ORIGINAL RESEARCH

LungMAP Portal Ecosystem

Systems-level Exploration of the Lung

Nathan Gaddis[1]*, Joshua Fortriede[3]*, Minzhe Guo[2]*, Eric E. Bardes[3], Michal Kouril[3], Scott Tabar[3], Kevin Burns[3], Maryanne E. Ardini-Poleske[1], Stephanie Loos[3], Daniel Schnell[3], Kang Jin[3], Balaji Iyer[3,4], Yina Du[2], Bing-Xing Huo[5], Anukana Bhattacharjee[3], Jeff Korte[5], Ruchi Munshi[5], Victoria Smith[5], Andrew Herbst[5], Joseph A. Kitzmiller[2], Geremy C. Clair[6], James P. Carson[7], Joshua Adkins[6], Edward E. Morrisey[8,9], Gloria S. Pryhuber[10], Ravi Misra[10], Jeffrey A. Whitsett[2,11], Xin Sun[12,13], Trevor Heathorn[14], Benedict Paten[14†], V. B. Surya Prasath[3,4,11], Yan Xu[2,11], Tim Tickle[5], Bruce J. Aronow[3,4,11], and Nathan Salomonis[3,4,11]; on behalf of the NHLBI LungMAP Consortium

[1]RTI International, Research Triangle Park, North Carolina; [2]Division of Pulmonary Biology, The Perinatal Institute, and [3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [4]Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, Ohio; [5]Data Sciences Platform, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts; [6]Biological Science Division, Pacific Northwest National Laboratory, Richland, Washington; [7]Texas Advanced Computing Center, University of Texas at Austin, Austin, Texas; [8]Department of Medicine and [9]Penn-CHOP Lung Biology Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; [10]Department of Pediatrics, University of Rochester Medical Center, Rochester, New York; [11]Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, Ohio; [12]Department of Pediatrics and [13]Department of Biological Sciences, University of California, San Diego, San Diego, California; and [14]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, California

ORCID IDs: 0000-0003-0478-8520 (G.C.C.); 0000-0002-9185-3994 (G.S.P.); 0000-0001-9689-2469 (N.S.).

Abstract

An improved understanding of the human lung necessitates advanced systems models informed by an ever-increasing repertoire of molecular omics, cellular imaging, and pathological datasets. To centralize and standardize information across broad lung research efforts, we expanded the LungMAP.net website into a new gateway portal. This portal connects a broad spectrum of research networks, bulk and single-cell multiomics data, and a diverse collection of image data that span mammalian lung development and disease. The data are standardized across species and technologies using harmonized data and metadata models that leverage recent advances, including those from the Human Cell Atlas, diverse ontologies, and the LungMAP CellCards initiative. To cultivate future discoveries, we have aggregated a diverse collection of single-cell atlases for multiple species (human, rhesus, and mouse) to enable consistent queries across technologies, cohorts, age, disease, and drug treatment. These atlases are provided as independent and integrated queryable datasets, with an emphasis on dynamic visualization, figure generation, reanalysis, cell-type curation, and automated reference-based classification of user-provided single-cell genomics datasets (Azimuth). As this resource grows, we intend to increase the breadth of available interactive interfaces, supported data types, data portals and datasets from LungMAP, and external research efforts.

Keywords: single-cell atlas; genomics; proteomics; LungMAP; human cell atlas

(Received in original form April 16, 2022; accepted in final form November 21, 2022)

*These authors contributed equally to this work.

†Data Advisor to California Institute for Regenerative Medicine.

Supported by the National Heart, Lung, and Blood Institute (U01HL122700 [G.S.P.], U01HL148856 [J.A.W. and Y.X.], U01HL148857 [E.E.M.], U01HL148860 [J.P.C., G.C.C., and J.A.], U01HL148861 [G.S.P.], U01HL148867 [X.S.], and U24HL148865 [T.T., B.P., B.J.A., and N.S.]) and the National Human Genome Research Institute (U41HG010972 [B.P.], R01HG010485 [B.P.] , U24HG010262 [B.P.], and U24HG011853 [B.P.]).

Author Contributions: Conception and design: N.G., J.F., M.G., E.E.B., M.K., S.T., K.B., M.E.A.-P., S.L., D.S., K.J., B.I., Y.D., B.-X.H., A.B., J.K., R.M., V.S., A.H., J.A.K., G.C.C., J.P.C., J.A., E.E.M., G.S.P., R.M., J.A.W., X.S., T.H., B.P., V.B.S.P., Y.X., T.T., B.J.A., and N.S. Drafting the manuscript for important intellectual content: N.G., J.F., J.A.W., G.S.P., R.M., V.B.S.P., Y.X., T.T., B.J.A., and N.S.

Correspondence and requests for reprints should be addressed to Nathan Salomonis, Ph.D., Cincinnati Children's Hospital Medical Center, 3333 Burnett Avenue, Cincinnati, Ohio 45229. E-mail: nathan.salomonis@cchmc.org.

This article has a related editorial.

This article has a data supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Cell Mol Biol Vol 70, Iss 2, pp 129–139, February 2024
Copyright © 2024 by the American Thoracic Society
Originally Published in Press as DOI: 10.1165/rcmb.2022-0165OC on November 22, 2022
Internet address: www.atsjournals.org

Clinical Relevance

While several dynamic lung molecular omics research portals have been described, these have remained largely uncoordinated efforts, often examining a single type of molecular entity. Here, we describe the LungMAP Portal Ecosystem, an interconnected group of interactive research resources that leverage collections of mammalian lung samples joined by a common metadata schema, spanning hundreds of donors and animals and a dozen datatypes. This effort aims to provide streamlined access to diverse molecular omics and imaging datasets, and detailed metadata for the mammalian lung to facilitate broad reuse and interactive analysis of lung datasets by any researcher.

The lung is among the most complex organs in the human body, with more than 40 cell types, each with specialized functions to support gas exchange and protect the lung against environmental challenges (1). Understanding the biological factors that govern lung health is complex and requires a multifaceted approach focused on clinical or pathological samples as well as informative animal models that allow for the rigorous validation of new hypotheses. The NHLBI (National Heart, Lung, and Blood Institute) LungMAP (Molecular Atlas of Lung Development Program) consortium was created to centralize the creation of standard reference maps that span regions of the lung and peripheral airway from mouse lifespan and human preterm viability through late childhood development (2). In its first phase, LungMAP produced a collection of diverse complementary molecular and imaging datasets for mouse and human samples and built a centralized repository of donor samples generously consented for research by family members of deceased infants, children, and young adults (BRINDL [Biorepository for Investigation of Neonatal Diseases of the Lung]) (3–5). Now in its second phase, a major focus of the LungMAP consortium is on the harmonization of datasets across lung location, developmental time points, ancestral diversity, and technology. In parallel with LungMAP, other significant lung research networks and individual laboratories are tackling many of these same questions with the goal of improving human lung health and patient survival. With the global concern and new challenges related to the coronavirus disease (COVID-19) pandemic, efforts have exponentially increased to catalog and describe the unique functions of innate and infiltrating cell types in response to pathogen infection (6). Importantly, many recent atlas-level studies have produced detailed single-cell genomics data for different lung cell lineages (e.g., endothelial, fibroblast, and immune), patient cohorts, and regions of the airway, providing exciting new opportunities to explore lung biology at a high resolution.

Although a number of dynamic lung research portals have been described, such as LungMAP.net, Lung Gene Expression Analysis (LGEA) (7), ToppCell (A Hierarchical Modular Single Cell Gene Expression Analysis System) (8), and the idiopathic pulmonary fibrosis Atlas (9), these have been largely uncoordinated resources (built using separate tools with independent metadata schemas and focused on distinct molecular omics and imaging technologies or specific regions or cell lineages in the lung). Hence, there is a strong need for a centralized, integrated resource that connects disparate lung and airway datasets, derived from distinct laboratories, species, and disease conditions and provides standardized metadata, data, and online interactive interfaces. Such integrated resources are critical and enable the research community to discover preliminary evidence for research proposals, validate findings from animal models, and identify critical gaps in current knowledge. Such challenges are nontrivial, given the hundreds of single-cell genomics samples that have been profiled, corresponding to millions of cells, with corresponding imaging, proteomic (bulk and single-cell), lipidomic, and epigenomic data produced on distinct but complementary technologies.

Here, we describe the LungMAP portal ecosystem, an interconnected group of interactive research resources that leverage collections of mammalian lung samples joined by a common metadata schema. This effort aims to provide streamlined access to diverse molecular omics and imaging datasets and detailed metadata for the mammalian lung. This effort has been designed to be synergistic with sister consortia, in particular, HCA (Human Cell Atlas) lung network and NIH's Human BioMolecular Atlas Program (HuBMAP) lung research initiative. In its current iteration, eight independent lung research portals have been integrated with LungMAP, leveraging a centralized tissue repository: BRINDL (https://brindl.urmc.rochester.edu). LungMAP.net, the gateway to this ecosystem, hosts both the LungMAP network and lung community datasets (standalone and integrated) for deeper exploration and advanced interactive visualization. To ensure data are findable, accessible, interoperable, and reusable, LungMAP has adopted a core set of community schemas to enhance and enrich its associated metadata, file formats, cell-type descriptions, and protocols and enable integration and reanalysis datasets in the cloud. We intend for this web ecosystem to enable user journeys with diverse pathways beginning with either specific genes or proteins, cell types, anatomical regions, dataset collections, diseases, or developmental time points, to discover affected molecular and cellular programs, and to enable orthogonal validation in independent cohorts in diverse species.

Some of the results of these studies have been previously reported in the form of a preprint (bioRxiv, 6 December 2021, https://www.biorxiv.org/content/10.1101/2021.12.05.471312v1).

Resource Description

The Portal Ecosystem

The LungMAP ecosystem was developed as an interconnected set of research resources that provide both unique and overlapping functions (Figures 1 and 2). LungMAP.net is the newly redesigned gateway portal in this ecosystem, providing access to centrally developed and affiliate portals (Table 1). This gateway portal has been redesigned to provide intuitive and streamlined access to diverse components of this ecosystem. At the hub of this network is a common-metadata schema model to curate and describe datasets, both consortium and community-generated. This ontology-aware model is interoperable with the HCA metadata schema to allow datasets to be analyzed independently or in an integrated manner, to align results across developmental age, technologies, clinical demographics, and protocols. To ensure that deposited data are

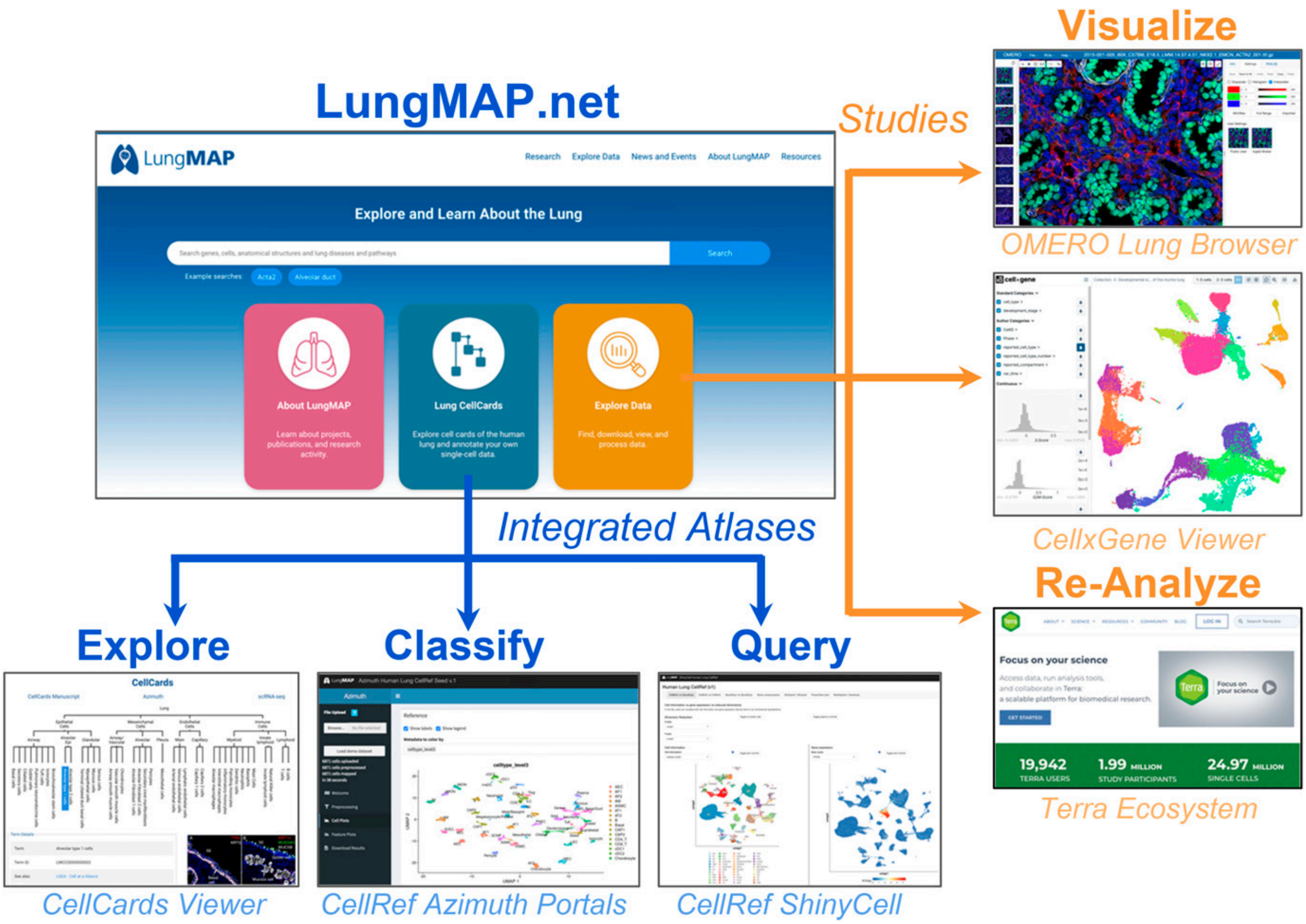


Figure 1. The LungMAP gateway portal. Example portals and applications are available from the LungMAP gateway portal and accessory downstream data portals. These are separated into standard approved reference atlas versions from the LungMAP consortium and individual data collections (studies, samples, and community datasets) available for exploration and interactive visualization or reanalysis in the cloud using tools in the Terra ecosystem. OMERO = open microscopy environment.

accessible and compliant with the originating data consents, raw data are available in both an open and a managed manner through alignment with the NHLBI BioDataCatalyst initiative.

To enable agile access to diverse datasets and ontologies, data can be queried directly through the LungMAP.net portal or through a dedicated Omics data browser. Metadata at LungMAP.net is served from a graph-relational database that allows for diverse technologies, annotations, and assay types to be interrelated on the basis of molecular features (i.e., genes and antibodies), cell types, or anatomical regions. Interoperable LungMAP and HCA lung network single-cell genomics data (https://data.humancellatlas.org) are additionally provided in a separate dedicated data browser portal that leverages the HCA data coordination platform (DCP) faceted search approaches, making available a well-documented application programming interface (for computational access) and a graphical user interface (Table 1). To reuse omics data, this initiative leverages the database of Genotypes and Phenotypes (dbGAP), BiodataCatalyst, and Terra to access and leverage open and managed access data. Terra provides a cloud-native analysis environment, enabling horizontally scalable reprocessing and analysis supporting initial quantification, quality control, summarization, integration, and visualization. Similarly, a dedicated lung image portal has been developed to provide the capability to not only query but reanalyze, curate, and combine images across multiple technologies. Finally, a series of dynamic apps for visualization and community analyses are provided, allowing lung researchers to produce customized views of community lung datasets and annotate their own datasets on the basis of LungMAP aggregate references (Table 2).

Community Contribution

Beyond LungMAP consortia-produced genomics datasets, the LungMAP Data Coordination Center (DCC) collaborates with diverse initiatives and individual investigators to incorporate community datasets as a sharable resource within our NIH genomic standards-compliant system, thus ensuring donor Protected Health Information (PHI) privacy. This effort includes *1*) submission by lung researchers into the LungMAP ingest broker; *2*) integration of datasets into LungMAP.net by staff curators; and *3*) incorporation of public atlas community lung datasets (Figure 2B). This flexible structure allows for the integration and display of datasets in a consistent manner. Lung researchers can

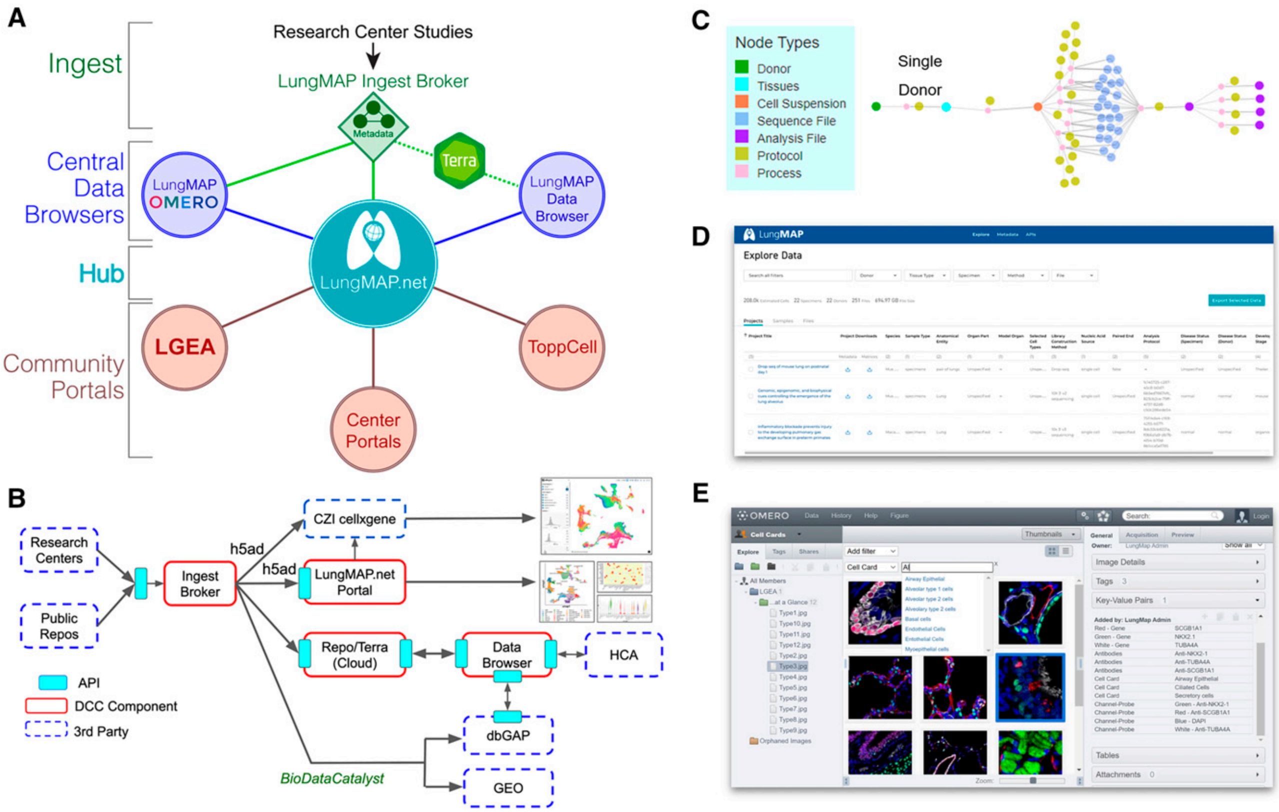


Figure 2. Advanced omics and image analysis through the LungMAP portal ecosystem. (*A*) Overview of the LungMAP.net ecosystem, highlighting LungMAP central data browsers for image analysis with the OMERO, dataset browsing through the LungMAP.net graph database or Lung Data Browser Azure database, and connected community portals (Terra, LGEA, ToppCell, ToppGene, and research center portals). (*B*) Data flow of LungMAP and community data through the centralized LungMAP ingest broker. Standardized metadata is submitted via LungMAP-compatible APIs for deposition in Terra, LungMAP.net, and the LungMAP data browser, as well as controlled-access repositories (dbGAP and BioDataCatalyst) and third-party technologies/portals (ShinyCell, CZI cellxgene, and Azimuth) using LungMAP extended data formats (e.g., h5ad). (*C*) Standardized metadata graph (on the basis of the HCA metadata schema) for an example biological sample, describing the sample, omics evaluation, protocols, study, and connectivity (process) between entities in the graph. (*D*) LungMAP data portal to quickly search, download, and reanalyze data (Terra) from LungMAP deposited experiments. (*E*) LungMAP OMERO image analysis portal to query, explore, combine, reprocess, and export imaging data from diverse imaging modalities. API = application programming interface; CZI = Chan Zuckerberg Initiative; dbGAP = database of Genotypes and Phenotypes; DCC = Data Coordination Center; GEO = Gene Expression Omnibus; HCA = Human Cell Atlas; LGEA = Lung Gene Expression Analysis.

contact LungMAP curators directly through LungMAP.net to begin submitting and curating datasets within the ingest broker using the LungMAP metadata schema.

LungMAP-supported Data Modalities and Metadata

Diverse complementary omics, imaging metadata, and quantitative data types are provided for interactive exploration in LungMAP.net, all explorable through a common interface (Visualize Data). In addition to bulk and single-cell RNA sequencing (RNA-Seq), proteomics, lipidomics, DNA methylation, microRNA, and metabolomics measurements are provided for diverse samples, including those matched for the same samples (multiomics) in LungMAP.net (Figure 3). The bulk assays are derived from laser capture microscopy and sorted populations to enable comparative analyses between distinct molecular programs; these targeted samples are complemented with similar representative assays on whole lung tissue. To facilitate reuse and reanalysis, mouse and human omics data are provided as downloadable results (raw and processed data) and a growing number of interactive assays (interactive heatmaps and HighChart graphs).

LungMAP.net hosts adjacent imaging sections at multiple image resolutions for direct interrogation to understand the spatial and cellular localization of proteins, mRNA transcripts, and metabolites in the lung. Multiple imaging technologies are provided, including immunofluorescence confocal, histological stains, *in situ* hybridization (ISH) (10), computed tomography (CT) and micro-CT, vibratome-assisted subsurface imaging microscopy, and nanospray desorption electrospray ionization (11), in addition to recently produced mass spectrometry-based imaging data such as the matrix-assisted laser desorption/ionization-mass spectrometry at both two and three-dimensional resolutions. This imaging data resource (LungMAP OMERO [Open Microscopy Environment] portal, http://omero.lungmap.net) will continue to be updated with diverse imaging technologies and samples.

The LungMAP OMERO portal provides tools for image and data analysis,

Table 1. Lung Cancer Master Protocol Portals Ecosystem Components

Portal	Interactive Single-cell	Image Exploration	Curation and Ontologies	Raw Data Download	Data Reprocessing	Epignomics Exploration
LungMAP central portal	X	X	X	—	—	—
Omics Data Browser	X	—	X	X	X	—
CellCards Azimuth	X	X	X	—	—	—
OMERO image browser	—	X	X	X	X	—
CellCards browser	X	—	X	—	—	—
LGEA	X	X	X	—	—	X
ToppCell	X	—	X	—	—	—
Terra	X	—	X	X	X	—

Definition of abbreviations: LGEA = Lung Gene Expression Analysis; OMERO = Open Microscopy Environment.

including API endpoints for programmatic- and pipeline-based analyses. Advanced search capabilities within OMERO are aided by the rich metadata curated from each image. Plugins such as omero.figure allow these source images to be used to build publication-quality, multipanel images with annotations, labels, and scale bars. Channel-level intensities can also be adjusted to allow for the best view. These omics and imaging datasets are organized to a centralized metadata schema with associated rich protocols directly embedded in the website or linked to protocols.io. The current version of LungMAP OMERO hosts 1,755 immunofluorescent images from diverse sources, laboratories, and species for in-depth analyses, with a growing number of human images (Table 3). This centralized organization of raw data and metadata allows users to perform downstream analyses and apply artificial intelligence techniques on imaging and multiomics data. This metadata model is extensible to enable extension and linkage to emerging metadata standards and file formats, such as HuBMAPs common coordinate framework json files (https://hubmapconsortium.github.io/ccf/) and ILI volumetric CT scan information, which are needed for three-dimensional visualization of source biopsy location (12).

Interactive Single-cell Genomics Analysis

Many invaluable single-cell atlases have been produced over the last several years, spanning health and diverse lung diseases, including COVID-19, chronic obstructive pulmonary disease, pulmonary fibrosis, and neonatal/pediatric death. To explore discrete transcriptomic differences in such atlases and from a single website, LungMAP.net provides both harmonized and independent views of diverse human, mouse, and nonhuman primate single-cell atlases. Individual dataset views are presented as technology-specific study-level pages. These pages include dynamic visualization of individual genes and gene sets according to study-specific covariates, such as age, disease, developmental stage, and drug treatment. These interactive viewers include dynamic Uniform Manifold Approximation and Projection (UMAP) visualization via the Chan Zuckerberg Initiative (CZI)-developed cellxgene (13, 14) and R ShinyCell apps (13), heatmaps using Morpheus browser technology, and frequency bar charts for precomputed cell type and/or covariate signatures (Figure 4A). When the same dataset is available for interactive exploration in LungMAP affiliate or external portals, these datasets are provided as links with graphical previews (e.g., LGEA and ToppCell). Such interactive views are enabled through the creation of standardized gene expression matrix files (h5ad) with metadata related to the sample, cell, and study (*see* data supplement). For samples derived from LungMAP studies, donor IDs linked in each study can be further queried across all LungMAP experiments to find related orthogonal omics datasets (bulk RNA-Seq, lipidomics, and proteomics) or imaging datasets to find data from the same donor (Figure 4).

For many users, using single-cell datasets to explore cell identity, disease, and/or developmental differences can be daunting. To this end, the LungMAP web portal includes a series of tutorials that walk through specific frequent-use cases (https://lungmap.net/tutorials/). Generally, these use cases fall into *1*) hypothesis-focused research questions (i.e., the role of a specific molecule in a specific cell population in disease vs. healthy); *2*) broad exploratory research questions (i.e., developmentally regulated markers within a cell lineage); and *3*) dataset supervised classification. For example, one can interactively explore a mouse developmental single-cell atlas for maximally differentially expressed genes between two different developmental stages (e.g., early embryonic vs. adult club cells) using the cellxgene viewer (15). Such differences can be further quantified and visualized in LungMAP.net using the heatmap, bar chart, or covariate line plots viewers with gene sets immediately exported to ToppGene for comprehensive enrichment analyses (16). Subsequently, the user has the option to reprocess the source sequence data from the LungMAP data browser via Terra using open-access single-cell analysis workflows in the cloud (e.g., Optimus and Cumulus) or locally on their system to extend these findings. Multiple similar user journeys are possible for different species, developmental, and disease study designs (*see* data supplement).

Lung CellCards

Single-cell genomics enables the identification of distinct cell populations and regulatory programs from transcriptomic, epigenomic, and proteomic measurements. Over a dozen atlas-level single-cell datasets of the lung now exist, spanning healthy states, development, and disease (15). However, the precise identity of cell populations still remains largely speculative, as identities will vary on the basis of which annotation and cell clustering methods are used. Although curation initiatives, such as Cell Ontology, aim to standardize the description of cell types within all organ systems (17), there is a significant need to curate cell types from an organ-specific perspective, considering the well-established literature and new experimental predictions. The LungMAP CellCards initiative was developed to respond to this need by developing a

Table 2. Interactive Web Technologies

Portal	CellxGene	Azimuth	ShinyCell	Morpheus.js	HighCharts	Gene-set Enrichment	Ontology Browser	Genome Browser	Cell–Cell Interaction	Terra Reprocessing	OMERO.iviewer
LungMAP Central Portal	X	X	X	X	X	X	X	—	—	—	—
Omics Data Browser	X	—	X	X	X	—		—	—	X	—
LGEA	—	—	—	—	X	—	X	X	X	—	—
ToppCell/ToppGene	—	—	—	X	—	X	—	—	X	—	—
OMERO Image Browser	—	—	—	—	—	—	—	—	—	—	X

centralized resource for current and future lung curation efforts (1). Cell types and anatomical terms in CellCards that are not represented in the cell ontology or anatomy ontology, respectively, are being submitted to these resources. The CellCards resource at LungMAP.net is browsable through a cell lineage tree to provide functional, location, developmental, experimental, regenerative, and disease associations together with corresponding cell surface and molecular markers for each curated lung cell type (Figure 5). In this web interface, such markers dynamically link to hundreds of available imaging datasets from mice and humans throughout their lifespans (Figure 3).

An important application of CellCards is to guide the annotation of user-provided single-cell RNA-Seq datasets. Hence, we integrated data from diverse single cell RNA sequencing (scRNA-Seq) studies and identified cell populations associated with discrete CellCards markers, in which there is a one-to-one association between each CellCard and each lung cell type. Version 1.0 beta of this lung reference, which we call LungMAP CellRef, integrates 505,256 cells from 104 donors from eight published and two previously unpublished LungMAP cohorts, spanning four regions of the lung. This integrated dataset consists of nondiseased adult and pediatric lung scRNA-Seq captures (different 10× Genomics library chemistries) to provide a comprehensive adolescent/adult reference cell atlas. The mouse CellRef includes 40 CellCards-focused consensus populations derived from 95,658 cells from 17 samples, spanning eight developmental time points (E16.5 to postnatal Day 28) using Drop-seq. Optimal Leiden cluster definitions were supervised on the basis of their specificity for CellCards literature-defined markers, corresponding to 48 distinct cell populations, after batch effects correction across cohorts, donors, and regions of the airway (Guo and colleagues in review). These reference maps are provided as multiple interactive single-cell browsers (LungMAP.net and LGEA) and as an R shiny application (Azimuth) (18) to enable any users to map their own single-cell RNA-Seq data to these references. To use this functionality, users can directly upload preprocessed scRNA-Seq datasets (counts-level) into the CellRef human or mouse Azimuth interfaces, perform basic quality control analyses to exclude outlier cells, project cell population labels from these atlases onto their own or downloaded datasets, compare with prior assigned cluster annotations, and iteratively query/visualize their data for individual genes (*see* data supplement). Compatible input files are provided for a growing number of LungMAP scRNA-Seq datasets (e.g., h5ad) in addition to direct links to these files deposited in other public repositories, such as Gene Expression Omnibus (GEO), through the LungMAP visualize data interface (Figures 3A and 3B). Such methods allow computational and noncomputational biologists to quickly curate and explore their own datasets in just a few minutes.

LungMAP Community Portals

LungMAP connects an ecosystem of community portals developed by the LungMAP consortium and community partners. These portals include the popular Lung Gene Expression iN Single-cell (LungGens), ToppGene, Terra, and the recently developed Omics data browser and ToppCell portals, which provide distinct and complementary functionality for shared and unique lung datasets, as well as the capability to interactively reanalyze diverse genomics datasets. For each portal, data and metadata are interconnected through LungMAP.net.

LungMAP omics data browser. The Omics data browser (https://data-browser.lungmap.net) provides a faceted search capability over genomic datasets that have been ingested with conformance to metadata standards. The Omics data browser metadata schema is closely aligned with the Human Cell Atlas schema and provides a rich set of metadata properties with which to describe biological entities and processes. Examples of searchable facets include donor, tissue type, specimen disease, sequencing method, and file type. The Omics data browser is implemented as a highly scalable cloud-based indexing system and web service; it indexes key metadata entities of broad interest, providing a very fast search capability that allows users to filter data of interest over a large range of datasets, generating custom cohorts for further analysis. Selected data may be downloaded to a local system; alternatively, the data may remain in the cloud and be processed *in situ* by user-selected workflows within the Terra system. A RESTful programmatic API provides computational users with full search-and-retrieve capabilities.

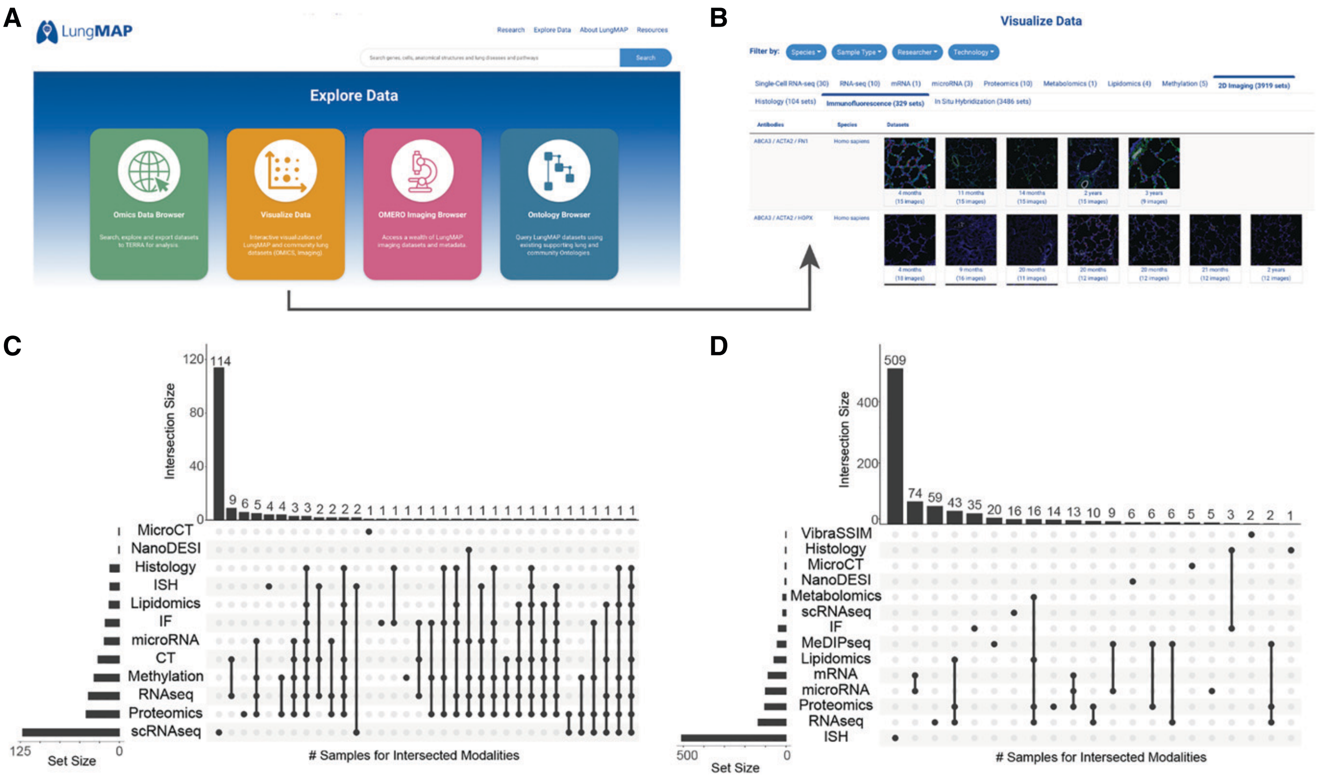


Figure 3. Shared samples across diverse LungMAP molecular and imaging assays. (*A*) LungMAP.net explore data menu items. (*B*) Visualize data menu, highlighting immunofluorescence image exploration. (*C* and *D*) Upset plot of the intersection of data types available for (*C*) human donors and (*D*) mice at LungMAP.net. The *x*-axis is the graphical intersection of the shared samples across the indicated data types. Many samples often exist for a single donor or sample (e.g., ISH). The set size bar chart provides the number of human donors or mice with each type of data. The intersection size bar chart provides the number of human donors or mice with the intersection of types indicated by dots in the matrix beneath the chart. CT = computed tomography; IF = immunofluorescence; ISH = *in situ* hybridization; MeDIPseq = Methylated DNA immunoprecipitation sequencing; MicroCT = Microcomputed tomography; NanoDESI = Nanospray desorption electrospray ionization; RNAseq = RNA sequencing; scRNAseq = single-cell RNA sequencing; VibraSSIM = Vibratome serial section imaging microscopy.

Terra. For datasets available in the Omics data browser as well as research community datasets, advanced analyses of the raw data can be facilitated through the Terra cloud environment. Terra (19) is a data platform codeveloped by the Broad Institute, Microsoft, and Verily that combines cloud-native services and access to scientific datasets into an unprecedented resource. This environment is open-source, modular, and community-driven, enabling fast computational analyses of linked datasets in the cloud (Figure 6). Beyond LungMAP, Terra serves the needs of thousands of scientists worldwide and requires no infrastructure on the part of the scientist. This environment was intentionally designed to be open and promote collaboration. Terra is a critical component of several data centers supporting atlas building, including the HCA DCP, the BCDC (Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network [BICCN]'s Brain Cell Data Center), LungMAP2 DCC and the Single Cell Opioid Responses in the Context of HIV (SCORCH) Data Center. Beyond atlas building projects, Terra is a key component to AnVIL (NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space), All of Us, and National Cancer Institute (NCI) Cloud Resources programs. Terra enables the standardized reanalysis of datasets from diverse research consortia, together with private datasets produced from individual laboratories.

Championing Global Alliance for Genomics and Health (GA4GH) standards, Terra is a part of a larger federated community driving many interoperability standards. Using these standards, Terra enables access to datasets from many scientific projects (e.g., BICCN, HCA, AnVIL, 1000 Genomes, BioData Catalyst, ENCODE, TARGET, TCGA, and TOPMed). These datasets can include open access as well as managed access datasets. Scientists are able to upload their own data to Terra (setting the data private or sharing it with others) and complement their data with

Table 3. OMERO Image Inventory

	Human				
Age	Black	Other	White	Mouse	*n*
20 mo	91	—	—	—	91
3 yr	—	—	53	—	53
4 mo	—	102	—	—	102
4 yr	115	—	—	—	115
9 mo	75	—	—	—	75
E16.5	—	—	—	47	47
E18.5	—	—	—	96	96
P01	—	—	—	204	204
P03	—	—	—	199	199
P07	—	—	—	135	135
P10	—	—	—	244	244
P14	—	—	—	235	235
P28	—	—	—	159	159
n	281	102	53	1319	1755

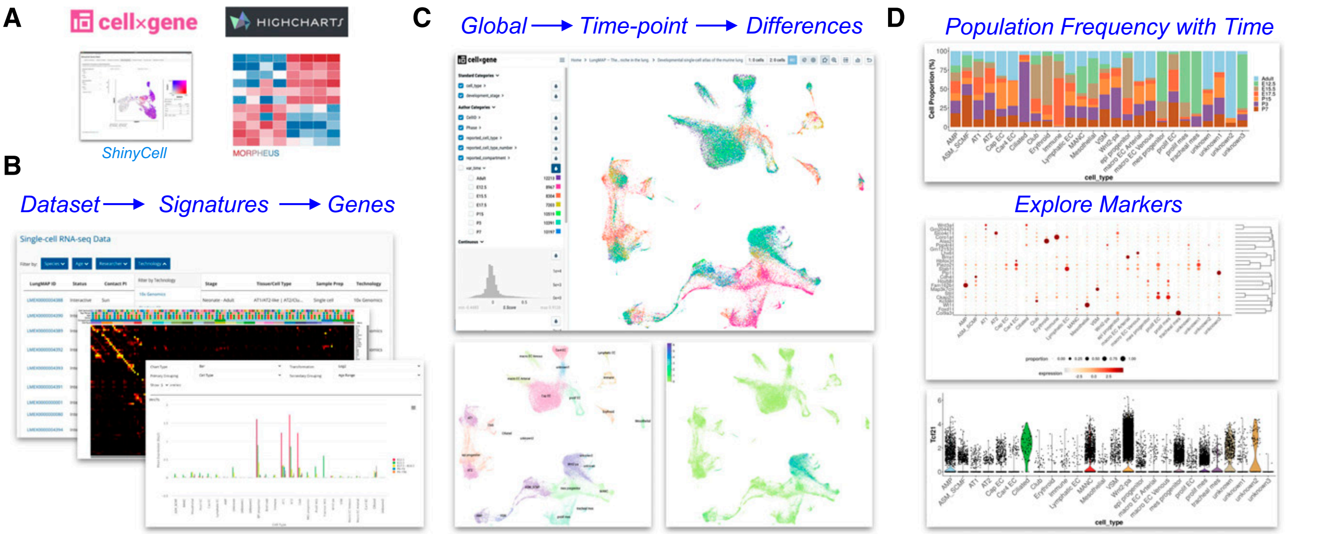
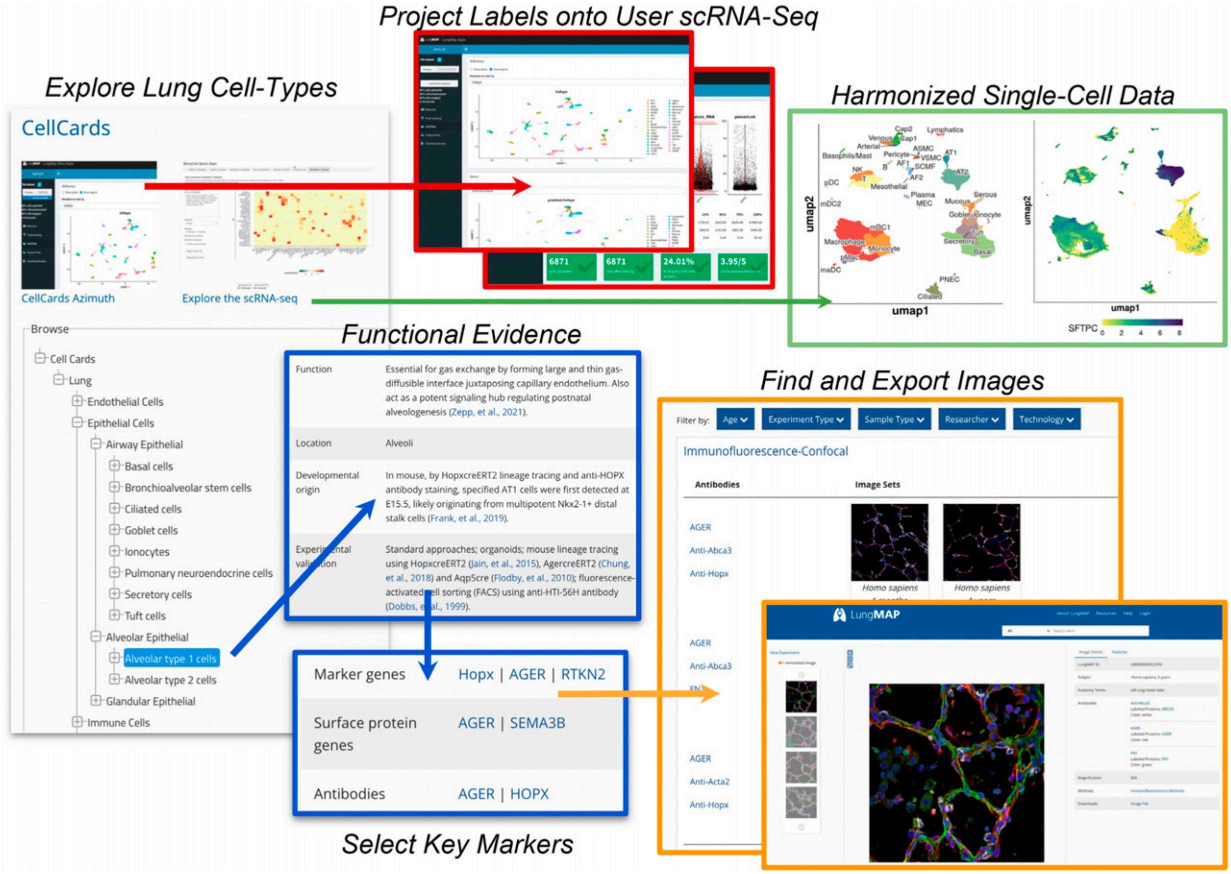


Figure 4. Comprehensive visualization of single-cell lung atlases. (*A*) Opensource computational tools for genomics analysis used in LungMAP.net. (*B*–*D*) Single-cell atlas of seven time points of mouse development (16). (*B*) Dataset browsing within the LungMAP.net portal to select datasets by technology, species, or sample age (bottom); Morpheus heatmap browser of available time point and cell-type signatures (middle); HighCharts visualization of gene expression across developmental age. (*C*) CZI cellxgene dynamic Uniform Manifold Approximation and Projection (UMAP) visualization of integrated developmental time points (top), cell types (bottom left), and gene expression for prior defined candidates or cellxgene time-point differential expression-identified genes. (*D*) R ShinyCell browser plots for cell-type frequency at different developmental time points (top), proportion bubble plot for selected genes (middle), and violin plot of expression for a selected gene (bottom).

existing public and shared datasets. Terra orchestrates scalable resources, including scalable computing for batch processing workflows and interactive environments (e.g., Jupyter Notebooks, RStudio, RShiny apps, and Galaxy [20]). These resources are leveraged with configurable resources (e.g., CPU/GPUs, memory, and Spark backends). Workflows can be added by scientists and shared or made public for wider use. Dockstore integration enables workflow import into Terra, and together both Dockstore and Terra make available approximately 3,000 public workflows, including GATK Best Practices (21), Cumulus (22), and workflows used by the HCA DCP and the BICCN. Terra supports the Single Cell Portal (currently hosting 1,105 public and private studies spanning 35 million cells), which is built on top of the Terra infrastructure and offers both user-friendly interactive visualization capabilities through the portal user interface and scalable analysis capabilities through Terra. For interactive data analysis, computational biologists can create and share analyses as notebooks, R Studio scripts, Galaxy tools, or RShiny apps in environments natively supporting R, C++, Python, or custom containerized programming environments. These resources are complemented by a user support team that provides documentation, operates a community forum and help desk, and creates targeted tutorials, showcases, and workshops. Terra maintains the highest degrees of security as a Federal Information Security Modernization Act (FISMA) moderate authorized and a Federal Risk and Authorization Management Program (FedRAMP) moderate impact authorized environment with the authority to operate from several NIH Institutes. Tutorials on how genomics datasets can be imported and analyzed with Terra from the LungMAP Omics data browser are provided in the data supplement and the LungMAP website.

LungGENS. Data in the LGEA portal is referenced in LungMAP.net for matching studies, ontologies, and samples. Unique features of this environment include access to unique sample sets and disease cohorts and additional interactive visualization interfaces and workflows that enable deep exploration of transcriptomics, proteomic, epigenetic, and systems-level impacts in healthy and diseased lungs. The current LGEA web portal provides access to a diverse number of query and analytic tools, including LungGENS, LungSortedCells, LungDTC (lung developmental time course), LungDiseases, LungEpigenetics, LungImage, LungProteomics, LungOntology, LGEA-Project, and LGEA-ToolBox. The newly released feature toolset "lung-at-a-glance" contains four interactive components: region at a glance, cell at a glance, gene at a glance, and lung single-cell reference (7, 23, 24). Queries on this website can be performed from the perspective of individual genes to view their expression globally (i.e., UMAP plots), regionally (i.e., bar charts), or sort for the top markers per cell population or browse by cell types. This website is organized in a storybook manner to enable simplified user journeys. For example, lung at a glance provides an easy-to-use web toolset for knowledge interrogation and visualization of complex omics and imaging data, providing an interactive web interface to bridge lung anatomic ontology classifications to lung structure, histology, immunofluorescence confocal images, and cell-type–specific gene expression. For multiomics analysis, LungProteomics facilitates integrative analyses of transcriptomic and proteomic expression profiling for the major cell types in human and mouse lungs. Protein query retrieves expression information and provides comparisons of protein and mRNA pairs in conjunction with associated statistical summaries, enabling users to quantitatively examine the consistency or discordancy of protein–mRNA pairs and the selectivity of queried proteins for each cell type. The cell type query retrieves signature genes from



Function	Essential for gas exchange by forming large and thin gas-diffusible interface juxtaposing capillary endothelium. Also act as a potent signaling hub regulating postnatal alveologenesis (Zepp, et al., 2021).
Location	Alveoli
Developmental origin	In mouse, by HopxcreERT2 lineage tracing and anti-HOPX antibody staining, specified AT1 cells were first detected at E15.5, likely originating from multipotent Nkx2-1+ distal stalk cells (Frank, et al., 2019).
Experimental validation	Standard approaches; organoids; mouse lineage tracing using HopxcreERT2 (Jain, et al., 2015), AgercreERT2 (Chung, et al., 2018) and Aqp5cre (Flodby, et al., 2010); fluorescence-activated cell sorting (FACS) using anti-HTI-56H antibody (Dobbs, et al., 1999).

Marker genes	Hopx | AGER | RTKN2
Surface protein genes	AGER | SEMA3B
Antibodies	AGER | HOPX

Figure 5. CellCards: A curated database of lung cell types. The Lung CellCards browser is shown (left) with displayed curated cell types, definitions, literature, markers, and link-outs to LungMAP.net (right). Link outs include all associated LungMAP database entities (gene and protein information, molecular omics datasets, and images). Interactive query of the integrated CellCards single-cell atlas (top) for supervised annotation and exploration of user-supplied single-cell RNAseq datasets (Azimuth, middle) and exploration of the integrated reference datasets (ShinyCell, right).

each cell type and displays signature proteins and mRNAs using an interactive heatmap and provides cell-level correlations on the basis of the expression of coherent mRNA/protein signatures (25).

ToppCell. ToppCell (http://toppcell.cchmc.org) was launched in 2018 as a biologist-oriented web portal for conducting differential expression tests and downstream analysis for single-cell RNA-Seq data with complex metadata. ToppCell is a component of the ToppGene ecosystem of websites, with a shared set of gene signatures, annotations, and backend software infrastructure. In November of 2021, there were more than 70 datasets publicly available on the ToppCell web portal, including BrainMap, Cardiovascular Atlas, ImmuneMap, LungMap, and other projects, each of which consists of various single-cell datasets for a specific cell lineage, tissue, or organ. Twenty-two mouse and human lung scRNA-Seq datasets are currently available at (https://toppcell.cchmc.org/biosystems/go/index3/LungMap), ranging from normal lung tissues (26, 27) to lungs with disease, such as pulmonary fibrosis (28), tumors (29), and virus infections (30).

This portal enables the extensive exploration of cell-type module signatures in the community and, in the near future, user-uploaded datasets for automated subclustering, module identification, and network analysis. Each dataset has a user-friendly interface with hundreds of gene modules and sets of differentially expressed genes (200 by default) arranged in a hierarchical manner according to user-defined cell metadata. For example, in human fetal lung single-cell data (27), age and compartment information were used as sample-specific metadata for cells, whereas lineage and cell type were used as cluster-specific information. Both kinds of information were used to generate differentially expressed genes in gene modules, allowing direct comparisons of cells in different ages or regions. In addition, with a seamless link between ToppCell, ToppGene, and ToppCluster, users can enrich an individual gene module or multiple customized modules, which can be used for functional comparisons or protein–protein analysis.

Discussion and Future Development

As lung research efforts incorporate progressively greater multimodal, spatial, three-dimensional, and temporal resolution data, a continuing challenge and emphasis of the LungMAP portal ecosystem will be to produce increasingly more comprehensive data views for both individual datasets as well as aggregate compendiums. Central to this work is the deployment of automated

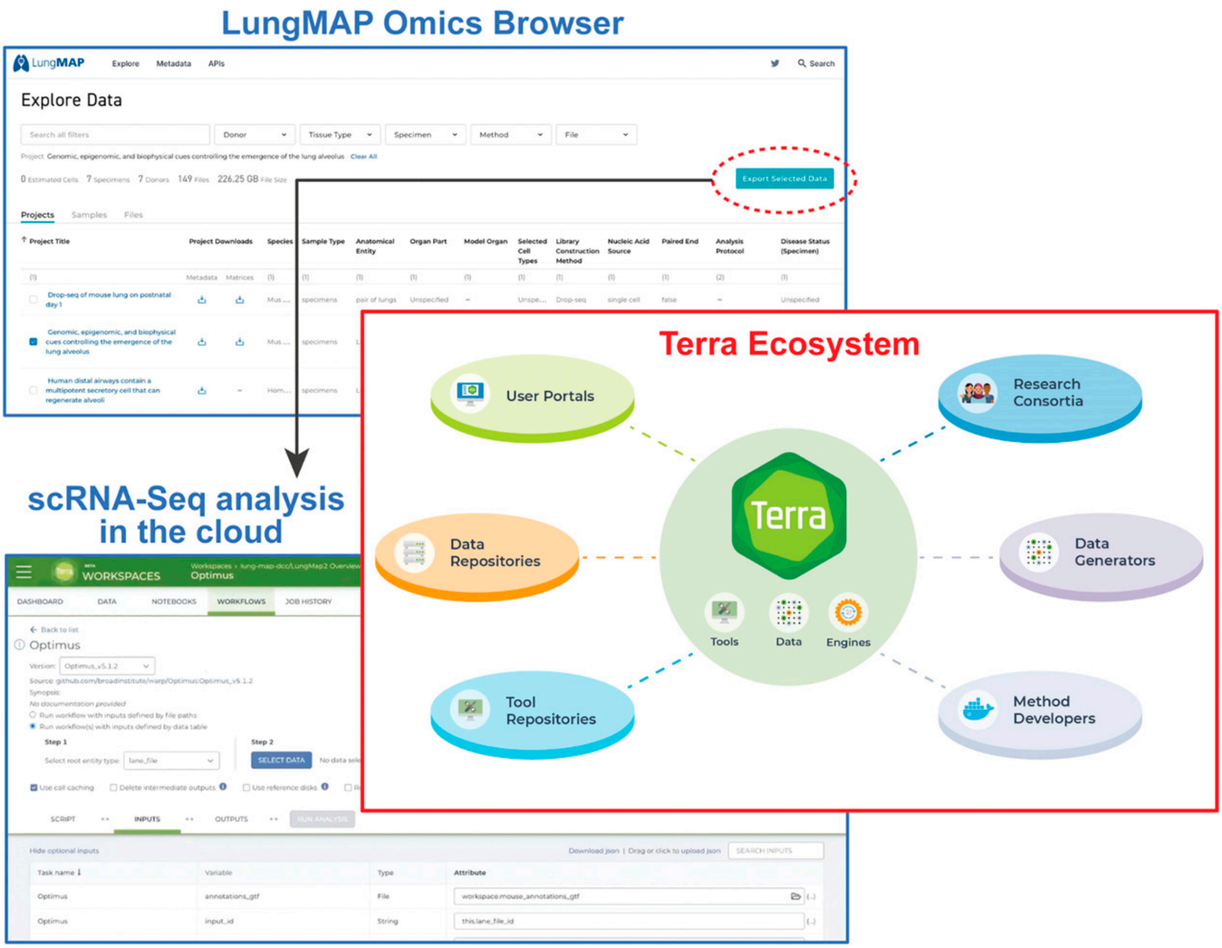


Figure 6. Enabling scalable cloud computing using Terra. The LungMAP Omics browser allows LungMAP raw and processed genomics datasets to be downloaded or analyzed further in the cloud. This includes reanalysis and integration of scRNAseq datasets, including those in BioData catalyst. Terra combines cloud, storage, and computing in a secure environment to create a scalable infrastructure that spans institutions and connects scientific projects. Enabling researchers to bring tools and data to leverage cloud computing resources, Terra connects scientists to a federated ecosystem of portals, data repositories, and tool repositories using Global Alliance for Genomics and Health (GA4GH) standards. Terra is used by research consortia to perform joint analysis, data generators to manage and share data, and methodologies to create, disseminate, and leverage at-scale pipelines and analyses.

analytical pipelines to produce standardized datasets that span cohorts and platform versions, support new technologies, advanced analytical methods, and the engagement of new community partners with independent portals. A central component of this ongoing work is the development of revised lung tissue single-cell genomics references and curated cell types for human, mouse, and other model organisms. Such work requires new iterations of these compendiums as knowledge related to lung cellular diversity and function continues to advance in parallel with new multimodal single-cell, spatial transcriptomics, and proteomics technologies. To develop these annotations, LungMAP has begun a crossconsortium curation initiative with partners from the HCA-Lung and HuBMAP, leveraging new insights from the literature, single-cell genomics, and curation initiatives such as Anatomical Structures, Cell Types, plus Biomarkers (ASCT+B) and the Cell Ontology. Interacting with these data will require new interactive web applications for supervised multimodal classification and exploration, which will likely include extensions to our supported instances of Azimuth, ShinyCell, cellxgene, OMERO, and emerging interactive web interfaces, such as HuBMAP's Vitessce's spatial single-cell analytical framework (31). Furthermore, data integration continues to be a key focus within LungMAP, in particular for disparate datasets produced from different laboratories using different technologies. As more controlled-access single-cell datasets are migrated into BioData Catalyst, users with approved access (dbGAP) will be able to integrate and reanalyze and visualize raw data from diverse studies interactively, using new and emerging cloud-based workflows (e.g., Optimus and Cumulus), using cloud workbenches (e.g., Terra and Seven Bridges) (22).

Standardized and iterative curation metadata curation will remain a significant focus of these efforts, with a particular focus on new supported technologies and sample types. This will include support for common coordinate framework and volumetric CT data in LungMAP for three-dimensional visualization of source biopsy anatomic locations, together with updatable dataset metadata for distributed genomics datasets (e.g., versioned single-cell population identities). To ensure that such increasingly complex datasets can be queried and interpreted by lung researchers from diverse analytical backgrounds, we further aim to improve communication of these resources and conduct targeted virtual education efforts. We welcome community contributions and collaborations with individual laboratories and consortiums, which are vital to the success of these efforts. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors would like to thank the Cincinnati Children's Research Foundation for their additional support and the Data and Technology Services and Research IT Services in the Division of Biomedical Informatics at Cincinnati Children's Hospital Medical Center (CCHMC) for their contributions.

References

1. Sun X, Perl AK, Li R, Bell SM, Sajti E, Kalinichenko VV, *et al.*; NHLBI LungMAP Consortium. A census of the lung: CellCards from LungMAP. *Dev Cell* 2022;57:112–145.e2.
2. Ardini-Poleske ME, Clark RF, Ansong C, Carson JP, Corley RA, Deutsch GH, *et al.*; LungMAP Consortium. LungMAP: the molecular atlas of lung development program. *Am J Physiol Lung Cell Mol Physiol* 2017;313:L733–L740.
3. Moghieb A, Clair G, Mitchell HD, Kitzmiller J, Zink EM, Kim YM, *et al.* Time-resolved proteome profiling of normal lung development. *Am J Physiol Lung Cell Mol Physiol* 2018;315:L11–L24.
4. Endale M, Ahlfeld S, Bao E, Chen X, Green J, Bess Z, *et al.* Temporal, spatial, and phenotypical changes of PDGFRα expressing fibroblasts during late lung development. *Dev Biol* 2017;425:161–175.
5. Masci AM, White S, Neely B, Ardini-Polaske M, Hill CB, Misra RS, *et al.*; LungMAP Consortium. Ontology-guided segmentation and object identification for developmental mouse lung immunofluorescent images. *BMC Bioinformatics* 2021;22:82.
6. Muus C, Luecken MD, Eraslan G, Sikkema L, Waghray A, Heimberg G, *et al.*; NHLBI LungMap Consortium; Human Cell Atlas Lung Biological Network. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat Med* 2021;27:546–559.
7. Du Y, Kitzmiller JA, Sridharan A, Perl AK, Bridges JP, Misra RS, *et al.* Lung gene expression analysis (LGEA): an integrative web portal for comprehensive gene expression data analysis in lung development. *Thorax* 2017;72:481–484.
8. Jin K, Bardes EE, Miteplubkt A, Wang JY, Bhatnagar S, Sengupta S, *et al.* Implicating gene and cell networks responsible for differential COVID-19 host responses via an interactive single cell web portal [preprint]. bioRxiv; 2021 [accessed 2022 Apr 1]. Available from: https://www.biorxiv.org/content/10.1101/2021.06.07.447287v2.
9. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020;6:eaba1983.
10. Ljungberg MC, Sadi M, Wang Y, Aronow BJ, Xu Y, Kao RJ, *et al.* Spatial distribution of marker gene activity in the mouse lung during alveolarization. *Data Brief* 2018;22:365–372.
11. Nguyen SN, Kyle JE, Dautel SE, Sontag R, Luders T, Corley R, *et al.* Lipid coverage in nanospray desorption electrospray ionization mass spectrometry imaging of mouse lung tissues. *Anal Chem* 2019;91:11629–11635.
12. Protsyuk I, Melnik AV, Nothias LF, Rappez L, Phapale P, Aksenov AA, *et al.* 3D molecular cartography using LC-MS facilitated by Optimus and 'ili software. *Nat Protoc* 2018;13:134–154.
13. Ouyang JF, Kamaraj US, Cao EY, Rackham OJL. ShinyCell: simple and sharable visualisation of single-cell gene expression data. *Bioinformatics* 2021;btab209.
14. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, *et al.* Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices [preprint]. bioRxiv; 2021 [accessed 2022 Apr 1]. Available from: https://www.biorxiv.org/content/10.1101/2021.04.05.438318v1.
15. Zepp JA, Morley MP, Loebel C, Kremp MM, Chaudhry FN, Basil MC, *et al.* Genomic, epigenomic, and biophysical cues controlling the emergence of the lung alveolus. *Science* 2021;371:eabc3172.
16. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;37:W305-11.
17. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, *et al.* Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol* 2021;23:1129–1135.
18. Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573–3587.e29.
19. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, *et al.* Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genome* 2022;2:100085.
20. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, *et al.* The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;46:W537–W544.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
22. Li B, Gould J, Yang Y, Sarkizova S, Tabaka M, Ashenberg O, *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat Methods* 2020;17:793–798.
23. Du Y, Guo M, Whitsett JA, Xu Y. 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax* 2015;70:1092–1094.
24. Du Y, Ouyang W, Kitzmiller JA, Guo M, Zhao S, Whitsett JA, *et al.* Lung gene expression analysis web portal version 3: lung-at-a-glance. *Am J Respir Cell Mol Biol* 2021;64:146–149.
25. Du Y, Clair GC, Al Alam D, Danopoulos S, Schnell D, Kitzmiller JA, *et al.* Integration of transcriptomic and proteomic data identifies biological functions in cell populations from human infant lung. *Am J Physiol Lung Cell Mol Physiol* 2019;317:L347–L360.
26. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 2020;587:619–625.
27. Miller AJ, Yu Q, Czerwinski M, Tsai YH, Conway RF, Wu A, *et al.* In vitro and in vivo development of the human airway at single-cell resolution. *Dev Cell* 2020;54:818.
28. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;199:1517–1536.
29. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;11:2285.
30. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* 2020;26:842–844.
31. Keller MS, Gold I, McCallum C, Manz T, Kharcjenko PV, Gehlenborg N. Vitessce: a framework for integrative visualization of multi-modal and spatially-resolved single-cell data [preprint]. OSF Preprints; 2021 [accessed 2022 Apr 1]. Available from: https://osf.io/y8thv.