

A cellular reference atlas of the human lung

The Human Lung Cell Atlas (HLCA), which includes data from 36 studies, is the most comprehensive representation of cellular gene expression in the human respiratory system to date. The HLCA serves as a reference for future cellular studies of the lung, enabling a better understanding of lung biology in health and disease.

This is a summary of:

Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* <https://doi.org/10.1038/s41591-023-02327-2> (2023).

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 12 June 2023

The mission

Technical advances over the past decade have enabled researchers to study tissues at the resolution of single cells. Measuring the gene expression of thousands of cells simultaneously has led to an improved understanding of cellular heterogeneity in the lung¹ and the discovery of novel cell types, such as the pulmonary ionocyte². However, there is still a lack of consensus between studies on cell type definitions. Furthermore, owing to cost and time limitations, individual studies often include only a small number of individuals, which limits our understanding of how cell type profiles differ between individuals. In this context, the Human Cell Atlas (HCA) announced the goal of creating a central reference of human cellular biology³. With many available single-cell datasets of the human lung to choose from, which dataset is most suited to this task?

The solution

Rather than choosing a single dataset, recent advances in computational methods enable us to combine multiple datasets into one integrated single-cell reference. These data integration methods remove technical differences between studies while preserving meaningful biological variation in the data. To choose which of the currently available methods is best suited to build an integrated lung cell reference, we compared 12 popular methods with our recent benchmarking pipeline⁴. Using the best performing method, we integrated 14 high-quality lung datasets to create the core of the HLCA (Fig. 1). We then analyzed and reannotated the HLCA with input from the HCA lung community to propose the first consensus-based annotation of the human lung. The large scale of the atlas, including data of healthy lung tissue from more than 100 individuals, enabled us to characterize the diversity in cellular profiles between individuals.

We identified 61 cell types in the human respiratory system, each detected in multiple datasets, including rare cell types that had not been previously identified. Leveraging the diversity of individuals included in the atlas, we found cell-type specific effects of demographic variables such as age, sex and BMI. For example, individuals with higher BMI showed an increase in inflammatory gene expression programs in alveolar macrophages, which characterizes a cellular state that may be involved in

chronic obesity-associated systemic inflammation. With recent advances in transfer learning methods⁵, we showed that the HLCA can be used as a reference for mapping and annotating new data and for the automated identification of disease-specific cell states. Extending the HLCA with data from more than 10 different diseases, we identified altered cellular states common to multiple diseases, such as a subset of lung monocyte-derived macrophages observed in COVID-19, lung carcinoma and pulmonary fibrosis, which are likely involved in the process of scar formation across all three diseases.

The implications

Similar to the first draft of the human genome project, the HLCA is a reference for single-cell studies of lung tissue that will change the way we analyze new lung data. Instead of starting an analysis from scratch, we can now map new data to the HLCA for the rapid analysis and annotation of new data and for highlighting cell states in these data that differ from healthy cells in the atlas. In addition, just as genome annotations evolved over time, the HLCA will enable a similar process by providing a framework for discussing and challenging consensus on cell type annotations on the basis of integrated datasets.

Although we have mapped over 30 datasets from different technologies to the HLCA, approaches for mapping new data to the HLCA still have a way to go to reach the usability of genome-alignment methods. Currently, mapping data from model systems such as organoids and cell lines remains a challenge, and mapped data still need manual quality checks. Furthermore, the HLCA is fundamentally an observational study and resource. Although we can characterize cell states shared between diseases, further experiments are needed to draw conclusions on common causal disease mechanisms.

The HLCA cannot be static: cell type consensus will evolve as we explore spatial organization and epigenetic profiles of lung cells. And if we can do this for the lung, why stop there? As part of the HCA integration team, we are now expanding the HLCA cookbook to build references for further tissues and organs.

Lisa Sikkema¹ & Malte D. Luecken^{1,2}

¹Institute for Computational Biology, Helmholtz Munich, Neuherberg, Germany.

²Institute for Lung Health & Immunity, Helmholtz Munich, Neuherberg, Germany.

EXPERT OPINION

“This manuscript presents a comprehensive single-cell atlas of the human lung on the basis of several datasets. The manuscript implements a detailed protocol for batch corrections, filtering, annotations and analyses of biological and technical covariates and of cell types potentially

associated with disease genes. I believe it will be an important resource for lung biologists and for the single-cell community in general, as many of the methodologies will be applicable to other HCA efforts.”
Shalev Itzkovitz, Weizmann Institute of Science, Rehovot, Israel.

FIGURE

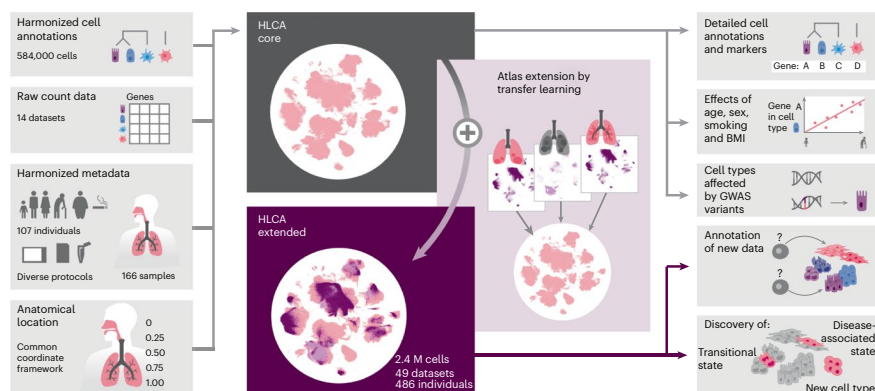


Fig. 1 | Building and using the HLCA. Left: the HLCA core was built from single-cell RNA sequencing count data from 14 datasets and includes detailed donor and sample metadata. Middle: using transfer learning, 35 more datasets were mapped to the HLCA. Right: the HLCA includes cell annotations and marker genes and can be used to model effects of demographics on cell types, link disease genetic information to cell types, annotate new data and discover unknown and disease-affected cell types. © 2023, Sikkema, L. et al., CC BY 4.0.

BEHIND THE PAPER

The HCA, a consortium that aims to construct reference atlases for human tissues and organs, has proven to be an incredibly fruitful basis on which to build the HLCA. Before the start of the SARS-CoV-2 pandemic we had begun collecting datasets to build the HLCA. The pandemic then brought new questions to the forefront: which cell types are being infected first and does the susceptibility to infection differ with age, sex or other demographic factors? We put the HLCA on

hold to answer these questions using some of the data we had collected and much more unpublished data that was shared by the community. Although delaying the HLCA timeline, it also showcased the power of data sharing and collaboration to answer big questions. In the HLCA, we made use of this willingness to collaborate and share that had been kindled by the pandemic. Overall, this sharing of data and expertise enabled us to provide a true community reference atlas. **M.D.L. & L.S.**

REFERENCES

1. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
One of the earliest single-cell RNA sequencing papers systematically profiling the healthy human respiratory system.
2. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
One of the two papers that first described the pulmonary ionocyte, which was discovered using single-cell data.
3. Regev, A. et al. The Human Cell Atlas white paper. Preprint at <https://doi.org/10.48550/arXiv.1810.05192> (2018).
The HCA white paper, describing its goals and outlining how those goals can be achieved.
4. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
A systematic benchmark study of single-cell data integration methods, used for selecting the HLCA integration method.
5. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
Transfer learning tool for mapping query datasets to a reference atlas, used for extending the HLCA with 35 additional datasets.

FROM THE EDITOR

“This study was of interest to us both for the resource value and because it presents a blueprint for generating integrated single-cell atlases that can be applied to other organs. By integrating single-cell RNA sequencing data from more than 400 individuals and different lung regions, it enables the identification of transcriptional programs associated with demographic covariates and anatomical location within the respiratory system.” **Editorial Team, Nature Medicine.**