Original Draft, D.C.S., J.M., M.B., and J.P.C.; Writing – Review & Editing, all authors; Supervision, M.A., A.C.P., M.B., and J.P.C.; Project Administration, I.K., M.L.B., and M.Y.; Funding Acquisition, M.A. and A.C.P.

Acknowledgements

We are extremely grateful to Jason Swedlow, Chris MacLeod, and Emil Rozbicki of Glencoe Software, who provided crucial technical and troubleshooting support and enthusiasm for our project. We also thank Dr. Swedlow for his critical reading of this manuscript.

Authors are working on making FFIND open-source; please visit the web-accessible Git repository (http://gitlab.com/powersbrissova/FFIND) for updates.

This work was performed using resources and/or funding provided by The Leona M. and Harry B. Helmsley Charitable Trust, by the NIDDK-supported Human Islet Research Network (RRID:SCR_014393; http://hirnetwork.org; UC4 DK104211, DK108120, DK112232, DK123716), by DK106755 and DK20593, and by the Department of Veterans Affairs (BX000666).

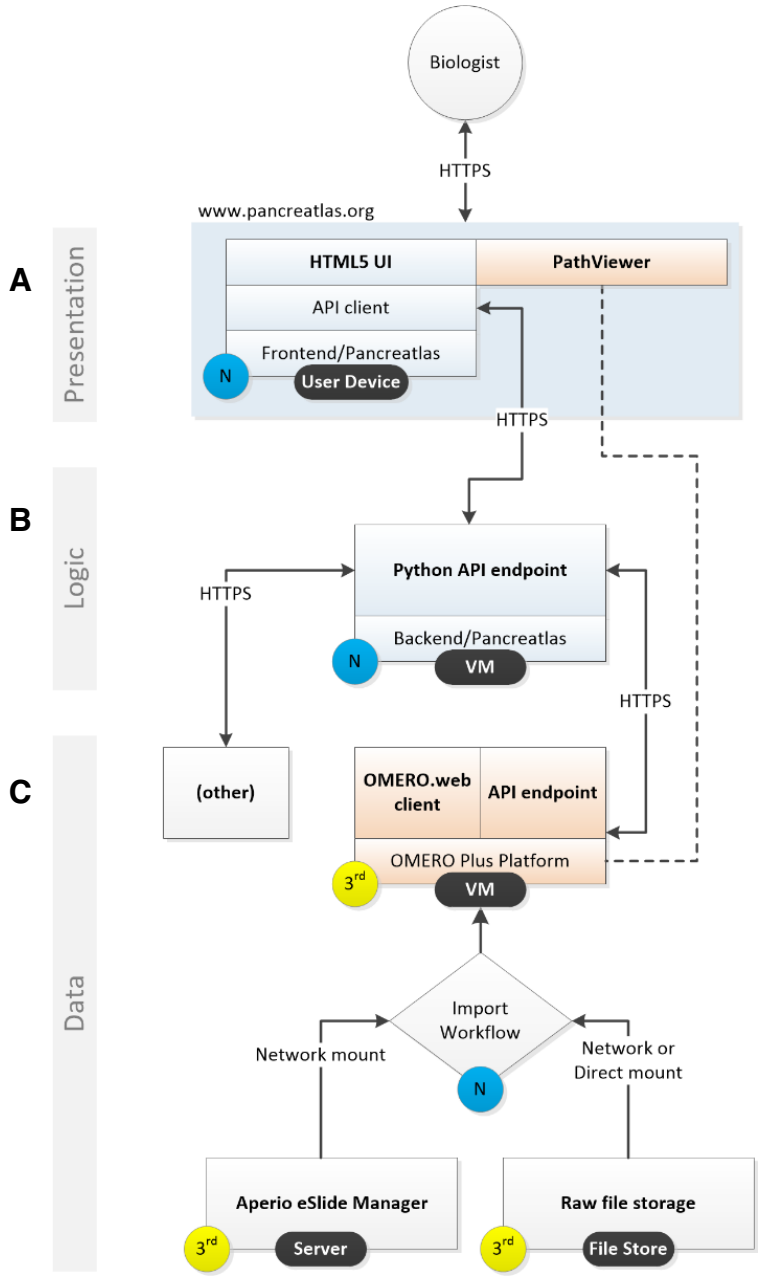

Figure 1. Schematic showing general connectivity and data flow across the three-tier Pancreatlas platform. Presentation in panel **A**, Logic in panel **B**, and Data in panel **C**. Yellow circles indicate third party tools; blue circles indicate our novel work integrating the three-tier platform infrastructure. **(A)** Separated from data- and logic-burdens, the presentation layer is entirely focused on providing a powerful and user-friendly interface to interact with the data using design idioms and structures commonly found in popular web applications and websites. **(B)** Our Python-based API, built for handling business logic, integrates data from multiple sources rather than only that of the OMERO API endpoint. **(C)** Image data and associated metadata can be loaded directly from a file store or proprietary imaging server (e.g. Aperio eSlide Manager, HALO Link). This information is processed and ingested into the OMERO database, which is hosted on a virtual machine (VM).

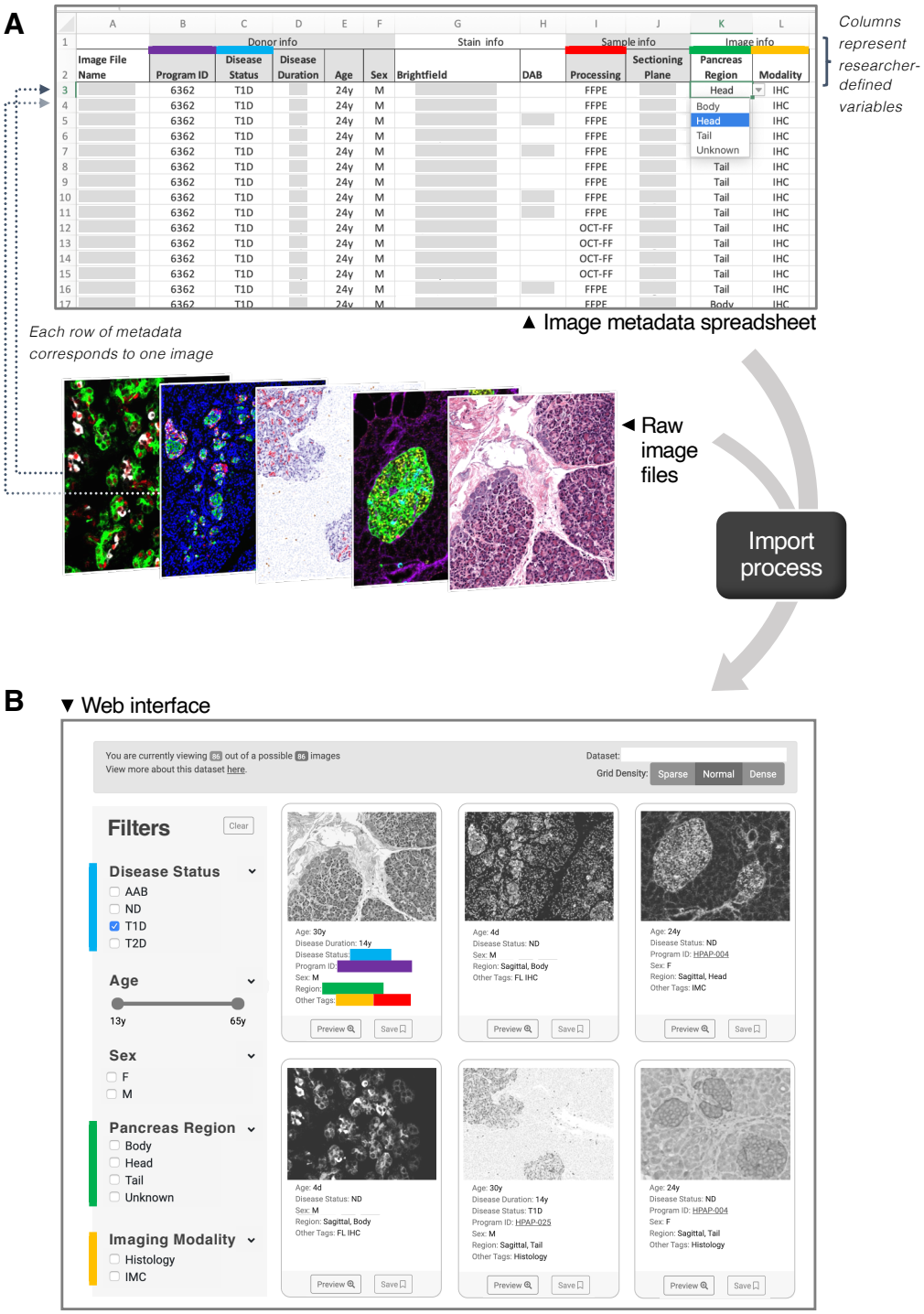

Figure 2. Spreadsheet-based metadata import process of FFIND. Data curators (scientists) input various clinical and experimental details (metadata) for each image file, creating a .csv file that will direct automatic import of both raw data (shown here: images) and metadata annotations **(A)**. The associated metadata is retrieved, along with data previews, in the custom web application, and is also used to create simple filtering options (shown in the left column of web application) **(B)**.

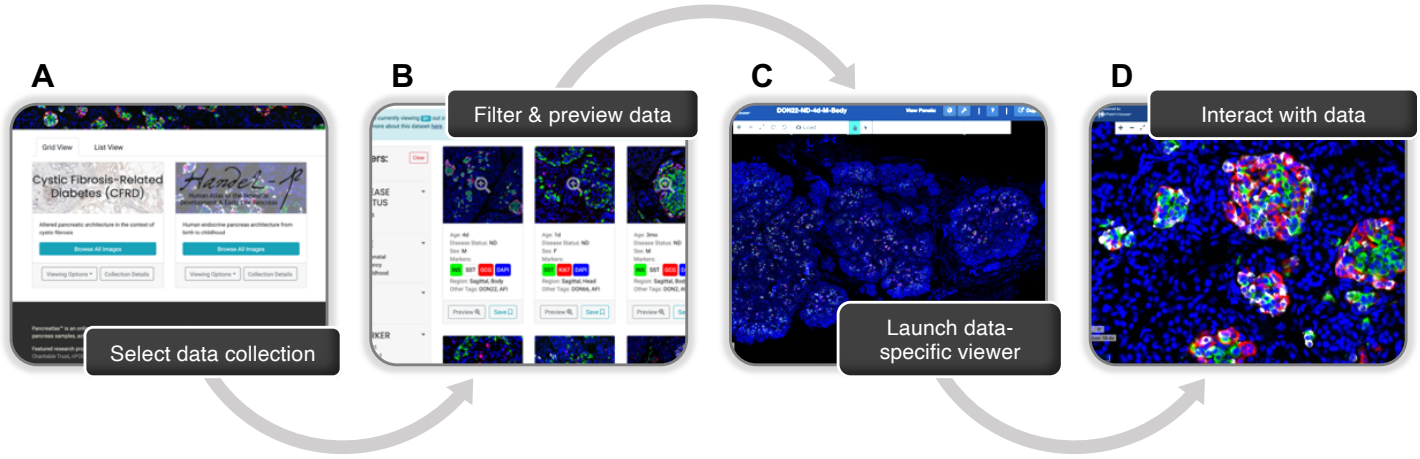

Figure 3. Stages of user-directed navigation in FFIND via custom user interface. Shown are **(A)** Collection page, **(B)** default grid view displaying data snapshots for a single collection, and **(C-D)** interactive web viewer; in the case of Pancreatlas, this is an image browser (PathViewer). For more detailed description of UIs, please see **Table 1**.

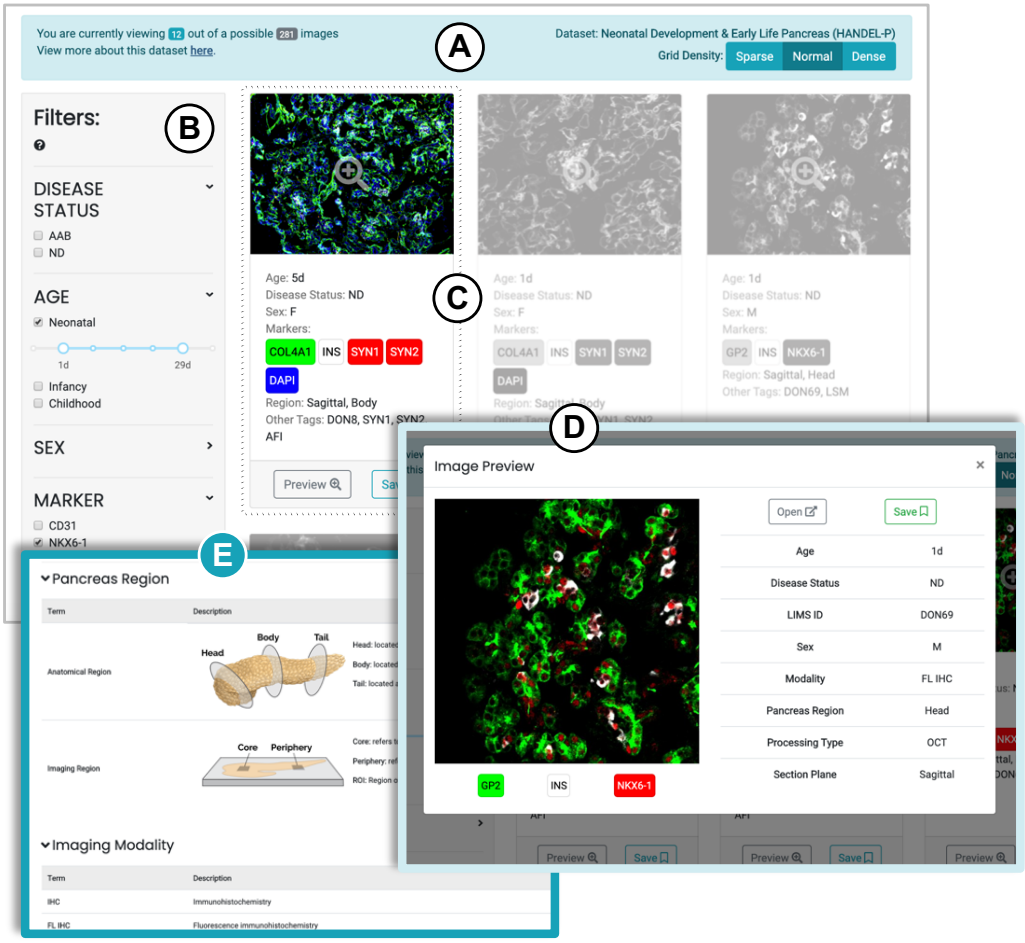

Figure 4. Image viewing within a selected collection. The default collection view features a clean grid with navigation information in the top toolbar **(A)**, a filtering menu on the left **(B)**, and small cards summarizing each datum **(C)**. Checking boxes in the filtering menu automatically repopulates cards only with data meeting the user-defined criteria. Clicking a data card launches a pop-up preview **(D)** that features an image thumbnail, experimental conditions, and other characteristics. From here, users can click on the thumbnail or Open button to launch PathViewer, or they can "save" the image to view later. Attribute terms in the filtering menu are defined on the Nomenclature page **(E)**, streamlining fieldwide metadata standards and providing useful diagrams to aid the user in interpreting images.

References

1. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).

2. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–9 (2012).

3. Ardini-Poleske, M. E. *et al.* LungMAP: The Molecular Atlas of Lung Development Program. *Am J Physiol-lung C* **313**, L733–L740 (2017).

4. McMahon, A. P. *et al.* GUDMAP: The Genitourinary Developmental Molecular Anatomy Project. *J Am Soc Nephrol* **19**, 667–671 (2008).

5. Harding, S. D. *et al.* The GUDMAP database--an online resource for genitourinary research. *Dev Camb Engl* **138**, 2845–53 (2011).

6. Ellenberg, J. *et al.* A call for public archives for biological image data. *Nat Methods* **15**, 849–854 (2018).

7. Linkert, M. *et al.* Metadata matters: access to image data in the real world. *J Cell Biology* **189**, 777–82 (2010).

8. Burel, J.-M. *et al.* Publishing and sharing multi-dimensional image data with OMERO. *Mamm Genome* **26**, 441–447 (2015).

9. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* **11**, 417–422 (2014).

10. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e15 (2018).

11. Hart, N. J. *et al.* Cystic fibrosis–related diabetes is caused by islet loss and inflammation. *Jci Insight* **3**, e98240 (2018).

12. Wang, Y. J. *et al.* Multiplexed In Situ Imaging Mass Cytometry Analysis of the Human Endocrine Pancreas and Immune System in Type 1 Diabetes. *Cell Metab* **29**, 769-783.e4 (2019).

13. Kaestner, K. H., Powers, A. C., Naji, A., Consortium, H. & Atkinson, M. A. NIH Initiative to Improve Understanding of the Pancreas, Islet, and Autoimmunity in Type 1 Diabetes: The Human Pancreas Analysis Program (HPAP). *Diabetes* **68**, 1394–1402 (2019).

14. Jackson, J., Posgai, A., Campbell-Thompson, M. & Kusmartseva, I. Insulitis in Autoantibody-Positive Pancreatic Donor With History of Gestational Diabetes Mellitus. *Diabetes Care* **40**, 723–725 (2017).

15. Canzano, J. S. *et al.* Islet Microvasculature Alterations With Loss of Beta-cells in Patients With Type 1 Diabetes. *J Histochem Cytochem* **67**, 41–52 (2018).

16. Beery, M. L., Jacobsen, L. M., Atkinson, M. A., Butler, A. E. & Campbell-Thompson, M. Islet amyloidosis in a child with type 1 diabetes. *Islets* **11**, 1–6 (2019).

17. Rigden, D. J. & Fernández, X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res* **48**, D1–D8 (2019).

18. Tagari, M., Newman, R., Chagoyen, M., Carazo, J.-M. & Henrick, K. New electron microscopy database and deposition system. *Trends Biochem Sci* **27**, 589 (2002).

19. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* **13**, 387–388 (2016).

20. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nat Methods* **14**, 775–781 (2017).

21. Tohsato, Y., Ho, K. H. L., Kyoda, K. & Onami, S. SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena. *Bioinformatics* **32**, btw417 (2016).

22. Martini, S., Nielsen, M., Peters, B. & Sette, A. The Immune Epitope Database and Analysis Resource Program 2003–2018: reflections and outlook. *Immunogenetics* **72**, 57–76 (2020).

23. Leng, J. *et al.* Securing the future of research computing in the biosciences. *Plos Comput Biol* **15**, e1006958 (2019).

24. Patwardhan, A. Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr Sect D Struct Biology* **73**, 503–508 (2017).

25. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

26. Uhlén, M. *et al.* A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Mol Cell Proteomics* **4**, 1920–1932 (2005).

27. Snyder, M. P. *et al.* The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).