EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Review

The discovAIR project: a roadmap towards the Human Lung Cell Atlas

Malte D. Luecken, Laure-Emmanuelle Zaragozi, Elo Madissoon, Lisa Sikkema, Alexandra B. Firsova, Elena De Domenico, Louis Kümmerle, Adem Saglam, Marijn Berg, Aurore C.A. Gay, Janine Schniering, Christoph H. Mayr, Xesús M. Abalo, Ludvig Larsson, Alexandros Sountoulidis, Sarah Teichmann, Karen van Eunen, Gerard H. Koppelman, Kourosh Saeb-Parsy, Sylvie Leroy, Pippa Powell, Ugis Sarkans, Wim Timens, Joakim Lundeberg, Maarten van den Berge, Mats Nilsson, Peter Horváth, Jessica Denning, Irene Papatheodorou, Joachim Schultze, Herbert B. Schiller, Pascal Barbry, Ilya Petoukhov, Alexander V. Misharin, Ian Adcock, Michael von Papen, Fabian J. Theis, Christos Samakovlis, Kerstin B. Meyer, Martijn C. Nawijn

Please cite this article as: Luecken MD, Zaragozi L-E, Madissoon E, *et al*. The discovAIR project: a roadmap towards the Human Lung Cell Atlas. *Eur Respir J* 2022; in press (https://doi.org/10.1183/13993003.02057-2021).

The discovAIR project: a roadmap towards the Human Lung Cell Atlas

Malte D. Luecken[1,*], Laure-Emmanuelle Zaragozi[2,*], Elo Madissoon[3,4,*], Lisa Sikkema[1,*], Alexandra B. Firsova[5,*], Elena De Domenico[6,*], Louis Kümmerle[1,*], Adem Saglam[6,*], Marijn Berg[7,8,*], Aurore C.A. Gay[7,8,*], Janine Schniering[9,*], Christoph H. Mayr[9,*], Xesús M. Abalo[10,*], Ludvig Larsson[10,*], Alexandros Sountoulidis[5,*], Sarah Teichmann[3,12], Karen van Eunen[13,14], Gerard H. Koppelman[8,13], Kourosh Saeb-Parsy[15], Sylvie Leroy[16], Pippa Powell[17], Ugis Sarkans[4], Wim Timens[7,8], Joakim Lundeberg[11], Maarten van den Berge[8,18], Mats Nilsson[10], Peter Horváth[19], Jessica Denning[17], Irene Papatheodorou[4], Joachim Schultze[6,20,21], Herbert B. Schiller[9], Pascal Barbry[2], Ilya Petoukhov[22], Alexander V. Misharin[23], Ian Adcock[24], Michael von Papen[25], Fabian J. Theis[1,26], Christos Samakovlis[5], Kerstin B. Meyer[3], Martijn C. Nawijn[7,8]

* These authors made an equal contribution to this manuscript

1. Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany
2. Université Côte d'Azur and CNRS, Institut de Pharmacologie Moléculaire et Cellulaire, Sophia Antipolis, France.
3. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
4. European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK
5. Science for Life Laboratory, Department of Molecular Biosciences, Wenner-Gren Institute, Stockholm University, Stockholm, Sweden.
6. Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany
7. University of Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, Groningen, the Netherlands
8. GRIAC research institute at the University Medical Center Groningen, Groningen, the Netherlands
9. Helmholtz Zentrum München, Institute of Lung Biology and Disease, Comprehensive Pneumology Center, Member of the German Center for Lung Research (DZL), Munich, Germany.
10. Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 171 65 Solna, Sweden
11. Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, 171 65 Solna, Sweden
12. Theory of Condensed Matter, Cavendish Laboratory, 19 JJ Thomson Avenue, Cambridge CB3 0HE, UK.
13. University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, Groningen, the Netherlands
14. UMCG Research BV, University Medical Center Groningen, Groningen, The Netherlands.
15. Department of Surgery, University of Cambridge, and Cambridge NIHR Biomedical Research Centre, Cambridge, CB2 0QQ, UK
16. Université Côte d'Azur and CHU Nice, Département de Pneumologie, FHU-OncoAge, , Nice, France.
17. European Lung Foundation, Sheffield, UK
18. University of Groningen, University Medical Center Groningen, Department of Pulmonology, Groningen, the Netherlands
19. Synthetic and Systems Biology Unit, Biological Research Center, 6726 Szeged, Hungary
20. PRECISE Platform for Single Cell Genomics and Epigenomics, Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) and the University of Bonn, Bonn, Germany

21. Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany.
22. A Beta World (former Principal at MIcompany), Amsterdam, the Netherlands
23. Division of Pulmonary and Critical Care Medicine, Feinberg School of Medicine, Northwestern University, Chicago, USA
24. Airway Disease Section, National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, United Kingdom.
25. Comma Soft AG, Bonn, Germany
26. Institute of Computational Biology, Helmholtz Center Munich (HMGU), Neuherberg, Germany

Acknowledgements

This work is supported by the European Union's Horizon2020 Research and Innovation Program under grant agreement no. 874656 (discovAIR) to ST, KSP, SL, WT, JL, MvdB, MS, PH, JD, IP, JS, HBS, PB, MvP, FJT, CS, KBM and MCN, and a Seed Network grant from the Chan Zuckerberg Initiative to PB, HBS, KBM, AVM, MCN and FJY. MvdB and MCN are also supported by grants (no. 5.1.14.020 and 4.1.18.226) from the Netherlands Lung Foundation. KBM and SAT are also supported by Wellcome (WT211276/Z/18/Z and Sanger core grant WT206194). JS is also supported by an ERS/RESPIRE4 Marie Sklodowska Curie fellowship (R4202007-00844). PB was also supported by Institut National contre le Cancer (PLBIO2018-156), FRM (DEQ20180339158), Inserm Cross-cutting Scientific Program HuDeCA 2018, and the National Infrastructure France Génomique (Commissariat aux Grands Investissements, ANR-10-INBS-09-03, ANR-10-INBS-09-02).

Conflict of Interest statement

Dr. Theis reports personal fees and non-financial support from Cellarity, and from Dermagnostix, during the conduct of the study; Janine Schniering was supported by a ERS/EU RESPIRE4 Marie Sklodowska Curie Postdoctoral Fellowship; Dr. Lundeberg reports personal fees from 10x Genomics Inc, outside the submitted work; P. Powell and J. Denning are employees of the European Lung Foundation; Dr. Timens reports personal fees from Merck Sharp Dohme, personal fees from Bristol-Myers-Squibb, outside the submitted work; Dr. Nilsson reports personal fees from 10x Genomics, outside the submitted work; Dr. Koppelman reports grants from Lung Foundation of the Netherlands, GSK, Vertex, TEVA the Netherlands, UBBO EMMIUS Foundation, European Union (H2020 grant), outside the submitted work and he has participated in advisory board meetings to GSK and PURE-IMS, outside the submitted work; Dr. van den Berge reports research grants paid to UMCG from GlaxoSmithKline, Genentech, Roche, Novartis, outside the submitted work; Dr. Nawijn reports grants from European Commission, grants from Chan Zuckerberg Initiative, and grants from The Netherlands Lung Foundation, during the conduct of the study; grants from GSK Ltd, outside the submitted work. All other authors do not have a conflict of interest.

Abstract
The Human Cell Atlas (HCA) consortium aims to establish an atlas of all organs in the healthy human body at single-cell resolution to increase our understanding of basic biological processes that govern development, physiology and anatomy, and to accelerate diagnosis and treatment of disease. The lung biological network of the HCA aims to generate the Human Lung Cell Atlas as a reference for the cellular repertoire, molecular cell states and phenotypes, and the cell–cell interactions that characterize normal lung homeostasis in healthy lung tissue. Such a reference atlas of the healthy human lung will facilitate mapping the changes in the cellular landscape in disease. The discovAIR project is one of six pilot actions for the HCA funded by the European Commission in the context of the H2020 framework program. DiscovAIR aims to establish the first draft of an integrated Human Lung Cell Atlas, combining single-cell transcriptional and epigenetic profiling with spatially resolving techniques on matched tissue samples, as well as including a number of chronic and infectious diseases of the lung. The integrated Lung Cell Atlas will be available as a resource for the wider respiratory community, including basic and translational scientists, clinical medicine, and the private sector, as well as for patients with lung disease and the interested lay public. We anticipate that the Lung Cell Atlas will be the founding stone for a more detailed understanding of the pathogenesis of lung diseases, guiding the design of novel diagnostics and preventive or curative interventions.

Introduction

Lung diseases are leading causes of death worldwide [1], with incidences increasing at an alarming rate while curative interventions are lacking for most of these disorders. Research into lung diseases lags behind compared to cancer and cardiovascular disease, the other two major causes of death [2]. The high anatomical complexity of lung tissue, with the bifurcating bronchial tree, facilitating transport of air, and the parenchyma containing the alveoli for gas exchange, as well as its extremely rich cellular heterogeneity, with more than 50 different cell types identified to date [3], are some of the obstacles hampering progress in understanding the mechanisms of the many different lung diseases. These discrete lung cell types are each defined by morphological features as well as constitutive gene expression patterns that are essential for cell type identity [4]. Each cell type can adopt various molecular cell states, defined by transient expression of additional, facultative gene modules to allow execution of specific functions, adaptations to environmental factors or stimuli, or transition to another cell type during differentiation or in disease pathogenesis [4]. Therefore, the exact molecular phenotype of lung cells in time and space is determined by their exact location within the highly complex, three-dimensional structure of the lung, their local interactions with other cells, with the lung matrix and with the external environment that is omnipresent in this organ. The rich cellular complexity and precise spatial organization are critical for proper lung function in health, and are often lost in disease. Hence, there is an urgent need to have a detailed description of this complexity in healthy lung tissue and propel basic, translational and clinical research in lung disease into a fast track for the development of precision diagnostics and therapeutics. To achieve this, we need a detailed understanding of the cells that make up the lung, their fixed and variable features, their interactions and their organization into local cellular neighborhoods and higher-order structures that make up the macroscopic tissue architecture in health and the deviations thereof in disease [4].

The central goal of the Human Cell Atlas (HCA; https://humancellatlas.org) is to establish such an atlas of all organs in the healthy human body [5]. Achieving this daunting task for the lung is the central goal of the Lung Biological Network of the HCA [6], an open research community that encompasses a large number of research groups, as well as several research consortia, including discovAIR (https://discovAIR.org), the CZI Seed Network for the Human Cell Atlas (http://bit.ly/HCALungCZI1), LungMAP (https://LungMAP.net) and HuBMAP (https://HuBMAPconsortium.org). The discovAIR consortium is one of six pilot actions for the Human Cell Atlas funded by the European Commission in the context of the H2020 framework program.

DiscovAIR aims to contribute to establishing the first draft of an integrated **Human Lung Cell Atlas**. In this atlas, discovAIR will contribute significantly to the multimodal single-cell omics data from healthy human lung, to mapping the cellular heterogeneity observed in single-cell RNA-seq (scRNA-Seq) datasets onto the lung tissue architecture, and to identification of changes of molecular cell phenotypes and their interactions in lung disease cohorts such as asthma, chronic obstructive pulmonary disease (COPD), pulmonary arterial hypertension (PAH), coronavirus disease of 2019 (COVID-19), and interstitial lung diseases (ILD) such as interstitial pulmonary fibrosis (IPF), enabling accelerated translational and clinical research into lung diseases. The discovAIR results will facilitate progress in regenerative and precision medicine by identifying novel candidates for precision diagnostics and curative interventions in lung disease. Here, we present the contributions of the discovAIR project to the roadmap of the Human Lung Cell Atlas from the HCA Lung Biological Network.

The Lung Biological Network of the Human Cell Atlas

The anatomy, physiology and cell type composition of the lung has been studied in great detail using classical immunohistochemical and pathological approaches, with well-characterized relationships between cellular phenotype and function [7]. The application of single-cell RNA-sequencing

techniques to human lung tissue, however, showed that our knowledge of the cellular composition of the lung is incomplete, and novel cell types such as the pulmonary ionocyte have been identified [8, 9]. Consequently, the HCA Lung Biological Network (HCA-Lung) aims to identify all cell types and their molecular states or activities present in healthy lung tissue, their interactions and organization into higher-order anatomical and/or functional units [4].

Lung tissue is exquisitely suitable for such a mapping effort, due to the presence of unambiguous tissue landmarks to relate local cellular neighbourhoods and larger-order cellular structures back to defined physical coordinates within the organ. Importantly, healthy lung tissue is available to a large number of research groups through lung resections in patients with lung disease (healthy lung tissue adjacent to well-defined regions of disease, such as lung cancer), but also from organ donation programs and in some cases through research bronchoscopy programs involving healthy control subjects. This combination of a highly ordered tissue architecture - facilitating the implementation of a common coordinate framework (CCF) [10] - and good community-wide availability of tissue makes lung especially well-suited as a lead organ for the HCA to develop the infrastructure, workflows, platforms and computational approaches needed for a community-driven tissue mapping effort as laid down in the vision of the HCA consortium [5]. Consequently, atlases of both the airways and parenchymal lung tissue have been selected by the HCA consortium as a priority effort, with the Lung Cell Atlas having developed into one of the HCA Flagship projects [6]. The infrastructure, workflows, roadmap and platforms developed within HCA-Lung can serve as blueprints for other biological networks of the HCA community.

The roadmap of the Lung Biological Network

The identification of the pulmonary ionocyte as a previously unknown cell type present in healthy lung tissue [8, 9], sparked a number of studies mapping the cellular heterogeneity of healthy lung tissue with increasing spatial and cellular resolution, and taking into account smoking as one of the most important factors driving specific cell states in lung tissue [3, 11–13]. An overview of current lung tissue scRNA-Seq datasets and their availability is provided in Appendix 1. Progress in the Lung Cell Atlases has also recently been summarized [14]. Notwithstanding the importance of these foundational datasets for the respiratory community, HCA-Lung aspires to move well beyond this current state-of-the-art that is hampered by relatively low number of donors, poor coverage of different ethnicities and ancestral diversity as well as age range of tissue donors, limited resolution across the CCF and lack of spatial mapping of the identified cell states onto the tissue architecture [4, 6].

Therefore, HCA-Lung aims to provide a true reference atlas of the lung that captures the heterogeneity of cell states associated with location within the organ, as well as the variation of cell states that can be observed within healthy lung tissue, as a function of genetic, demographic and geographical variables. One approach to establish such a reference atlas is to leverage recent advances in computational data integration techniques [15–19] to integrate the currently available single-cell RNA-seq datasets into a single embedding, thereby establishing an integrated Lung Cell Atlas that captures most of the available data and identifies undersampled regions and cell populations, and efficiently corrects for batch effects. In addition to capturing the full heterogeneity in transcriptional cell states of lung cells, HCA-Lung aims to incorporate a much larger variation in ethnicity and ancestral diversity, age range and geographical location of tissue donors, as well as multiple layers of omics data into an advanced draft of the Lung Cell Atlas, including epigenetic features such as CpG methylation, chromatin accessibility and histone modifications, or proteomic features at single-cell resolution. Leveraging the recent advances in single cell joint profiling protocols, paired modalities can enrich our RNA reference using transfer learning approaches that map the RNA measurements onto each other [20]. Finally, the lack of a systematic description of molecular cell phenotypes with spatially resolving methods is severely hampering the interpretation

of the wealth of data generated by the single-cell omics approaches. Clearly, adding such spatial maps of healthy lung tissue is one of the current priorities of HCA-Lung. Merging spatial datasets with an integrated Lung Cell Atlas based on multiple layers of omics data will then establish a true Human Lung Cell Atlas of healthy lung tissue, which will be an invaluable resource for basic, translational and clinical research into lung and its diseases.

The discovAIR approach to establish the Human Lung Cell Atlas

The discovAIR project aims to contribute to the HCA-Lung goals by delivering the first version of the Human Lung Cell Atlas towards the summer of 2022. Over the last two years a number of studies have presented tissue atlases of both healthy and diseased human lung [3, 11, 13, 21–30], using single-cell RNA-sequencing complemented with single molecule fluorescent in situ hybridization (smFISH) analyses for validation of the main results. While these studies provide a rich resource for mapping cell-type specific gene expression patterns, as exemplified by the rapid description of the expression patterns of genes encoding the SARS-CoV-2 cell entry factors by the HCA Lung community during the COVD-19 pandemic [31], these are usually single-center studies and limited by the relatively low number of biological replicates, the low resolution in spatial locations sampled, the annotation of 'novel' cell type labels lacking ontological or pathological context, and the lack of detailed spatially resolving methods to accompany the single-cell RNA-seq data.

To address this shortcoming, discovAIR will combine multimodal single-cell profiling of lung cells with multimodal spatial mapping of the lung cell types and their molecular states onto the lung tissue architecture (see Figure 1). Moreover, discovAIR partners will integrate multiple lung tissue single-cell datasets into a single embedding, establishing the first draft of an integrated Lung Cell Atlas, allowing cell-type specific analysis of the transcriptomic variation explained by demographic covariates, as was recently piloted for a limited number of SARS-CoV-2 cell entry genes by a highly collaborative HCA-Lung effort [32]. We have recently benchmarked computational approaches for data integration of existing single-cell RNA-seq datasets [19]. We will use this framework to identify the best-suited method for integration of available scRNA-seq datasets from healthy lung tissue. The resulting integrated dataset will be presented as the core reference of healthy lung tissue, and can be used as the first draft of an integrated Lung Cell Atlas. This atlas can then be used as a reference to which newly generated datasets from either healthy or diseased lung tissue can be compared using transfer learning methods such as scArches [20]. Taking this approach is especially powerful, as it allows unified use of cell type labels, direct comparisons of cellular composition across datasets and - in case of diseased tissue - direct identification of unique, disease-associated cell states absent from the healthy reference. Moreover, the integrated Lung Cell Atlas can help to structure discussions around cell type label harmonization and mapping of the hierarchical cell type labels used for data integration (see Figure 1) to existing ontologies of the cells of the lung [33].

Towards the end of the project, discovAIR will establish a second draft of this integrated Lung Cell Atlas incorporating all newly generated data within the consortium as well as data from the larger HCA Lung Biological Network community. The newly generated dataset from the discovAIR project will entail scRNA-seq data from at least 60 healthy controls spanning a large age range, both sexes and multiple ethnicities, as well as the detailed characterization of the 5-location dataset from healthy donor lung tissue encompassing multimodal data as well as matching spatial datasets. It is important to bear in mind that any tissue sample obtained for building the reference atlas of healthy lung tissue will be derived from either deceased individuals whose lungs are suitable for organ transplant, lung resection programs in routine clinical care (often as part of cancer treatment) or bronchoscopy studies performed for research purposes on healthy volunteers. Of these, only the latter source can be considered to reflect truly healthy, fresh lung tissue. The integrated Lung Cell Atlases will be made freely available to the respiratory and scientific community to ensure optimal impact of the discovAIR efforts. In addition to these integrated atlases of scRNA-seq data, discovAIR

aims to develop innovative visualisations for spatial datasets, including 3D reconstruction of lung tissue architecture and molecular phenotyping of local cellular neighbourhoods by multiplexed sm-FISH or immunofluorescence analysis.

Most datasets with a large number of biological replicates sample only a limited number of locations within the organ, and use a single-omics technique. To contribute to a more balanced coverage of multimodal sampling across the CCF in sufficient numbers of donors, discovAIR aims to establish a reference dataset offering an in-depth exploration of healthy lung tissue from 5 deceased transplant organ donors, each sampled at 5 discrete locations (see Figure 1) using both scRNA-Seq and scATAC-Seq to characterise their constant and variable features, as well as multimodal spatial profiling of lung cells on adjacent tissue sections, to map their organisation in a three dimensional tissue architecture, and their local and distant interactions. The generation of in-depth transcriptomic data will allow us to predict cell-cell interactions and the receptor-ligand pairs mediating these. The discovAIR approach will also allow validation of these predicted cell-cell interactions by spatial data that identifies, in a quantitative manner, cellular interactions.

The discovAIR project has selected 4 different spatially resolving methods. Of these, two are probe-based, allowing validation of the spatial expression pattern of a (limited) gene set, selected on the basis of the scRNA-seq data, by mapping these onto the tissue architecture. The other two methods are sequencing-based spatially resolving methods, allowing detection of spatial gene expression patterns without limiting the analysis to a prior selection of genes.

The two probe-based methods to map the transcriptional variation and cell-type heterogeneity observed in the single-cell datasets onto the spatial architecture of lung tissue differ in sensitivity and multiplexity. First, we will use a panel of 64 probes in the sm-FISH based method SCRINSHOT [34], that has high sensitivity, but depends on sequential rounds of hybridization and imaging of the same tissue section to achieve multiplexity. We have established two probe panels based on the transcriptional variation observed in a similar dataset covering these same 5 locations. One probeset was designed for airway wall-resident cell types and one for cell types of the lung parenchyma. In addition to SCRNSHOT, we will map the cell types and their molecular phenotypes in high detail with a panel of 150 probes using *in situ* sequencing (ISS), an alternative approach that has higher multiplexity, but not the sensitivity of SCRNSHOT [35]. Probe selection for both SCRINSHOT and ISS was performed using a multi-objective computational approach, which optimizes our ability to discern cell types of interest while recovering most of the transcriptional variation observed in these tissue samples with the selected probes (Appendix 2: the current discovAIR smFISH gene lists for probe design in ISS and SCRINSHOT).

In addition to these probe-based spatial approaches, discovAIR has selected two unbiased, sequencing-based methods based on their capacity to generate spatially resolved transcriptomic data from the adjacent tissue sections from the same lung tissue samples as used in SCRNSHOT and ISS. In this pilot project for the Human Lung Cell Atlas, we have chosen to focus on spatially resolved transcriptome-based data only, while protein-based analyses will need to be integrated at a later stage. First, spatial transcriptomics by Visium [36] will be performed on 6.5 mm x 6.5 mm x 10 micron sections from the same lung tissue blocks also used for SCRINSHOT and ISS. The Visium spots are 55 microns in diameter with a 100 micron center-center distance, and routinely yield >4000 genes per spot. Considering the architecture of the lung, we aim to profile pairs of adjacent sections (10 micron) followed by a gap of 100 micron so that a total of about 10 tissue sections will cover a tissue volume of over 1 cm^3. A data-driven 3D model will be created based on the gene expression data. We have selected two out of the five deep dive locations to be included in the 3D transcriptomic approach: location 3 (3rd/5th generation airway) and location 5 (lung parenchyma; see Figure 1D), which will thus generate a first draft map of the transcriptome in two main sampling

locations of the 5 deceased donor lungs. The 3D transcriptomic results will be integrated with the matching snRNA seq data generated on adjacent tissue sections to validate the spatial coordinates of the identified cell types. Moreover, SCRINSHOT and ISS data from adjacent sections will be available to guide further validation and higher-resolution mapping of the 3D transcriptomic map onto the tissue architecture. We will create probabilistic spatial cell maps of scRNAseq defined cell types using approaches like pciSeq [37] and Tangram [38].

The second sequencing-based approach with spatial resolution employs automated nuclear isolation using the laser capture microdissection (LCM) method [39], followed by snRNA-seq or snATAC-seq of the individual nuclei containing spatial coordinates. In this approach, a section of the lung tissue with a thickness of 10 microns and lateral sizes varying between few millimeters and 1 centimeter is imaged in up to three color channels. The nuclei of interest are detected by a machine learning method using the intensity information from the different color channels. The detected nuclei are then sequentially cut and collected in a collector plate filled with the appropriate lysis buffer and further processed for snRNA-seq or snATAC-seq. The spatial location of each nucleus is stored in a file by the LCM system. The RNA or ATAC sequencing data generated in this way will be correlated to the spatial location of the individual nuclei to provide insights on the distribution of particular cell populations across the human lung tissue. These datasets will be integrated with the matching snRNA-seq as well as the SCRINSHOT and ISS data, and 3D Visium data for the 2 locations where this is available. Together, this will generate a spatial map of cell types and cell states of the healthy human lung, including 3D reconstruction of small airways and lung parenchyma.

Finally, DiscovAIR aims to provide a detailed profiling of the cellular trajectories of healthy adult lung cells towards the disease-associated cell states. To this end, discovAIR will study a (limited) number of chronic and/or inflammatory lung diseases: asthma, COPD, ILD, PAH and COVID-19. Moreover, discovAIR will establish a lung cell perturbation map to chart cell-state transitions between health and disease (see Figure 1).

Lung diseases are highly heterogeneous, and several phenotypes as well as endotypes can be observed for nearly all chronic lung diseases. The scope of discovAIR is to provide proof-of-principle data that a high-quality reference cell atlas of healthy lung and can be used in combination with datasets from lung tissue of patients with lung disease to identify disease-associated cell states and cell-cell interactions, and the healthy-to-diseased cellular trajectories can be further characterized using the perturbation atlas. Therefore, discovAIR has chosen to include relatively small numbers of samples from patients with lung disease (10 tissue samples per disease condition). In addition, discovAIR has selected very specific disease groups, such as adult patients with (physician diagnosed) childhood-onset asthma without a history of cigarette smoking, in an attempt to minimize the chance that heterogeneity of the disease obscures any reproducible disease-associated transcriptional cell states in the final dataset.

The discovAIR perturbation atlas will use primary epithelial, immune and endothelial cells, as well as precision-cut lung slices, all isolated from lung tissue of healthy donors, and stimulated in advanced *in vitro* culture models [40, 41]. Time-series analysis of the *ex vivo* stimulated primary cells by scRNA-seq will allow the identification of transitional cell states in the cellular trajectories of healthy cells towards a potentially disease-associated transcriptional cell state. The discovAIR project focusses on testing this perturbation atlas concept in epithelial, endothelial and immune cells given the availability of well-developed cell culture models available to the consortium. However, other cell types such as mesenchymal cells including fibroblast subsets and smooth muscle cells are likely to play key roles in disease inception, progression and exacerbations, and will need to be studied in a similar approach in future follow-up projects if the concept of the perturbation atlas is validated. Integration of the datasets from *ex vivo* cultures of healthy primary lung cells and tissues with the

scRNA-seq datasets acquired in lung tissue samples obtained from patients with lung disease is expected to allow the identification of the optimal *in vitro* proxy for the diseased cell states observed *in vivo*.

These studies are expected to further uncover cellular mechanisms of disease, guide identification of potential targets for preventive or therapeutic intervention and reveal biomarkers that can be used to detect the presence of these specific diseased cell states for use in diagnosis or treatment response monitoring. Given that discovAIR is a 2-year pilot project, the approach has a strong focus on transcriptomic data to establish the first draft of the Human Lung Cell Atlas in health in disease. Notwithstanding the significant progress beyond the state-of-the-art such an Atlas would entail, these foundational transcriptomic datasets will need to be validated at the protein level, both for the Human Lung Cell Atlas describing the transcriptional heterogeneity in healthy lung tissue as well as for the changes therein associated with chronic lung disease, to be able to achieve their full impact for the respiratory community and patients with lung disease.

Taken together, the multimodal healthy lung tissue datasets, state-of-the-art integration methods, the spatial mapping of healthy lung tissue and the disease cohorts in combination with the perturbation atlas will allow discovAIR to generate a first draft of the Human Lung Cell Atlas (V1.0) as a standard reference for the respiratory community. This will not only encompass a healthy reference Lung Cell Atlas combining multimodal omics and spatial datasets, but also a comprehensive description of the changes thereof with disease, and the cellular trajectories that might lead to the acquisition of the unique, disease-associated cell states. This will be a key asset for the respiratory community and is expected to facilitate progress in regenerative and precision medicine for lung disease, and guide identification of novel candidates for precision diagnostics and curative interventions. All discovAIR methods and best practices will be openly shared through open access portals such as protocols.io (https://www.protocols.io/workspaces/hca) and Github (https://github.com/LungCellAtlas & appendix 3).

Dissemination, public engagement and outreach for the Human Lung Cell Atlas

To achieve full impact, discovAIR has the European Respiratory Society (ERS) and the European Lung Foundation (ELF) as project partners for dissemination of results and community involvement in the Human Lung Cell Atlas. ELF is a patient ambassador organization, aiming to bring together patients and the public with respiratory professionals to positively influence lung health, and is essential to safeguard patient involvement, public engagement and outreach for discovAIR. ERS is the largest scientific and clinical organisation in respiratory medicine in Europe. Involvement of ERS in the Human Lung Cell Atlas initiative through discovAIR is instrumental to inform and involve the basic, translational and clinical respiratory scientific community as well as diagnostic, regenerative medicine and pharmaceutical industries with respect to the progress and achievements of the Human Lung Cell Atlas.

Needs from different user communities of a Human Lung Cell Atlas

A critical part of the discovAIR dissemination strategy is to involve the different user communities that will benefit from the Human Lung Cell Atlas early on. The discovAIR consortium organized a user group meeting for the Lung Cell Atlas at the 2020 ERS Lung Science Conference in Estoril, Portugal, together with ELF. This meeting aimed to identify the needs of the different user groups with regard to content and design of the Lung Cell Atlas, as well as the best approach to ensure optimal use of the Lung Cell Atlas. The user groups represented in this meeting were patients with different respiratory diseases, basic and translational scientists active in the respiratory field, clinical experts, and representatives of the diagnostic and pharmaceutical industry. The results from the user group meeting clearly indicate that the expectations and needs for the Lung Cell Atlas differ between the

patient representatives and the lay public on the one hand, and experts including clinician, scientists and representatives from the private sector on the other hand.

Patient needs

The patient ambassadors clearly indicated that the added value of a Lung Cell Atlas for them is to better understand their lung disease. Patients with lung disease indicated that the Lung Cell Atlas might help them understand the structure and function of the lung, the changes in disease, as well as the identity and basic characteristics of the cells that make up the lung. Of special interest to the patient is the opportunity to understand how their disease condition affects the cells of the lung, how this is reflected in specific (diagnostic) test results and how prescribed drugs can work to restore normal cell functions and interactions. As such, the Lung Cell Atlas can serve as a tool in interactions between patients and their doctors. Thus, information needs to be presented in the context of a specific disease.

In addition, the Human Lung Cell Atlas could be used as an educational tool, to explain the details of a lung disorder to relatives and the lay public, and might help to educate the next generation of patient advocates and respiratory scientists. Finally, the patient representatives also recognized the Lung Cell Atlas as a potential tool for scientists, clinicians and partners in the private sector to develop novel treatments for lung disease that can improve the quality of life for patients with lung disease.

Scientific, clinical and industry needs

The needs indicated by the scientific, clinical and industry representatives at the Human Lung Cell Atlas user group meeting were much more detailed, with open access to data being considered as one of the most important features of the Human Lung Cell Atlas. Data access could be facilitated either by querying the platform, by downloading the data or by obtaining contact details of the data guardians. Furthermore, the respiratory scientists would like to see details on the aspects of the Atlas that are not well covered in the current draft of the Human Lung Cell Atlas, as well as opportunities to contribute their data to a next iteration of the atlas. Also, the Human Lung Cell Atlas should offer details on disease-induced molecular phenotypes of lung cells, as well as the cellular interactions causing disease or are altered by disease to increase its impact with the respiratory community.

Representatives from the pharmaceutical industry were highly interested to use the Lung Cell Atlas to map gene expression to specific locations in the bronchial tree or parenchyma, and to guide design of drug delivery methods for targeted therapies. Furthermore, detailed insight into the cellular transcriptomes, behaviors and interactions in the different regions of the lung and in subtypes of disease could accelerate drug design for precision medicine. The industry representatives further stressed the importance of access to the raw data and future expansions of the Lung Cell Atlas with, for instance, single-cell epigenetic or proteomic datasets, as well as datasets from a large number of lung diseases.

Functionalities to meet the different needs

Given the divergent needs indicated by the patient representatives and the individuals representing respiratory science, medicine and industry, all stakeholders agreed that in designing a Lung Cell Atlas, two separate portals might need to be developed. All user groups indicated that curiosity into the biology of the lung and its disorders is an important incentive to access the Lung Cell Atlas. Patient representatives indicated that clear illustrations and a step-by-step introduction into the anatomy and physiology of the lung would be extremely helpful before accessing the individual cell types and their changes in disease, with complexity slowly increasing as the user gets to the more detailed parts of the atlas. Accessing the atlas through different mobile devices, with strong visual

support and structured search functions, as well as the ability to leave feedback were also of importance.

Interactive resources with stories from patients around specific regions or structures in the lung would greatly increase the attractiveness of the Lung Cell Atlas, especially when these are updated regularly and kept up-to-date and relevant (for instance around cessation of cigarette smoking or the consequences of COVID-19). Finally, the Lung Cell Atlas would need to be available in different languages and hosted or mirrored by various local and national organizations for patients with lung disease, each in their own language, to make it truly accessible to patients and the public. The Lung Cell Atlas v1.0 may serve as a platform that could be expanded to accommodate these features in the future.

The individuals from academia, clinic and industry indicated that the portal should allow maximal interaction with the data through a variety of analysis tools, and the availability of the data for download to perform such analyses offline. Examples are analyses of gene co-expression networks, of trajectories along spatial, demographic or disease parameters, of DNA variant queries, of gene ontology and of cell-type specificity regarding gene expression networks. An interactive analytical tool or browser, that could be used to analyze the data in such a way that its results could be presented in scientific publications or presentations, would clearly increase the impact and recognition of the Human Lung Cell Atlas as an authoritative reference tool. In addition, such functionalities would enable the Human Lung Cell Atlas to maximally contribute to open data and open science. These data analysis tools will be made available through a dedicated webportal at the Single-Cell Expression Atlas at EBI [42], as well as through the FASTGenomics platform (https://fastgenomics.org/).

Conclusions

The Human Lung Cell Atlas is a shared goal of several international consortia, all of which are represented in the Lung Biological Network of the Human Cell Atlas. The discovAIR consortium is the main European consortium active within HCA-Lung, and is one of the six research and innovation actions funded by the European Committee in the Horizon2020 framework program. The discovAIR consortium aims to contribute to the goals and efforts of the Lung Biological Network of the Human Cell Atlas by providing the multimodal characterization of healthy lung tissue, including spatial mapping of cell types and cell states onto the tissue architecture. This will allow discovAIR to generate a first draft of the Human Lung Cell Atlas as a standard reference for the respiratory community.

In addition to the healthy reference Lung Cell Atlas combining multimodal omics and spatial datasets, discovAIR will also provide a first description of the changes thereof with several lung diseases, and the cellular trajectories that might lead to the acquisition of the unique, disease-associated cell states. This will be a key asset for the respiratory community and is expected to facilitate progress in regenerative and precision medicine for lung disease, and guide identification of novel candidates for precision diagnostics and curative interventions. DiscovAIR will develop portals for data exploration and analysis suitable for the academic and industrial end-users, as well as information portals for patients, their families and the interested lay audience in collaboration with the ELF, an ambassador organization for patients with lung disease. All discovAIR methods, datasets and atlases will be made freely available and serve as a basis for further expansions and updates by the Lung Biological Network of the Human Cell Atlas consortium. Future iterations of the Lung Cell Atlas are expected to increase the genetic diversity of the atlas, to incorporate fetal and pediatric lung development, to expand the number of diseases and diseased samples included in the atlas and to further develop the interactive and integrated features of the Human Lung Cell Atlas across transcriptomic, epigenomic, proteomic and spatial data modalities, evolving into a standard reference for the respiratory community.

References

1. Gibson GJ, Loddenkemper R, Lundbäck B, Sibille Y. Respiratory health and disease in Europe: the new European Lung White Book. *Eur. Respir. J.* 2013; 42: 559–563.

2. Fact sheets [Internet]. [cited 2021 Jul 19].Available from: https://www.who.int/news-room/fact-sheets/.

3. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J, Berry GJ, Shrager JB, Metzger RJ, Kuo CS, Neff N, Weissman IL, Quake SR, Krasnow MA. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 2020; 587: 619–625.

4. Schiller HB, Montoro DT, Simon LM, Rawlins EL, Meyer KB, Strunz M, Vieira Braga FA, Timens W, Koppelman GH, Budinger GRS, Burgess JK, Waghray A, van den Berge M, Theis FJ, Regev A, Kaminski N, Rajagopal J, Teichmann SA, Misharin AV, Nawijn MC. The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease. *Am. J. Respir. Cell Mol. Biol.* 2019; 61: 31–41.

5. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, et al. The Human Cell Atlas. *Elife* 2017; 6. Available from: http://dx.doi.org/10.7554/eLife.27041.

6. Regev A, Teichmann S, Rozenblatt-Rosen O, Stubbington M, Ardlie K, Amit I, Arlotta P, Bader G, Benoist C, Biton M, Bodenmiller B, Bruneau B, Campbell P, Carmichael M, Carninci P, Castelo-Soccio L, Clatworthy M, Clevers H, Conrad C, Eils R, Freeman J, Fugger L, Goettgens B, Graham D, Greka A, Hacohen N, Haniffa M, Helbig I, Heuckeroth R, Kathiresan S, et al. The Human Cell Atlas White Paper. arXiv 2018. Available from: http://arxiv.org/abs/1810.05192.

7. Franks TJ, Colby TV, Travis WD, Tuder RM, Reynolds HY, Brody AR, Cardoso WV, Crystal RG, Drake CJ, Engelhardt J, Frid M, Herzog E, Mason R, Phan SH, Randell SH, Rose MC, Stevens T, Serge J, Sunday ME, Voynow JA, Weinstein BM, Whitsett J, Williams MC. Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proc. Am. Thorac. Soc.* 2008; 5: 763–766

8. Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 2018; 560: 377–381.

9. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J, Rogel N, Burgin G, Tsankov AM, Waghray A, Slyper M, Waldman J, Nguyen L, Dionne D, Rozenblatt-Rosen O, Tata PR, Mou H, Shivaraju M, Bihler H, Mense M, Tearney GJ, Rowe SM, Engelhardt JF, Regev A, Rajagopal J. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 2018; 560: 319–324.

10. Rood JE, Stuart T, Ghazanfar S, Biancalani T, Fisher E, Butler A, Hupalowska A, Gaffney L, Mauck W, Eraslan G, Marioni JC, Regev A, Satija R. Toward a Common Coordinate Framework for the Human Body. *Cell* 2019; 179: 1455–1467.

11. Vieira Braga FA, Kar G, Berg M, Carpaij OA, Polanski K, Simon LM, Brouwer S, Gomes T, Hesse L, Jiang J, Fasouli ES, Efremova M, Vento-Tormo R, Talavera-López C, Jonker MR, Affleck K, Palit S,

Strzelecka PM, Firth HV, Mahbubani KT, Cvejic A, Meyer KB, Saeb-Parsy K, Luinge M, Brandsma C-A, Timens W, Angelidis I, Strunz M, Koppelman GH, van Oosterhout AJ, et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* 2019; 25: 1153–1163.

12. Goldfarbmuren KC, Jackson ND, Sajuthi SP, Dyjack N, Li KS, Rios CL, Plender EG, Montgomery MT, Everman JL, Bratcher PE, Others. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat. Commun.* 2020; 11: 1–21.

13. Deprez M, Zaragosi L-E, Truchi M, Garcia SR, Arguel M-J, Lebrigand K, Paquet A, Pee'r D, Marquette C-H, Leroy S, Barbry P. A single-cell atlas of the human healthy airways

14. Meyer KB, Wilbrey-Clark A, Nawijn M, Teichmann SA. The Human Lung Cell Atlas: a transformational resource for cells of the respiratory system. Lung Stem Cells in Development, Health and Disease 2021. p. 158–174.

15. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat. Methods* 2019; 16: 715–721.

16. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* 2021; 17: e9620.

17. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 2018; 15: 1053–1058.

18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. *Cell* 2019; 177: 1888–1902.e21

19. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ. Benchmarking atlas-level data integration in single-cell genomics. bioRxiv 2020 [cited 2021 Apr 8]. p. 2020.05.22.111161

20. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M. Query to reference single-cell integration with transfer learning. *bioRxiv* 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.07.16.205997v1.abstract.

21. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung M-I, Taylor CJ, Jetter C, Raju L, Roberson J, Ding G, Wood L, Sucre JMS, Richmond BW, Serezani AP, McDonnell WJ, Mallal SB, Bacchetta MJ, Loyd JE, Shaver CM, Ware LB, Bremner R, Walia R, Blackwell TS, Banovich NE, Kropski JA. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1972.

22. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, Fernandez R, Akbarpour M, Chen C-I, Ren Z, Verma R, Abdala-Valencia H, Nam K, Chi M, Han S, Gonzalez-Gonzalez FJ, Soberanes S, Watanabe S, Williams KJN, Flozak AS, Nicholson TT, Morgan VK, Winter DR, Hinchcliff M, Hrusch CL, Guzy RD, Bonham CA, Sperling AI, Bag R, Hamanaka RB, et al. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* 2018.

23. Morse C, Tabib T, Sembrat J, Buschur KL, Bittar HT, Valenzi E, Jiang Y, Kass DJ, Gibson K, Chen W, Mora A, Benos PV, Rojas M, Lafyatis R. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* 2019; 54.

24. Valenzi E, Bulik M, Tabib T, Morse C, Sembrat J, Trejo Bittar H, Rojas M, Lafyatis R. Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. *Ann. Rheum. Dis.* 2019; 78: 1379–1387.

25. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, DeIuliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, Washko GR, Homer R, Yan X, Rosas IO, Kaminski N. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1983

26. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, Harding P, Polanski K, Huang N, Nowicki-Osuch K, Fitzgerald RC, Loudon KW, Ferdinand JR, Clatworthy MR, Tsingene A, van Dongen S, Dabrowska M, Patel M, Stubbington MJT, Teichmann SA, Stegle O, Meyer KB. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* 2019; 21: 1.

27. Grant RA, Morales-Nebreda L, Markov NS, Swaminathan S, Querrey M, Guzman ER, Abbott DA, Donnelly HK, Donayre A, Goldberg IA, Klug ZM, Borkowski N, Lu Z, Kihshen H, Politanska Y, Sichizya L, Kang M, Shilatifard A, Qi C, Lomasney JW, Argento AC, Kruser JM, Malsin ES, Pickens CO, Smith SB, Walter JM, Pawlowski AE, Schneider D, Nannapaneni P, Abdala-Valencia H, et al. Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* 2021; 590: 635–641.

28. Delorey TM, Ziegler CGK, Heimberg G, Normand R, Yang Y, Segerstolpe Å, Abbondanza D, Fleming SJ, Subramanian A, Montoro DT, Jagadeesh KA, Dey KK, Sen P, Slyper M, Pita-Juárez YH, Phillips D, Biermann J, Bloom-Ackermann Z, Barkas N, Ganna A, Gomez J, Melms JC, Katsyv I, Normandin E, Naderi P, Popov YV, Raju SS, Niezen S, Tsai LT-Y, Siddle KJ, et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* 2021; Available from: http://dx.doi.org/10.1038/s41586-021-03570-8

29. Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, Debnath O, Thürmann L, Kurth F, Völker MT, Kazmierski J, Timmermann B, Twardziok S, Schneider S, Machleidt F, Müller-Redetzky H, Maier M, Krannich A, Schmidt S, Balzer F, Liebig J, Loske J, Suttorp N, Eils J, Ishaque N, Liebert UG, von Kalle C, Hocke A, Witzenrath M, Goffinet C, et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis [Internet]. Nature Biotechnology 2020.

30. Melms JC, Biermann J, Huang H, Wang Y, Nair A, Tagore S, Katsyv I, Rendeiro AF, Amin AD, Schapiro D, Frangieh CJ, Luoma AM, Filliol A, Fang Y, Ravichandran H, Clausi MG, Alba GA, Rogava M, Chen SW, Ho P, Montoro DT, Kornberg AE, Han AS, Bakhoum MF, Anandasabapathy N, Suárez-Fariñas M, Bakhoum SF, Bram Y, Borczuk A, Guo XV, et al. A molecular single-cell lung atlas of lethal COVID-19. *Nature* 2021; Available from: http://dx.doi.org/10.1038/s41586-021-03569-1.

31. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, Talavera-López C, Maatz H, Reichart D, Sampaziotis F, Worlock KB, Yoshida M, Barnes JL, HCA Lung Biological Network. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* 2020; 26: 681–687.

32. Muus C, Luecken MD, Eraslan G, Sikkema L, Waghray A, Heimberg G, Kobayashi Y, Vaishnav ED, Subramanian A, Smillie C, Jagadeesh KA, Duong ET, Fiskin E, Triglia ET, Ansari M, Cai P, Lin B, Buchanan J, Chen S, Shu J, Haber AL, Chung H, Montoro DT, Adams T, Aliee H, Allon SJ, Andrusivova Z, Angelidis I, Ashenberg O, Bassler K, et al. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat. Med.* 2021; 27: 546–559.

33. Pan H, Deutsch GH, Wert SE, Ontology Subcommittee, NHLBI Molecular Atlas of Lung Development Program Consortium. Comprehensive anatomic ontologies for lung development: A comparison of alveolar formation and maturation within mouse and human lung. *J. Biomed. Semantics* 2019; 10: 18.

34. Sountoulidis A, Liontos A, Nguyen HP, Firsova AB, Fysikopoulos A, Qian X, Seeger W, Sundström E, Nilsson M, Samakovlis C. SCRINSHOT enables spatial mapping of cell states in tissue sections with single-cell resolution. *PLoS Biol.* 2020; 18: e3000675.

35. Gyllborg D, Langseth CM, Qian X, Choi E, Salas SM, Hilscher MM, Lein ES, Nilsson M. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Research 2020. p. e112–e112.

36. Asp M, Bergenstråhle J, Lundeberg J. Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. *Bioessays* 2020; 42: e1900221.

37. Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat. Methods* 2020; 17: 101–106.

38. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, Avraham-Davidi I, Vickovic S, Nitzan M, Ma S, Buenrostro J, Brown NB, Fanelli D, Zhuang X, Macosko EZ, Regev A. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram bioRxiv 2020 [cited 2021 Jul 9]. p. 2020.08.29.272831.

39. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M, Sandberg R, Deng Q, Hedlund E. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.* 2016; 7: 12139.

40. Sachs N, Papaspyropoulos A, Zomer-van Ommen DD, Heo I, Böttinger L, Klay D, Weeber F, Huelsz-Prince G, Iakobachvili N, Amatngalim GD, de Ligt J, van Hoeck A, Proost N, Viveen MC, Lyubimova A, Teeven L, Derakhshan S, Korving J, Begthel H, Dekkers JF, Kumawat K, Ramos E, van Oosterhout MF, Offerhaus GJ, Wiener DJ, Olimpio EP, Dijkstra KK, Smit EF, van der Linden M, Jaksani S, et al. Long-term expanding human airway organoids for disease modeling. *EMBO J.* 2019; 38.

41. Ruiz García S, Deprez M, Lebrigand K, Cavard A, Paquet A, Arguel M-J, Magnone V, Truchi M, Caballero I, Leroy S, Marquette C-H, Marcet B, Barbry P, Zaragosi L-E. Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* 2019; 146.

42. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, Sarkans U, Brazma A. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* 2019; 47: D711–D715.

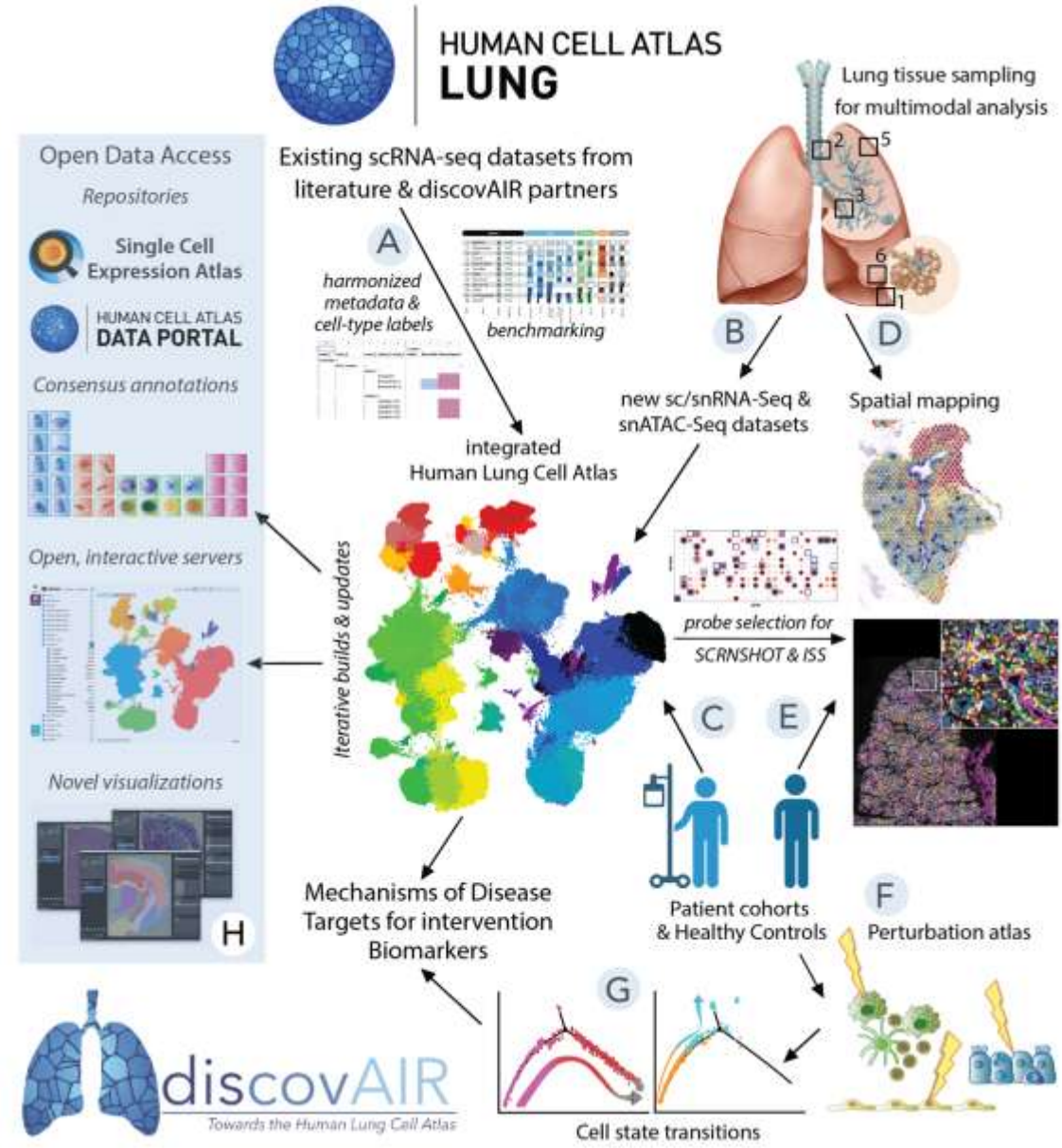

Figure 1 - discovAIR approach and workflow

The discovAIR consortium aims to establish a first draft of the Human Lung Cell Atlas, by integrating existing single-cell RNA seq datasets from the HCA Lung Biological Network into a single embedding, allowing analyses on the integrated dataset (A). This dataset will be enriched by additional datasets generated by the discovAIR consortium: a multi-omics characterization of lung tissue from 5 healthy donor lungs sampled at 5 locations ('deep dive'; B) and a cohort of healthy controls (at least n=50) and patients with lung disease (at least n=50; C). The 'deep dive' tissue samples will also be used to generate spatially resolving datasets on matched tissue sections using Visium, *in situ* sequencing, SCRINSHOT and LCM-guided snRNA-seq analysis (D). Samples from the discovAIR cohort (patients & controls) will be used for spatial mapping using SCRINSHOT (E) and for establishment of the perturbation atlas (F). In the perturbation atlas, stimulations of *ex vivo* cultured primary cells or precision-cut lung tissue slices will be used to map cell state trajectories of healthy cells to diseased cell states (G). These novel datasets will be ingested into next iterations of the integrated lung cell atlas, used to establish consensus annotations for the different lung cell types and submitted to the appropriate data repositories (H) as open data using novel visualization modalities for the spatial

datasets. Finally, the integration of the lung cell atlas (including datasets from lung disease) and the perturbation atlas will generate novel insights into mechanisms of disease, help guide identification of drug targets and biomarkers for disease inception or treatment response.