nature methods

Review article

https://doi.org/10.1038/s41592-024-02532-y

Considerations for building and using integrated single-cell atlases

Received: 31 October 2023

Accepted: 22 October 2024

Published online: 13 December 2024

Check for updates

Karin Hrovatin[1,2,14], Lisa Sikkema[1,2,14], Vladimir A. Shitov[1,3], Graham Heimberg[4,5], Maiia Shulman[1,2], Amanda J. Oliver[6], Michaela F. Mueller[1], Ignacio L. Ibarra[1], Hanchen Wang[5,7], Ciro Ramírez-Suástegui[1,6], Peng He[8], Anna C. Schaar[1,9], Sarah A. Teichmann[6,10,11,12], Fabian J. Theis[1,2,13] ✉ & Malte D. Luecken[1,3] ✉

The rapid adoption of single-cell technologies has created an opportunity to build single-cell 'atlases' integrating diverse datasets across many laboratories. Such atlases can serve as a reference for analyzing and interpreting current and future data. However, it has become apparent that atlasing approaches differ, and the impact of these differences are often unclear. Here we review the current atlasing literature and present considerations for building and using atlases. Importantly, we find that no one-size-fits-all protocol for atlas building exists, but rather we discuss context-specific considerations and workflows, including atlas conceptualization, data collection, curation and integration, atlas evaluation and atlas sharing. We further highlight the benefits of integrated atlases for analyses of new datasets and deriving biological insights beyond what is possible from individual datasets. Our overview of current practices and associated recommendations will improve the quality of atlases to come, facilitating the shift to a unified, reference-based understanding of single-cell biology.

Understanding the cellular composition of tissues and its variability across individuals is critical for understanding health and disease. Single-cell technologies have spurred important progress in our understanding of cellular heterogeneity by enabling researchers to study tissues at unprecedented scale and resolution[1–3]. However, while the number of single-cell datasets and the number of cells sequenced per study steadily increase, currently the median number of individuals sampled per study still does not exceed 14 (Fig. 1). Moreover, individual studies have study-specific biases related to, for example, cohort characteristics, sample handling and choice of single-cell technology. Integrating many studies into a single resource, here termed 'atlas', enables researchers to overcome these study-specific biases as well as to capture a larger number of individuals and more comprehensively profile cellular diversity.

A number of research initiatives, including the Human Cell Atlas (HCA)[4] and the Human Biomolecular Atlas Program (HuBMAP)[5], aim to create such single-cell atlases of the human body. Currently available

[1]Department of Computational Health, Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany. [2]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [3]Comprehensive Pneumology Center (CPC) with the CPC-M bioArchive / Institute of Lung Health and Immunity (LHI), Helmholtz Zentrum München; Member of the German Center for Lung Research (DZL), Munich, Germany. [4]Department of OMNI Bioinformatics, Genentech, South San Francisco, CA, USA. [5]Department of Biological Research | AI Development, Genentech, South San Francisco, CA, USA. [6]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [7]Department of Computer Science, Stanford University, Palo Alto, CA, USA. [8]Department of Pathology, University of California, San Francisco, San Francisco, CA, USA. [9]TUM School of Computation, Information and Technology, Technical University of Munich, Garching, Germany. [10]Theory of Condensed Matter Group, Department of Physics, Cavendish Laboratory, University of Cambridge, Cambridge, UK. [11]Cambridge Stem Cell Institute and Department of Medicine, University of Cambridge, Cambridge, UK. [12]CIFAR MacMillan Multiscale Human Programme, Toronto, Ontario, Canada. [13]Department of Mathematics, Technical University of Munich, Garching, Germany. [14]These authors contributed equally: Karin Hrovatin, Lisa Sikkema. ✉e-mail: fabian.theis@helmholtz-munich.de; malte.luecken@helmholtz-munich.de

https://doi.org/10.1038/s41592-024-02532-y

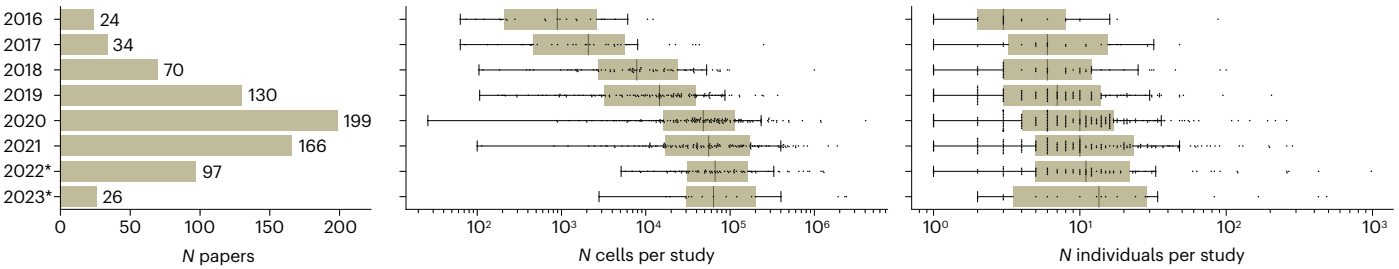

Fig. 1 | Single-cell dataset size trends over time. Left, total number of papers with single-cell data published in each year; middle, number of cells per dataset over time; right, number of individuals included per single-cell paper (Supplementary Methods). The list of publications was obtained from a curated database of single-cell studies[194]. Data after 2021 (asterisks) are likely not comprehensive; thus, the number of papers is likely underestimated. For both box plots, the boxes indicate the median and interquartile range. Whiskers extend to the furthest non-outlier data point. Individual data points are shown as dots.

BOX 1

Key quality standards for atlases

- Represent consensus cell-type nomenclature across the field
- Findings should be based on observations that hold true across datasets
- Availability of high-quality data and metadata
- Reliability ensured by stringent quality assessment

single-cell atlases[6–34] cover various tissues from mice, humans or both, and are almost exclusively composed of transcriptomic data (see Supplementary Table 1 for an overview of available atlases and their characteristics). These atlases have been used to address a range of biological questions and research challenges. For example, they serve as a consensus on cell-type definitions and disease-specific cell states[9,15–17], reveal heterogeneity in the population at a large scale[11,17,35], aid in the analysis of new datasets[11,13,15,17] and give guidance in study design[6,12,17]. Existing single-cell atlases thus already show their promise of advancing our understanding of human tissue in health and disease[36].

To serve as a community resource, integrated single-cell reference atlases should adhere to specific criteria. First, an atlas is meant to serve as a basis for community discussion on cell-type nomenclature and should, therefore, represent the current cell-type definition consensus as well as possible. Second, findings derived from the atlas need to be generalizable and should represent a consensus across studies, which requires the inclusion of numerous and diverse datasets encompassing a large set of individuals. Thus, recent large-scale efforts based on individual datasets[37–39] will in the future present a great basis for building multi-dataset atlases. Generalizability also requires that dataset-related or sample-related biases, such as data generation location and protocols, are documented and if possible, removed from the atlas. Third, atlases need to include extensive cell annotations and sample and subject metadata for future studies and analyses. Finally, atlases should be reliable, such that findings derived from the atlas are not based on artifacts in the data or mistakes in annotations, mandating stringent quality assessment. Such atlases could thus, similarly to reference genomes and other omics references[40–42], serve as the 'normal' basis against which new datasets are compared to address central questions in molecular biology and medicine (Box 1).

Despite the complexity and demands of the atlas building process, a clear overview of atlas building steps and associated considerations is currently lacking. Moreover, the potential applications of atlases in research are just starting to be explored. Here we review how previous reference atlases have been built and lay out guiding principles for the construction and sharing of future atlases to ensure quality and broad opportunities for application. We also provide a perspective on how atlases may be extended and updated in the future to stay up to date with new discoveries. Finally, we present a comprehensive overview of atlas use cases. Together, we envision this work will advance the progress of atlas-focused initiatives such as the HCA, HubMAP and others, thus contributing to moving the single-cell field toward cross-dataset, population-level reference atlases.

Building an integrated reference atlas

Building an integrated atlas requires biological and computational domain expertise and iterative optimization of the atlas. This process can be categorized into the following steps (Fig. 2), which are discussed in detail below: preparation, including choosing the focus and selecting datasets; data preprocessing, including metadata harmonization and quality control; data integration; atlas evaluation and reannotation; and atlas sharing and extension.

Atlas preparation

The envisioned downstream use of an atlas determines what technical decisions should be made when building it, and the atlas' goals should therefore already be taken into account during the preparation. For example, if one wants to build an atlas that enables modeling the effects of age on molecular phenotypes, it should ideally include pediatric samples. It is thus critical to determine the focus of the atlas before starting the building. Similarly, the included datasets should be selected carefully to align with the atlas focus and to maximize its quality. Below, we discuss important considerations in both of these processes.

Defining the focus. It might be desirable to make an atlas as general as possible, integrating data across technologies, organs or species. However, this may ultimately reduce its utility, as the removal of strong batch effects often also leads to excessive loss of biological variation[43]. Instead, the focus of the atlas must be chosen at the beginning to ensure that the final atlas will best be suited for the envisioned downstream applications. Whereas most atlases aim for a holistic understanding of a single organ, cell-type-specific atlases provide insights into cell-type-specific diseases affecting multiple tissues[6]. Moreover, while some atlases are focused only on healthy adult samples (Supplementary Table 1), the inclusion of multiple conditions, such as diseases or developmental stages, is crucial for cross-condition comparison. Similarly, atlases that include animal or in vitro model systems are vital for evaluating model utility, and can additionally be used to complement scarce human data. Finally, multi-omic atlases may increase resolution and reliability via a broader set of molecular features. We provide further guidance on defining atlas focus in Supplementary Note 1.

Selecting datasets. Once the goal of the atlas is clear, the datasets to be included must be selected, which importantly determine the atlas' quality and utility. For some atlases, data scarcity necessitates leniency during dataset selection, while in other cases, not all datasets that fit

https://doi.org/10.1038/s41592-024-02532-y

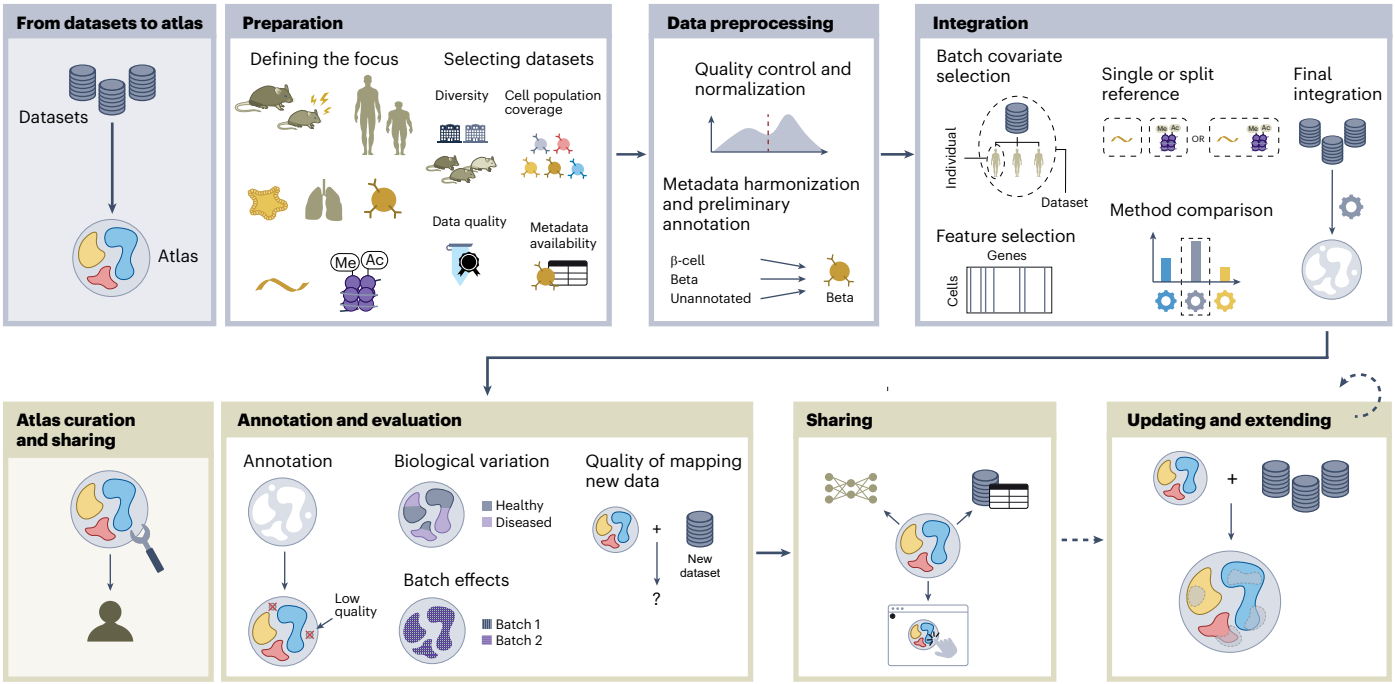

Fig. 2 | Workflow for building reference atlases. The atlas building process consists of constructing an atlas from individual datasets (top; including data preparation, preprocessing and integration), and atlas curation and sharing once the atlas is constructed (bottom; including atlas annotation and evaluation, sharing, and updating and extending the data captured within the atlas).

the goal of the atlas are suitable for inclusion. Here we discuss key considerations for selecting datasets.

Number and technical diversity of datasets. Including a large number of datasets is important not only to capture the variability in cell types and in tissue phenotypes, but also to cover a diverse range of technical variability, such as sample handling protocols and sequencing technologies. This will likely broaden the range of cell types covered, as different technologies better capture different cell types[17,44], will provide integration methods with more training data to distinguish between biological and batch effects and will enable assessing reproducibility of findings across studies during downstream analysis. A large number of datasets also allows for the removal of datasets that later on turn out not to integrate well without substantially reducing the size of the atlas. However, increasing the number of datasets will lead to longer data curation and preprocessing times and might eventually lead to prohibitive computational resource requirements.

Metadata availability. Sample-level metadata (such as the age or health status of individuals) and cell-level metadata (such as cell-type labels) are important for many steps in the atlas building process, as well as for its use. Metadata related to technical variables such as tissue sampling technique (for example, samples from autopsy and biopsy) will help distinguish experimental bias from biologically driven signal during atlas building and evaluation. Moreover, detailed donor metadata (for example, age, sex, body mass index, smoking history or disease stage and treatment history) will make the atlas more widely usable, particularly for understanding interindividual differences. Similarly, the availability of cell-level metadata, mainly cell-type labels, can aid in the atlas building process in several ways, including during quality control (for example, labels of doublets), integration (for methods using cell-type labels[17,43,45,46]) and atlas evaluation and use.

Demographic diversity. For a truly comprehensive representation of a given tissue, an atlas should cover the diversity of the human population in terms of age, sex, genetic ancestry and other demographic variables. The same holds for other types of biological variation, such as multiple mouse models or strains and various environmental conditions. Increasing the diversity of samples makes atlas-based findings more generalizable, and might enable stratification into, for example, patient groups.

Cell-type coverage. An atlas ideally represents the full diversity of cell types that are part of the organ or cellular compartment of interest, which can be achieved by diversifying the selected datasets in several ways. When cell-type composition differs widely in different parts of an organ, different anatomical locations should be covered. Similarly, the way a tissue is sampled (for example, with a brushing, biopsy or surgical resection) affects the cell types that are captured in a sample, as do dissociation protocols, with certain cell types only being detected using specific protocols[11,17,47]. Spatial assays, not biased by tissue dissociation protocols, can be used to determine which cell types should be detected in a tissue.

Study design. When building an atlas, untangling batch-related variables from biological signals of interest is key to fruitful downstream analyses. However, without proper study design, batch correction methods can unintentionally remove biological signals along with batch effects. For example, if one dataset is made up only of samples from donors with a disease not seen in any other dataset, integration could inadvertently mistake this disease-specific biology for a dataset-specific batch effect and remove this signal from the integration. Therefore, ideally each biological group of interest is represented in multiple datasets, and each dataset includes multiple conditions so that dataset-specific batch effects can be separated from effects of interest.

Data quality. Including datasets of high quality in the atlas will enhance the reliability of downstream analyses and will moreover ease the data integration process. Datasets can differ substantially in quality, for example, in terms of sequencing depth, fraction of mitochondrial reads per cell and detail of cell-type annotation. Datasets with lower cellular resolution, for example, due to low sequencing depth, have been shown to integrate more poorly with other datasets[48] (Box 2).

BOX 2

Key takeaways

- A specific atlas focus should be defined, such as a selected tissue and selected biological conditions. The broader the focus of an atlas, the more complex the integration process is likely to be, which can result in increased loss of information during integration.
- The datasets to be included in the atlas should be selected according to its focus.
- Biological conditions of interest should ideally be represented in multiple datasets, to untangle biological from batch effects during atlas integration and evaluation. Similarly, datasets should ideally span multiple conditions (for example, disease and control).
- Sample-level and cell-level metadata availability should be considered for each dataset as it is key for atlas building, evaluation and use.

Data harmonization and preprocessing

For datasets to be jointly analyzed, the metadata of subjects, samples and cells that are provided with each dataset must be encoded consistently across datasets and count data should be computationally preprocessed in similar or identical ways. Individual datasets are typically preprocessed in different ways and use different metadata nomenclature, contributing to batch effects and hampering downstream interpretation, respectively. While metadata must always be unified, it is not yet clear how specific differences in preprocessing affect the final atlas. Below we discuss considerations in prioritizing different atlas building steps for harmonization.

Data preprocessing. Preprocessing of FASTQ sequencing data into count matrices is the first step in preprocessing single-cell sequencing data. Atlases are often built from count matrices rather than raw FASTQ files, as these are easier to share and combine, but inconsistencies in how the count matrices were generated can lead to batch effects in the data[49]. Whereas realignment of all data might not always be feasible, using gene identifiers rather than gene names could already mitigate batch effects. Once count matrices have been generated, low-quality droplets need to be removed from the data. In some cases, this will already have been done by dataset generators, and the exact method by which low-quality droplets have been removed from the data might result in differences between datasets. Moreover, the decision to instead only annotate but not remove low-quality droplets from the atlas can enable quality-control transfer to new datasets, although it may also affect the integration. Finally, counts should be normalized and corrected for, for example, ambient RNA, in the same way across datasets, taking into account suitability and scalability of the normalization methods. Preprocessing-related considerations are further detailed in Supplementary Note 2.

Harmonizing sample and subject metadata. Sample and subject metadata are essential for both the atlas building process and for downstream analyses on the atlas. However, often inconsistent nomenclatures across datasets can make them challenging to use. Therefore, all metadata from individual datasets should be mapped to standardized categories. For many forms of metadata, standardized nomenclature already exists in the form of ontologies, such as for disease[50], ancestry[51] or single-cell protocols[52]. As these existing ontologies have been constructed by specialists, adhering to them will in most cases give a better classification and naming system than a manually set-up categorization would. Moreover, it will ease future communication and comparison within the field. Single-cell data platforms such as CELLxGENE[53] and the HCA Data Repository[54] already conform to these ontologies.

For human metadata, one should ideally also track if these data are based on self-report, assignment by physicians or genomic information. As standards regarding ethnicity and genetic ancestry categorization are still evolving, it is useful to collect these data in 'raw' form before harmonizing them into predefined, possibly broader groups. Alternatively, data-driven methods for sex or ancestry inference from raw sequencing data[55] can complement the reported metadata.

Harmonizing cell-type annotations and annotating unlabeled datasets. Preliminary, author-provided, cell-type annotations can be beneficial in many ways. Firstly, they can help with the data integration itself, as some data integration methods allow for the use of cell-type labels to guide the integration[45,46,56]. Secondly, these labels aid in evaluating the quality of the atlas once the data are integrated, although it must be ensured that the same labels are not used for both integration and its evaluation to prevent evaluation biases. Thirdly, the comparison with original labels enables evaluating the impact of the consensus reannotation in an atlas (see section 'Exploring the information within the atlas').

For the above tasks, it is crucial to have a consistent set of cell-type labels. As studies are often inconsistent in cell-type nomenclature and annotation resolution[16,17], it is helpful to map all annotations to a common, cell-type nomenclature reference 57. When original annotations are of sufficient resolution, this may already result in a good-quality preliminary atlas annotation[10,12,17]. The cell-type reference can be 'hierarchical', thus accommodating annotations from different datasets at different resolutions. As manual harmonization of cell-type labels is laborious, automated construction of such cell-type hierarchies could aid in the process, for example, using CellHint[58]. Furthermore, community resources such as the Cell Annotation Platform (CAP[59]) are being developed to facilitate author-guided, consensus-based cell annotation and to define label synonyms for cell types and states.

If author-provided, cell-type annotations are unavailable or of insufficient quality, one can annotate the data specifically for the atlas[60,61]. Because manual annotation of each dataset individually is very time-consuming, it can instead be done for only a representative dataset subset[15]. Alternatively, automated annotation[62,63] combined with manual curation can be used to annotate individual datasets before integration[64,65], which can be combined with preexisting annotations[24] (Box 3).

Data integration

The core of any atlas building project is the integration of the data, involving the computational removal of batch-related variation, which can be attempted with a variety of methods[43]. Integration enables joint analysis of all data in a shared space, based on biological signals rather than batch-specific transcriptomic artifacts. Below we describe several important aspects of atlas-level data integration.

Determining the level of integration by setting the batch covariate. All data integration methods aim to identify and subsequently remove batch-specific transcriptomic shifts based on a predefined batch covariate. The choice of batch covariate will greatly affect which variation is removed from the atlas and how the integrated atlas will look. It should be noted that the variation determined as 'batch effect' is inherently subjective and is a reflection of what the atlas builder deems unwanted. Thus, the choice of batch covariate should be in line with the scope of the atlas.

Batch effects can occur at the level of the dataset, subject or even sample. Many experimental and preprocessing factors vary predominantly at the dataset level, such as tissue dissociation protocol, single-cell chemistry or reference genome. Therefore, batch correction at the dataset level can already remove a large part of the variation

BOX 3

Key takeaways

- Differences in dataset preprocessing could lead to batch effects and may be mitigated by preprocessing harmonization. Furthermore, some preprocessing steps may differ for atlases compared to standard analysis protocols, such as preservation of low-quality droplets for annotation transfer.
- As high-quality metadata are key for atlas use, the metadata should be harmonized across datasets and aligned with standard ontologies.
- Preliminary cell-type annotation is an important source of prior knowledge in atlas building as it can guide integration and is often required for atlas evaluation. It can be obtained by harmonizing available annotations across datasets, transferring annotations from individual datasets, or de novo manual or automated annotation.

caused by technical factors. Even when all dataset-level batch effects are removed from the data, additional sources of transcriptomic variation due to artifacts can exist at the sample and subject levels. For example, samples might have undergone different sample handling causing distinct transcriptomic changes, and cells that are sequenced on the same lane, which can originate from a single tissue sample or multiple tissue samples (for example, when multiplexing), may have specific technical effects. Moreover, individuals might display subject-specific batch effects, for example, due to different postmortem intervals[66]. To remove these sources of batch effects, sample and/or subject would need to be set as the batch covariate. Notably, some methods enable the use of multiple batch covariates at once, thus enabling nested (dataset–sample) or combinatorial (dataset–assay) batch effect designs[46,67].

The extent to which technical covariates contribute to batch effects can vary. For example, if a study has samples generated in multiple institutes, they might or might not show institute-specific batch effects. Similarly, batch covariates may affect individual cell types with different strengths[68]. The covariate contribution to batch effects can be approximated quantitatively by computing the percentage of variance explained by a particular technical covariate[17,69]. Alternatively, one can check which sample groups are easily distinguished from the rest of the dataset using a classifier, thus determining whether and how a dataset should be split up into separate batches[12]. Recent efforts have also attempted to automate the batch selection procedure[68].

Pitfalls in the choice of batch covariate arise predominantly due to the possibility of removing biological signals during the integration process, as they can covary with the chosen batch covariate. A particular challenge arises when batches directly correspond to biological variables of interest. This is the case when integrating datasets of, for example, different organoid protocols, species or organ locations, or samples of patients with different diseases. Here, removing dataset-level or sample-level variation could remove the related variance of interest (for example, protocol-specific states) from the integrated atlas, hampering downstream analyses. In such cases, one may select a coarser batch covariate that is not directly confounded by the covariate of interest, such as dataset instead of sample. It is important to note, however, that some existing atlases show biological preservation of sample-level variation even when using sample as the batch covariate[11,15].

Selection of genes for data integration. Similar to other single-cell RNA-sequencing (scRNA-seq) dimensionality reduction techniques, data integration can benefit from being performed on a subset of genes. Benefits range from improving the signal-to-noise ratio, removing non-informative signals given the atlas scope, and improved computational efficiency, resulting in improved integration[43]. However, when removing genes, one has to keep in mind that existing integration models cannot be adapted to add features (genes) later on, such as features that may be important to future samples mapped to the atlas.

Currently, the most common practice for gene selection is selecting 'highly variable genes', that is, genes that show higher variance than would be expected based on their mean expression levels in the data[70,71]. Most atlases select 2,000 to 5,000 genes[10,15,16,24], with higher numbers preferred for atlases with broader scope[13]. Often, genes are selected as the intersection of genes that vary within individual batches to avoid selecting genes varying due to batch effects[72,73]. Several methods that aim to improve the robustness or biological meaningfulness of gene selection have also been proposed[72–77]. These include removing batch-affected or quality-metric-associated genes and selecting genes related to signals of interest such as individual cell lineages, rare cell types or diseases (Supplementary Note 3). Given the diversity of proposed approaches, there is likely a large potential for optimizing atlas building in this step.

Selecting an optimal data integration strategy. The choice of integration method and its parameters will have a substantial effect on the outcome of the integration[12,17,43,78]. As different integration methods work best in different scenarios[43], it is important to select the optimal method and parameter settings for the data at hand, for example, using existing integration benchmarking platforms[43]. The methods scANVI[45], Scanorama[79] and scVI[67] have been shown to perform well on complex integration tasks[43] and could thus be prioritized if time constraints do not allow for extensive method benchmarking. Moreover, integration strategies for multimodal data are discussed in Supplementary Note 4.

The integration process also involves a decision on which of the available data to include in a single integrated representation. In some cases, batch effects are too strong to be removed while still preserving the desired level of biological information. In this case, the atlas may be split into multiple parts, for example, to create one sub-atlas per species[6,9], per lineage[80] or for cells versus nuclei[26]. Notably, recent efforts have been devoted toward facilitating the integration of more biologically and technically diverse datasets[81,82]. It is yet to be determined how global integration strategies compare to a split approach. While atlases consisting of multiple integrated representations could provide better resolution, they are less user-friendly, requiring separate analyses and comparisons for each sub-atlas.

Several atlases have taken the approach of separating their data into core datasets that are normally integrated, and extension datasets that are mapped onto the core[8,11,12,17] using query-to-reference mapping methods[83–85]. These methods make it possible to project new data onto an existing reference while removing batch effects from the resulting representation. In some cases, it may be desirable to separate core and extension datasets based on biology, for example, by using only healthy adult data in the reference to ease the learning of batch effects during integration with minimal biological confounding[17]. The core-extension approach also allows more flexibility in adapting and extending an atlas. Given a fixed core reference, datasets can be independently mapped onto the core. However, as the atlas model core is never updated with the newly mapped data, new biological variation in these datasets may not be sufficiently captured, or the model might not be able to sufficiently remove new batch effects. Additionally, most currently available query-to-reference mapping methods are designed to work only with selected integration methods or models (for example, Symphony with Harmony[86], scArches with conditional autoencoder-based methods[83]). These compatibilities should be kept in mind when selecting a data integration method to ease mapping to the atlas by future users ('Projecting new single-cell data into an atlas space').

Integration method performance can differ widely and thus the best integration should be selected for a given collection of datasets.

BOX 4

Key takeaways

- Data integration is the key step of any atlas building project.
- The choice of batch covariate importantly affects the integration outcome. Confounding of batch and biological covariates may lead to the removal of relevant variation from the data during integration.
- Gene selection helps to reduce noise in the data and limits the computational resources required for integration. Several gene selection approaches exist to improve the outcome of the integration.
- Gene selection must be performed with future query datasets in mind, as these might contain unique condition-specific genes.
- Atlases can exist as a single integration of all datasets, multiple partial atlases from datasets that are easier to integrate separately or a core integrated atlas from selected datasets extended with additional data via reference mapping.
- Integration approaches should be compared using metrics that assess both batch effect removal and the preservation of biological variation.

Importantly, visual inspection of the result of integration (for example, using uniform manifold approximation and projection (UMAP)) to assess performance can be misleading[43,87] and hard to apply to large amounts of integration outputs. Therefore, one should combine visual inspection with quantitative metrics to assess the integration quality. Several metrics aim to quantify either how well batch effects were removed or how well biological variation was retained during data integration[43]. As different metrics measure different aspects and resolutions of the integrated representation and have their specific benefits and limitations[43,69,82,88], it is important to consider which metrics to use. For example, both metrics that rely on prior biological knowledge (for example, cell labels) and those that are independent of it should be included. Notably, performing a benchmark on data integration methods for an individual atlas is resource intensive. Therefore, for atlases with a large number of cells, a subset of the data may be used for the integration method benchmark. We provide further details on integration benchmark metrics and on data subsetting in Supplementary Note 5 (Box 4).

Atlas evaluation and reannotation

The quality of an atlas is critical for its utility as a reference. Because automated evaluation of integration methods, as described above, provides no guarantees on the top-performing method being of sufficient quality for atlas use, it should be complemented with manual atlas evaluation. This also includes atlas-level cell-type reannotation, which serves both for the evaluation of the integrated representation quality and as a basis for downstream analyses.

Evaluation of overall atlas representation quality. The final atlas evaluation must be done on the basis of prior biological knowledge. This ensures that the atlas correctly represents biological information from the data and that batch effects have been sufficiently removed, as discussed below. The evaluation step serves as the last checkpoint in the optimization of the atlas representation (Fig. 3a–d) and may lead to revisiting and adjusting earlier atlas building steps (Fig. 3e–k).

To derive new insights from the atlas, one must ensure that the integration did not remove key biological information from the data (Fig. 3a). This can be evaluated based on the co-occurrence or separation of cells in the integrated representation in relation to known biological factors, such as cell type, age or disease. As the first step, the expected biological effects within the representation should be evaluated based on the presence of clusters corresponding to known cell types. For example, rare and transitioning cell clusters are commonly merged with other populations due to over-integration[17,43]. The integration benchmarking metrics described above can be further used here to highlight cell subsets that show poor integration quality, necessitating further manual exploration.

When analyzing the presence of biology-driven cell states and subtypes within cell-type clusters, caution must be taken not to interpret batch effect-driven separation as biological differences. The separation of cell representations based on specific covariates, such as disease, should therefore be supported across replicate samples and datasets and the cell populations should also be distinguishable by the expression of specific markers.

To ensure that downstream analyses are driven by biological rather than residual batch effect variation, it is necessary to evaluate how well batch effects have been removed from the atlas (Fig. 3b). For a detailed and thorough evaluation of the remaining batch effects in the atlas, the integrated representation needs to be checked for cell separation driven by technical effects. These include sample-specific or dataset-specific clusters that cannot be explained biologically. One way of identifying batch-driven separation of cells is using the correlation of cluster assignment with the expression of known technical effect genes, which will often be sample specific. This includes ambient genes[89,90], genes associated with tissue handling, such as stress genes induced by dissociation and extended processing time[91–94], or, when integrating single-cell and single-nucleus data, genes known to be differentially expressed between the two assays, such as mitochondrial genes[95]. However, it should be noted that these genes can also be involved in biologically relevant processes, such as disease-related cellular changes.

It is possible that the overall quality of an atlas integration is excellent, despite a small subset of samples or subjects, or a single dataset in the atlas not being well integrated, due to stronger batch effects in that data subset. Visual inspection can sometimes already highlight poorly integrated subsets of the atlas. Furthermore, metrics assessing the mixing of batches[17] within cell populations should be used to identify outliers. It is important to pinpoint the reason for reduced integration as it may result from past disease or outlier demographics that warrant distinct localization in the atlas. Several steps can be considered if finding outlier datasets. First, a reintegration without the outlier dataset, subject or sample can be considered, depending on their relative importance to the focus of the atlas (Fig. 3e). Second, reintegrating with a tailored batch covariate (Fig. 3h), method or parameter setting (Fig. 3i) can be considered, such that more emphasis is placed on the removal of outlier batch effects. Third, if the source of the batch effect is clear, adding more datasets of the same type and data reintegration might help mitigate batch imbalance[96].

Even after the removal of outliers and tuning of the integration approach, some batch effects will always remain. The residual batch effects determine how fine the cell annotation resolution can be before cells separate into clusters based on technical rather than biological effects. Therefore, it is essential to keep in mind how this affects the representation and thereby the downstream analyses.

Evaluation of reference quality for mapping new data. As one of the main uses of atlases is the analysis of new datasets with the atlas serving as a reference, the atlas must be suitable for high-quality alignment of the new data to the atlas via 'query-to-reference mapping' (Fig. 3c). This mapping projects any unseen single-cell dataset into the preexisting low-dimensional space of the integrated atlas, thus allowing joint analysis of the atlas and the new data. Poor reference mapping performance can result in faulty interpretation of the mapped query data. Resolving poor performance may require adapting the integration itself, and can include revisiting previous steps, from dataset selection

https://doi.org/10.1038/s41592-024-02532-y

to integration hyperparameters (Fig. 3e–i), to better capture the range of potential technical and biological effects in the integration itself already. Reference mapping also largely depends on both the used mapping algorithm and the underlying integration method ('Selecting an optimal data integration strategy'), and a different integration method that enables better mapping may be required.

Determining whether an atlas is suited for reference mapping involves considering what kind of datasets may be mapped in the future, with potential differences in technical factors (for example, sequencing protocols, genome versions) as well as biological differences (for example, tissue from donors with diseases, different developmental stages). Importantly, data from very different biological contexts might complicate mapping (Fig. 3k), just as widely different datasets can complicate the initial atlas building (as discussed in 'Defining the focus'). To evaluate the atlas' reference mapping potential, the concepts described in 'Selecting an optimal data integration strategy' and 'Evaluation of overall atlas representation quality' can be applied: assessing biological preservation and batch correction on the combined atlas and mapped query dataset. Several dedicated metrics and approaches can be used to estimate the quality of a mapping, such as an estimate of the preservation of neighborhoods or clusters before compared to after mapping, the confidence and accuracy of cell-type label transfer from reference to query via metrics that measure uncertainty, and the distance from query cells to reference cells[83–85].

Annotating the integrated atlas. Once the data have been successfully integrated, a reannotation of the cells should be performed to improve the quality and resolution of the annotations. The increase in total cell number enables the detection of rare cell types and states that might not have been annotated in the individual datasets, including groups of low-quality droplets[12,13,15,17]. Furthermore, the joint representation enables resolving contradictory annotations of the same cell type, as is often observed between datasets[17]. If parts of the cells were not labeled before integration, they can now be labeled on the basis of their similarity to labeled cells in the integrated representation. Importantly, if the reannotation is based on the original annotations of individual datasets or multiple independent expert opinions, the reannotated atlas constitutes a first step toward a consensus-based annotation of a given tissue[17].

Low-quality droplets (for example, empty droplets and doublets) will likely still be present in the data at this stage and should be identified to be separated from viable cells before annotation (Fig. 3d,j and Supplementary Note 6). As previously discussed ('Data harmonization and preprocessing'), it is still unclear to what extent quality control can be done entirely after integration, rather than before per dataset or sample. The former not only saves time during preprocessing, but annotating rather than removing low-quality droplets could also enable automated quality control of new datasets mapped onto the atlas via label transfer[97].

Atlas reannotation can be done manually, automatically or by a community-based crowdsourcing approach. The classical and most labor-intensive approach is to manually annotate all cells of the atlas based on their clustering in the integrated representation and marker gene expression[6,8,9,13,15]. Alternatively, preexisting cell-type labels from different datasets can be harmonized to a cell-type hierarchy manually or automatically[58,98,99]. Cells can also be automatically annotated using marker genes[12,24] or via label transfer[10,83,85]. Finally, crowdsourcing approaches enable the collection of annotations from larger groups or networks of experts[59]. These approaches are further discussed in Supplementary Note 7.

To ensure the quality of cell annotations, they must be evaluated from different perspectives. The grouping of cells into cell types should not be driven by technical effects (see section 'Evaluation of overall atlas representation quality'), avoiding, for example, annotating clusters that do not have cells from multiple donors and datasets. Furthermore, annotation labels should be robust, confirmed by the expression of known markers and in broad concordance with prior annotation of individual datasets, including coverage of cell types expected to be

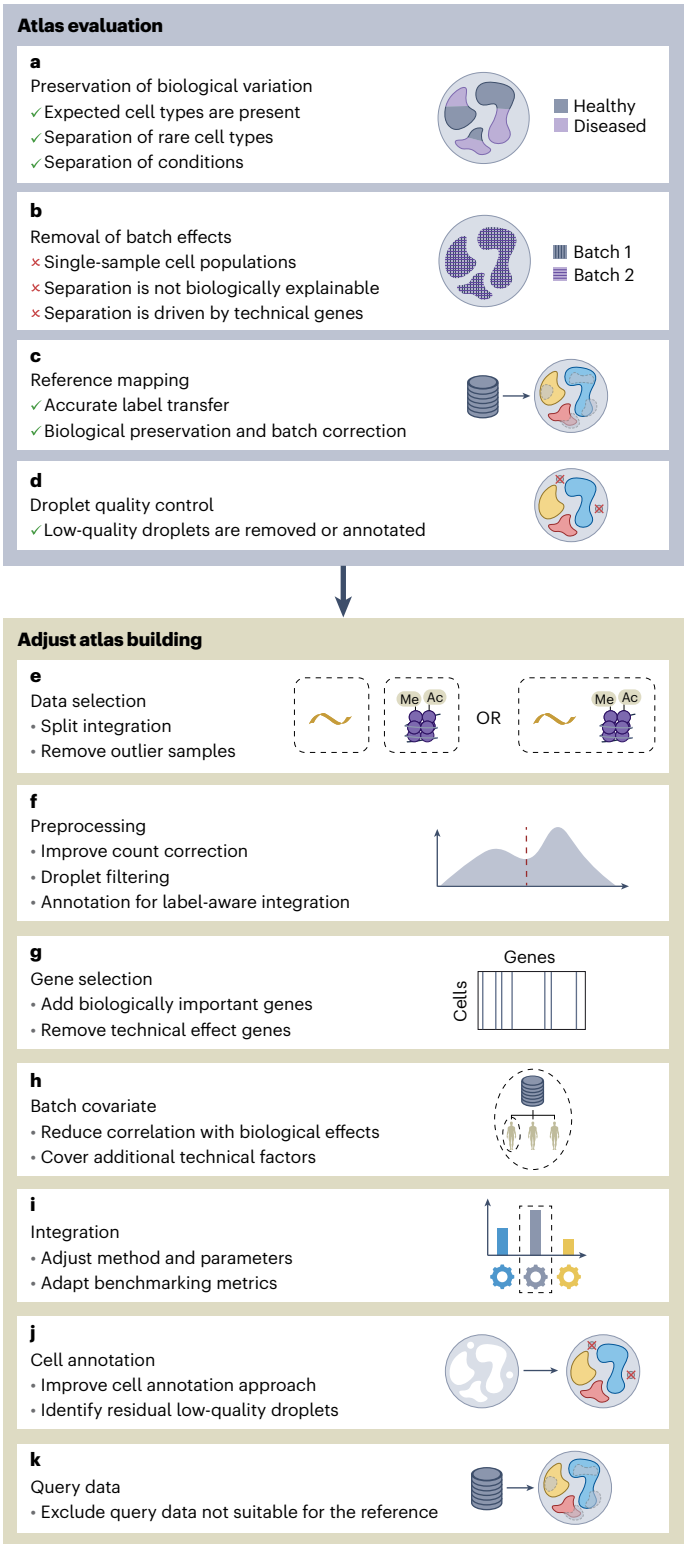

Fig. 3 | Workflow for evaluating and improving the atlas. a–k, The quality of the integrated atlas must be evaluated from different perspectives before proceeding to downstream tasks. This evaluation should assess biological preservation (**a**), batch correction (**b**), reference mapping (**c**) and cell contents of the atlas (**d**). If necessary, atlas quality can then be improved by modifying individual steps (**e–k**) of the atlas building workflow (Fig. 2).

BOX 5

Key takeaways

- Atlas evaluation is key to avoid low-quality integrations that might lead to false interpretations. If the atlas quality is insufficient, individual steps of the atlas building must be adapted and reconducted.
- Manual inspection of the atlas is required to assess that prior knowledge is sufficiently preserved and that batch effects do not bias conclusions drawn from the atlas.
- As reference mapping is one of the key use cases of atlases, it should be evaluated how well new data can be mapped to the atlas.
- Query datasets may be too different from the reference atlas to be successfully mapped. Thus, determining the dataset characteristics required for reliable mapping will improve atlas usability.
- Atlas cell-type reannotation, including the annotation of residual low-quality droplets, is part of atlas evaluation. It is also necessary for ensuring final label quality and for establishing annotation consensus.

present[12,17]. Involving multiple biological experts, as is the aim for the CAP[59], will likewise increase the reliability of the annotations (Box 5).

When the atlas is completed: sharing and extending the atlas

The finalized atlas will represent a community reference that will serve as a resource and that will continue to evolve with new discoveries in the field. To that end, the atlas needs to be made available to diverse user groups upon publication. Moreover, in the long term, the atlas will need to be extended with newly available data and information to stay up to date.

Making the atlas available to different user groups. To ensure that atlases can serve their primary role as a community resource, it is crucial that they are easily accessible and reusable (Table 1). This involves two main requirements. First, the published data should be well documented. This includes the description of all atlas components, including metadata covariates, and the sharing of all atlas-related code. Metadata covariates should moreover adhere to existing ontology nomenclature where possible. Second, while count matrices are commonly shared on portals such as the Gene Expression Omnibus (GEO)[100], BioStudies[101] and the HCA data portal[54], the data should also be made easily accessible to different user groups. For this purpose, specialized tools and frameworks have been developed. For simple queries, such as the visualization of gene expression levels or metadata categories across cells, interactive platforms can be used[53,102–105]. For more specialized analyses, the data should be easy to download, and should be formatted such that it is compatible with standard data analysis platforms[106–108]. Finally, atlas-related models (for example, for query-to-reference mapping) should be shared publicly[109,110], and a framework for automated mapping can be made available[85,111,112]. Notably, it is not yet clear how to share results from downstream atlas analyses in a standardized way, such as for custom marker lists. Further considerations on atlas sharing are elaborated on in Supplementary Note 8.

Extending and updating the atlas. Atlases can be living resources that evolve as new datasets become available[4]. The inclusion of new datasets as they are released adds more individuals or conditions, thus enhancing the statistical power of metadata covariate analyses and improving coverage of cell types and states across biological conditions. Similarly, cell annotations or metadata descriptors can be updated to adhere to evolving ontologies[113] or to include newly discovered cell types from recent studies[114]. Importantly, keeping reference atlases up to date will require considerable community-wide and consortia-wide efforts[115] as is the case in the genomics field, where standard genome builds are iteratively refined and used by the whole scientific community.

New data can be added to the atlas by mapping the new data onto the old atlas using query-to-reference mapping algorithms[83–85]. When more new data accumulates, the reintegration by retraining of the atlas model, rather than atlas extension by query-to-reference mapping, will be necessary to capture or correct for new biological and technical variation. Additionally, reference atlases can be extended with new modalities, such as single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq) and cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)[116]. Therefore, using an integration model that enables the mapping of different modalities may soon become of great importance due to the increased number of single-cell datasets of non-transcriptomic modalities[117].

An intriguing possibility for upscaling and streamlining atlas extension would be to let users, who map their new data to the reference for analysis of their own data, also share the representation of their mapped data on a reference portal. Each mapping, even if not intended for atlas extension, would thus further expand the atlas for all users. Such a continuous community effort would greatly increase the amount of data captured in the atlas at a rate that is not achievable for a single atlas curation team (Box 6).

Table 1 | Different databases and platforms enable sharing of atlas data for different purposes

Atlas sharing purpose	Type of tool/ platform	Examples
Fast and easy access to atlas for simple queries	Interactive platform	CELLxGENE[53], Single Cell Portal[104], UCSC Cell Browser[102], Vitessce[105], and Scope+[103]
Downloadability of atlas for detailed analysis	Single-cell database	GEO[100], HCA data portal[54], CELLxGENE[53]
Reference model sharing for query-to-reference mapping	Model database	Zenodo[109], HuggingFace[110]
Automated query-to-reference mapping	Online mapping platform	Azimuth[85,112], ArchMap[111]
Access to detailed downstream results	No dedicated databases	Paper supplements, Figshare[193]
Reproducibility of analyses and results	Public code repository	GitHub

For each sharing aspect, the type of tool or platform needed is specified, as well as examples of those tools and platforms.

Using integrated atlases

The value of an atlas derives from the biological insights it offers and its role as a consensus reference (Fig. 4). Atlases have the potential to answer pressing biomedical questions, aiding in understanding disease mechanisms, developing new treatments, improving model systems and advancing disease prognosis or diagnosis[36]. Furthermore, atlases can be used to study organism-wide cell function[113,118], development[119], organoid protocol design[120] and evolution across species[121]. Projecting new datasets to the atlas moreover enables atlas-guided analysis of new data. To promote the adoption of atlases across different fields, we here provide an overview of domain-agnostic biological and technical questions that can be answered using integrated atlases alone or as a complement to new data.

Exploring the information within the atlas

Marker genes, gene programs and the effects of biological and technical factors on cell types are routinely investigated in single-cell datasets.

BOX 6

Key takeaways

- Atlases should be publicly accessible both computationally and interactively. However, standards for atlas sharing are not yet fully established and data are currently often scattered across databases.
- Community efforts will be needed to keep atlases up to date with new datasets and associated discoveries, such as newly identified cell types. Atlas updates can be based either on mapping new data onto the atlas or on rebuilding the atlas.

Atlases provide a uniquely comprehensive resource for these analyses due to their greater coverage of biological and technical factors.

Cell identities and their markers. Cell-type annotations across single-cell datasets rarely agree. This is partly due to biological differences in cellular states, but also due to the lack of standardization in cell-type nomenclature and resolution[16]. By combining multiple annotated datasets from different laboratories, conditions, anatomical regions or sample handling protocols[10–12,15,16] as well as different expert opinions on cell-type labels, cell atlases present an opportunity for establishing consensus cell-type annotation[8,14,16,17].

Cell-type markers identified via an atlas are likely to be more specific, sensitive and robust as they are consistent across datasets and thus across protocols[10,16]. Moreover, as atlases pool data across multiple studies, they can reveal rare cell types that are often missed when analyzing individual datasets[8,9,12,16,17,65]. Thus, atlas-based markers are particularly valuable for cell-type annotation in new datasets[122] and evaluation of newly identified or previously proposed markers[14–16], as well as the selection of markers used for tissue staining[7,15], cell sorting[6,123] and probe design for spatial transcriptomics[124]. While common marker identification strategies (benchmarked in ref. 125) can be applied to atlases, additional considerations are required due to the large number of relevant clusters and their relative hierarchy, as well as the large number of datasets and related batch effects (Supplementary Note 9).

Description of gene function and regulation. The gene regulatory landscape and molecular pathways within individual cell types are often inferred via coexpression analyses[2,15,24,126,127]. These analyses benefit from large heterogeneous data collections with many samples. Thus, atlases can be used to robustly identify gene–gene relationships[128,129] and multicellular programs, which are groups of genes co-regulated across different cell types[130]. To better model regulatory relationships between genes, measurements from multiple omics layers can be used[131]. Multi-omic atlases that cover the full omic landscape of emerging multi-omic data types[132–134] could thus serve as a bridge between different omics layers[135].

Molecular and cellular changes across conditions. To understand the molecular characteristics of phenotypes, such as disease, age or sex, one must analyze associated changes in gene expression and cell-type composition[6–8,11,122] (called 'covariate analysis' henceforth). Atlases improve covariate analysis for multiple reasons. They capture a large number of subjects and datasets, which results in better generalization and higher power to detect associations between phenotypes and gene expression[7,17]. These associations can also serve as an additional layer of gene functional information beyond the commonly used pathway databases[15]. The large subject number also results in better coverage of continuous clinical or demographic trajectories, such as aging or disease progression[136]. Likewise, increased patient coverage may reveal heterogeneity between patients with the same disease, enabling patient stratification for personalized medicine[7,11]. Moreover, as atlases combine data from multiple studies, they bring together biological conditions that could not be compared within individual datasets[6,13,15,65,122]. For example, shared molecular characteristics across conditions may be informative for drug repositioning across diseases or tissues[7,17,137]. Similarly, cross-condition differences may aid in selecting preclinical models[9,15]. In the future, atlases may even be used to build predictive models for the clinical classification of patients based on their single-cell profile[8,11,138].

There are multiple challenges of covariate analyses within atlases. The datasets in an atlas were not generated with a single question in mind and thus do not follow a single optimal experimental design to answer any specific question[11]. Furthermore, when building an atlas, batch effects are often only corrected in an integrated representation, while the gene expression counts are left uncorrected[67,86,139,140]. This renders gene expression values incomparable across batches[141]. Similarly, cell proportion analysis may be affected by batch-related differences in sampling protocols (for example, dissociation technique) and tissue sampling locations[11,17]. For these reasons, atlas-level covariate analyses require the incorporation of confounders in statistical models[141,142]. This becomes particularly challenging in the case of partial confounding between biological and technical factors, such as when a cellular trajectory is divided across datasets. Alternatively, one may consider performing the analysis per dataset and afterwards combining the results[143]. Furthermore, modeling assumptions established based on individual datasets are not always met in atlases. For example, cell–cell communication tools assume that all cells were located together in the tissue, which is not true for an atlas as a whole. Thus, standard analysis approaches need to be adjusted with atlas-specific considerations.

Guiding future experimental design. Atlases offer several opportunities to improve the design of future experiments. For example, while individual datasets are rarely generated to assess how different technical parameters affect the data, atlases bring together multiple datasets that enable such analyses[11,12,16,17]. This can reveal which technical factors should be optimized to prevent cell stress, doublets or ambient contamination, or to better capture specific cell types[17]. Furthermore, atlases can be used for power analyses, that is, to estimate the number of cells, samples or donors that need to be profiled to answer specific questions. This can be useful when studying rare cell types or when determining the optimal combination of counts per cell and number of profiled cells for differential gene expression analysis[133,144]. Finally, atlases highlight which cell types, diseases, demographics or other categories are understudied in the current data[12,13,17,145] and need to be better captured in the future (Box 7).

Developing new single-cell methods and machine-learning models

The development of new single-cell methods heavily depends on the availability of high-quality datasets for method testing and benchmarking[146,147]. Highly curated reference atlases are particularly suitable for this for several reasons. First, they contain high-quality data in a standardized format, reducing the need for data wrangling. This is of particular interest for the development of large-scale generalizable 'foundation' models for single-cell biology (Supplementary Note 10). Second, they contain diverse large-scale data and thereby present realistic challenges (for example, batch effects) for methods, revealing potential method limitations. Third, they contain diverse data appropriate for various benchmarking tasks. This covers different analysis types, including trajectory inference across continuous covariates, differential analysis across conditions and integration across batches. Fourth, due to their size, atlases can easily be split at random or in a stratified manner (for example, by datasets and lineages) to conduct benchmarks for time efficiency and data complexity. Fifth, atlases are

https://doi.org/10.1038/s41592-024-02532-y

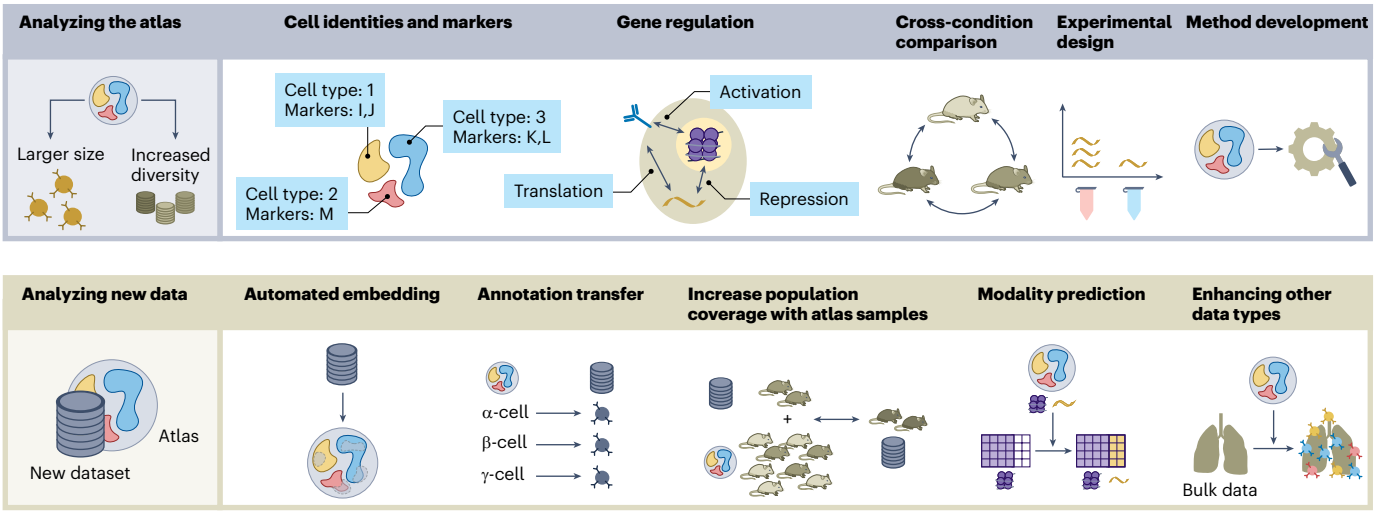

Fig. 4 | Use cases of integrated atlases. a,**b**, The rich information captured within the atlas can provide new biological or technical insights in multiple ways and can serve as a baseline for method development (**a**) or can be used as a reference for analyzing new single-cell, spatial or bulk datasets (**b**).

BOX 7

Key takeaways

- Atlases will play a key role in establishing cell annotation standards within tissues and across conditions. The number and diversity of datasets improve marker robustness and ease identifying rare cell populations.
- Atlases will improve our understanding of gene regulation, especially when multi-omic atlases become more widely available.
- Atlases can enable the association of demographic, clinical and other biological covariates with gene expression changes, due to large sample numbers and multi-condition coverage. However, batch effects may need to be explicitly modeled.
- Atlases can reveal how experimental protocol characteristics affect data quality and cell population capture.
- Atlases provide insights into under-sampled conditions and cell populations.

BOX 8

Key takeaways

- The quality and diversity of the data within atlases make them especially well suited for the development and comparison of new methods and models.

often well explored, thereby providing a reliable approximation of ground truth in terms of biological and technical effects present in the data. Thus, atlases have the potential to serve as standard benchmarking datasets, which are common in the machine-learning field[148], but still rare in single-cell data science[149–151](Box 8).

Analyzing new single-cell, spatial or bulk data with atlases as references

Analysis of new data provides many challenges, such as data integration and de novo cell-type annotation, and may in some cases be limited due to a low number of cells and samples. These and many other challenges can be alleviated by leveraging atlases as a basis for the analysis of new data. Atlases can also supplement new data with additional biological information, such as in cross-modal expression prediction, expansion of the control sample pool and contextualization of bulk data with single-cell information.

Projecting new single-cell data into an atlas space. One of the central goals of scRNA-seq analysis is obtaining a high-quality, low-dimensional representation that enables the identification of cell types, states and trajectories. Atlases provide such high-quality representation and by using query-to-reference mapping methods, new datasets (queries) can be positioned within the atlas (reference) representation space. This has multiple advantages over analyzing query data alone[12,152,153]. First, rare cell populations are better represented in the atlas and can thus be better identified in the mapped query dataset as well. Second, the atlas representation was optimized to distinguish biological from technical variation using many training datasets. Mapping to this representation can improve batch correction in query data, in particular if query batch effects are directly confounded with biological variation and can thus not be disentangled using the query alone. Third, mapping into the atlas representation space enables a rapid, joint analysis of the new dataset and the atlas, for example, for cell identity annotation transfer and comparison (more details in 'Annotating cellular identities' and 'Comparisons with a control population'). Atlas-based representations are thus likely to become a standard step in future scRNA-seq analyses.

Successful mapping of a query to a reference depends on a number of factors, including the sample characteristics, data preprocessing and the mapping method. The mapped samples should be sufficiently similar to the reference, both in terms of sample biology as well as the measured features and preprocessing choices. For example, a human reference may not work optimally for mapping animal data or data from other preclinical models, such as organoids. Such mapping can result in either too much separation of the query and reference in the resulting representation due to under-correction or merging of distinct cell populations due to overcorrection when attempting to increase integration strength. In both cases, the integration failure hampers the correct interpretation of the results. Furthermore, while the reference mapping method that can be used with a given atlas is in many cases dependent on the model used for atlas building, some of the methods allow tuning of the mapping parameters to tailor the mapping to a given context. This includes modifying integration strength and adding biologically relevant features (genes) missing from the ref. 97. Finally, while reference mapping should be always evaluated, tools enabling this are still lacking.

BOX 9

Key takeaways

- Mapping new datasets onto an atlas can improve their data representation. Successful mapping depends on the correspondence between the atlas and the new data in terms of biological, technical and data preprocessing characteristics.
- Data quality issues can be revealed by unexpected localization of new data points in the atlas representation space after mapping.
- New datasets can be rapidly annotated based on automated cell label transfer from the reference.
- Atlas-guided case–control comparisons in new data can improve population and condition coverage. However, atlases cannot fully replace matched control populations for new data.
- Multimodal atlases may in the future enrich unimodal datasets via cross-modal imputation.
- Atlases can help to infer cell-type proportions in bulk and spatial data.
- Atlases can be used to identify cell populations expressing genes of interest, such as drug targets or genes with disease-associated polymorphisms.

With the increased number of datasets capturing different omics layers, it will be of interest to use references consisting of one modality (currently this is usually the transcriptome) to analyze queries from different modalities. This enables various downstream analyses, such as cross-modality annotation transfer and identification of cross-omic feature correlations. While different strategies for cross-omic mapping were proposed[116,154], they have not yet been widely applied in the atlasing field.

Annotating cellular identities. While manual cell annotation is cumbersome and prone to mistakes[17,65], atlases enable automated transfer of high-quality reference annotations to new datasets. This is commonly done by transferring annotations from reference cells that are close by in the representation to mapped query cells[8,10,65,83–85,155,156]. Furthermore, annotation of low-quality or doublet droplets is often cumbersome, unreliable and inconsistent due to manually set thresholds[157]. Atlases that have annotated such populations may be used to automatically annotate low-quality droplets in the new datasets[97]. Finally, uncertainty metrics[158] can be used to identify annotations transferred with high uncertainty and, therefore, with a higher likelihood to be incorrect. Cells with high uncertainty labels have been shown to represent unseen cell identities in the new data (that is, cell identities not present in the reference atlas), such as new cell types or disease-related cell states[17]. Atlases are thus expected to serve as the first step in the annotation and analysis of future datasets, guiding the manual fine-tuning of the annotation[159].

Comparisons with a control population. Identifying the difference between healthy cellular phenotypes and those specific to a disease based on a single dataset can be complicated by within-cohort batch effects, incomplete coverage of healthy cell populations and small sample sizes. Using atlases as a basis for the analysis of new query data can mitigate these limitations. For example, mapping query samples from disease conditions on top of a healthy reference can directly identify cell types that have an altered, non-healthy phenotype[14,17,158]. The atlas size reduces the chance for falsely interpreting healthy variation as disease effects due to the lack of comprehensive controls. Nevertheless, atlases alone cannot yet fully replace control samples for new datasets[152]. Furthermore, using an atlas as a reference enables comparing cells across a wide range of conditions included in the atlas. This has been used to optimize organoid protocols[120] or compare model systems[160]. However, one must stay cautious in jointly analyzing atlases and mapped data, as atlases and mappings never perfectly remove all batch effects, which may lead to biases in the interpretation[152].

Reference atlases may also serve as a control to assess sample quality and thus prevent mistaking unanticipated technical variation for biological differences. For example, if cells from a new sample map to an unexpected location in the atlas representation, away from the reference cells of the corresponding cell-type and biological condition, this may indicate low sample quality or strong technical artifacts.

Cross-modal imputation. To improve the understanding of cellular states and regulation across modalities, data imputation across modalities can enrich unimodal datasets. Multimodal atlases can be useful for imputation due to the large number of contained datasets, thus increasing the reliability of imputation models[161,162]. Imputation can be used in many different settings, such as denoising, cross-omic prediction, prediction of non-measured genes in spatial data or imputation of spatial location in nonspatial data[12,135,161–166]. However, special care needs to be taken to assess the reliability of the imputation, especially when imputing conditions that are not closely related to the training data[167].

Analysis of non-single-cell data. Many bulk datasets are available across modalities and individuals[11], but they lack cell-type information that is crucial for understanding tissue function and disease. Similarly, spatial data often do not reach single-cell resolution. To enable interpretation at the level of cell populations, these datasets can be deconvolved based on single-cell data[168–171]. Atlases are uniquely suited for bulk[8,11,17,172,173] and spatial[8,13,24] data deconvolution due to their comprehensive coverage of the cell types present in a tissue or organ, and their robust multi-batch cell-type profiles.

Atlases can also help to better understand data that are neither single-cell nor transcriptomic. For example, they can help identify cell types that may be affected by perturbations, such as disease-associated genomic variants in genome-wide association studies[17,35,122,174,175] and cell types that may be targeted by specific drugs[176]. Eventually, when sufficient matched single-cell and clinical data are collected, it may become possible to develop models that could infer cell-level features, such as cell proportions, from images or other clinical measurements[177] (Box 9).

Conclusion and outlook

With the maturation of data integration methods and the wide availability of single-cell datasets, atlasing studies are becoming increasingly common. Atlas resources promise to build consensus across communities and impact biomedical research[36]. However, standards for building atlases are lacking and atlas use cases are still being explored. In this Review, we discussed considerations and opportunities for building and using atlases to initiate a discussion on standards in the field.

There are still many open questions in the field of atlas building that would benefit from benchmarks and that call for new datasets, technologies and methods. First, a systematic comparison of different atlas building pipelines is lacking. Second, as atlases become more comprehensive and complex, including cross-species, longitudinal, whole-organism and multimodal data, the need for integration methods that can accommodate such complex scenarios will grow. Recent developments in the machine-learning community, such as foundation models that are able to generate broadly usable representations for large and diverse datasets, may thus also be useful in the single-cell community[178,179]. Third, cost and labor reduction of single-cell profiling technologies will be needed to enable population-wide and cross-omic atlases as well as to popularize their use in clinics. These open questions and potential solutions are further discussed in Supplementary Note 11.

Similarly, given that atlases are designed as community resources and their usability is of crucial importance, we identify key areas for

usability enhancement. First, the performance of interactive interfaces and standard analysis pipelines diminishes with the size of datasets. To overcome this, wider adoption of graphics processing unit (GPU)-accelerated tools[180], developing more compact data representations, such as compressing cells into meta cells, encoding data into foundation models or simpler generative models, or providing standardized, human or machine-readable descriptions of cell and gene landscapes, would be beneficial. Second, as atlases increase in complexity, their visualization and interpretation also do so. Workflows should, therefore, be adopted to ease interpretation and visualization of atlases spanning tissue resolutions and omics layers. Third, existing workflows for analyzing new data based on mapping onto reference atlases are still in prototype stages and require further development and testing. Fourth, single-cell datasets covering underrepresented donor populations, such as specific ancestries, are needed to make atlases more generalizable and robust. Fifth, although atlases hold great potential for various fields including molecular biology, medicine and computational sciences, current access interfaces are mainly tailored to the bioinformatics community[181]. Therefore, it is necessary to further develop data-access options tailored to different user needs, including interactive platforms and application programming interface access points.

As new single-cell datasets are generated, atlases will also grow in size and complexity. This will bring with it questions regarding optimal atlas size and the point at which an atlas can be considered 'complete'. Future studies will need to systematically assess at what point adding more data no longer improves the coverage of biological information (for example, cell states and ancestries) or the quality of the integration. Currently, it is still unclear how the optimal atlas size can be determined in practice[182]. In part, this is due to the diversity of goals of atlas studies. Healthy cell-type variation, including rare cell states, may be comprehensively retrievable with currently available datasets. In contrast, comprehensive coverage of genetic and phenotypic diversity across populations and conditions will require a large number of samples, which is unlikely to be achieved in the near future[36].

Despite the promises of reference atlases, they also come with limitations. First, atlases rely on integration to remove batch effects between datasets. However, this rarely works perfectly and, especially when batch effects are strong, also removes biological variation. This can limit the resolution of retrievable cell populations. Second, just like any individual single-cell dataset, atlases are designed with particular goals in mind and thus may be unsuitable to answer certain biological questions. For example, if atlas-builders focus on providing a healthy reference, this may limit atlas-based analysis of data from other conditions. Third, atlas building demands substantial human and computational resources, which is likely to increase as atlases grow in size. Thus, recent work has proposed to complement high-quality reference atlases with automated pipelines that enable more modular and rapid data integrations tailored to a specific biological question at hand[68,183,184]. Fourth, the quality of atlases and their long-term maintenance vary as best practice standards are currently lacking. In this Review, we aim to make a first step toward establishing these standards.

While atlases are expected to have a profound effect on medicine, ranging from disease target identification and toxicity prediction to direct applications in clinics[36], medicine is not the only field that is anticipated to be transformed by atlases. For example, cross-species atlases may provide insights into phylogeny[121,185,186] and environmental niches[187–189]. Similarly, ecology and agriculture atlases[190,191] could integrate cross-areal datasets to reveal the interactions between environment and organism[188,192]. However, before such atlases can be created, more single-cell datasets must be generated in these domains. In the near future, atlases outside the biomedical field will, therefore, likely be focused on model organisms, for which sufficient data and community interest are present.

In this Review, we outlined current considerations and recommendations in building, using and sharing atlases, and highlighted aspects of these processes that merit further research and development. We envision that the insights collected here will aid in setting common standards for atlas building and will fuel the broad use of atlases in single-cell research. Together, this will pave the way to a consensus-based approach for describing cellular biology, increasing the impact of atlases on molecular biology and medicine.

Data availability

The final results of the analysis of the published scRNA-seq datasets are collected in Supplementary Table 2 and the intermediate results are available at https://github.com/lueckenlab/single-cell-papers-trends/.

Code availability

The code for the analysis of the published scRNA-seq datasets depicted in Fig. 1 is available at https://github.com/lueckenlab/single-cell-papers-trends/.

References

1. Qiu, C. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54**, 328–341 (2022).
2. Tritschler, S. et al. A transcriptional cross species map of pancreatic islet cells. *Mol. Metab.* **66**, 101595 (2022).
3. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
4. Regev, A. et al. The Human Cell Atlas White Paper. Preprint at https://arxiv.org/abs/1810.05192 (2018).
   Sets out the vision and goals of the HCA. The HCA consortium aims to create comprehensive reference maps of all human cells and represents the largest single-cell atlas initiative worldwide.
5. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
   Lays out the framework of the HuBMAP, one of the two largest single-cell atlas initiatives to comprehensively map the human body at single-cell level.
6. Buechler, M. B. et al. Cross-tissue organization of the fibroblast lineage. *Nature* **593**, 575–579 (2021).
7. Cheng, S. et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792–809 (2021).
   A pioneering cross-disease integrated atlas, characterizing tumor-infiltrating myeloid cells across 15 cancer types. Found myeloid cell phenotypes to be dependent on cancer type, possibly affecting responsiveness to cancer immunotherapies.
8. Nieto, P. et al. A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**, 1913–1926 (2021).
9. Schupp, J. C. et al. Integrated single-cell atlas of endothelial cells of the human lung. *Circulation* **144**, 286–302 (2021).
10. Swamy, V. S., Fufa, T. D., Hufnagel, R. B. & McGaughey, D. M. Building the mega single-cell transcriptome ocular meta-atlas. *Gigascience* **10**, giab061 (2021).
    One of the first studies to perform a systematic metric-based evaluation of data integration quality across integration methods, here for building a retina atlas. This is also one of the few cross-species atlases.
11. Salcher, S. et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503–1520 (2022).
12. Steuernagel, L. et al. HypoMap—a unified single-cell gene expression atlas of the murine hypothalamus. *Nat. Metab.* **4**, 1402–1419 (2022).

13. Suo, C. et al. Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).
Early atlas paper to focus on fetal development. The inclusion of datasets across developmental time and tissues enabled several new discoveries, such as the identification of fetal hematopoiesis across organs.
14. Guo, M. et al. Guided construction of single cell reference for human and mouse lung. *Nat. Commun.* **14**, 4566 (2023).
15. Hrovatin, K. et al. Delineating mouse β-cell identity during lifetime and in diabetes with a single cell atlas. *Nat. Metab.* **5**, 1615–1637 (2023).
Exemplifies several different use cases of integrated atlases, among which are the evaluation of different diabetes mouse models and the identification of shared pathways in beta cells across ages and in disease.
16. Novella-Rausell, C., Grudniewska, M., Peters, D. J. M. & Mahfouz, A. A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. *iScience* **26**, 106877 (2023).
17. Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
This was a pioneering integrated healthy reference atlas of a human organ. It showcases the value of healthy reference atlases in a variety of ways, including the use of healthy reference atlases for the understanding of disease datasets.
18. Chu, Y. et al. Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nat. Med.* **29**, 1550–1562 (2023).
19. Swamy, V. S., Batz, Z. A. & McGaughey, D. M. PLAE Web App enables powerful searching and multiple visualizations across one million unified single-cell ocular transcriptomes. *Transl. Vis. Sci. Technol.* **12**, 18 (2023).
20. Tang, F. et al. A pan-cancer single-cell panorama of human natural killer cells. *Cell* **186**, 4235–4251 (2023).
21. Yang, Y. T. et al. STAB2: an updated spatio-temporal cell atlas of the human and mouse brain. *Nucleic Acids Res.* **52**, D1033–D1041 (2023).
22. Li, Z. et al. An atlas of cell-type-specific interactome networks across 44 human tumor types. *Genome Med.* **16**, 30 (2024).
23. Reed, A. D. et al. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nat. Genet.* **56**, 652–662 (2024).
24. Ruiz-Moreno, C. et al. Harmonized single-cell landscape, intercellular crosstalk and tumor architecture of glioblastoma. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.27.505439 (2022).
25. He, Z. et al. An integrated transcriptomic cell atlas of human neural organoids. *Nature* **635**, 690–698 (2024).
One of the first integrated atlases of an in vitro model system, spanning 26 different neural organoid protocols. It compares the presence and transcriptional profiles of cell types in organoids to those of their counterparts in the developing brain.
26. Li, J. et al. Integrated multi-omics single cell atlas of the human retina. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.07.566105 (2023).
Multimodal integrated atlas. By leveraging both the high cell-type resolution of the large-scale transcriptomic data and the chromatin accessibility data, cell-subtype-specific gene regulatory networks are identified.
27. Miao, Z. et al. A brain cell atlas integrating single-cell transcriptomes across human brain regions. *Nat. Med.* **30**, 2679–2691 (2024).
28. Wu, Y. et al. uniLIVER: a human liver cell atlas for data-driven cellular state mapping. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.09.570903 (2023).
29. Zhang, X. et al. uniHEART: an ensemble atlas of cardiac cells provides multifaceted portraits of the human heart. Preprint at *Res. Sq.* https://doi.org/10.21203/rs.3.rs-3215038/v1 (2023).
30. Xu, Q. et al. An integrated transcriptomic cell atlas of human endoderm-derived organoids. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.20.567825 (2023).
31. Marečková, M. et al. An integrated single-cell reference atlas of the human endometrium. *Nat. Genet.* https://doi.org/10.1038/s41588-024-01873-w (2024).
32. Li, J. et al. Comprehensive single-cell atlas of the mouse retina. *iScience* **27**, 109916 (2024).
33. Netskar, H. et al. Pan-cancer profiling of tumor-infiltrating natural killer cells through transcriptional reference mapping. *Nat. Immunol.* **25**, 1445–1459 (2024).
34. Herpelinck, T. et al. An integrated single-cell atlas of the limb skeleton from development through adulthood. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.14.484345 (2024).
35. Muus, C. et al. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat. Med.* **27**, 546–559 (2021).
36. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
Describes the vision and potential of the HCA, currently the largest single-cell atlas initiative, in advancing medicine, as well as the achievements it has made so far toward that goal.
37. Yao, Z. et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).
38. Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385 (2023).
39. Zhang, B. et al. A human embryonic limb cell atlas resolved in space and time. *Nature* https://doi.org/10.1038/s41586-023-06806-x (2023).
40. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
41. ENCODE Project Consortiumet al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
42. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
Describes the Human Protein Atlas, a highly impactful initiative predating the single-cell era, aiming to map protein expression across organs, tissues and cells.
43. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
Presents a comprehensive atlas-level integration benchmarking framework 'scIB'. The framework can be applied to any combination of data and methods using a wide array of metrics.
44. Madissoon, E. et al. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat. Genet.* **55**, 66–77 (2022).
45. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
46. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
47. Uniken Venema, W. T. C. et al. Gut mucosa dissociation protocols influence cell type proportions and single-cell gene expression levels. *Sci. Rep.* **12**, 9897 (2022).

48. Mereu, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
49. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10**, 4667 (2019).
50. Vasilevsky, N. A. et al. Mondo: Unifying diseases for the world, by the world. Preprint at *medRxiv* https://doi.org/10.1101/2022.04.13.22273750 (2022).
51. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
52. Malone, J. et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
53. CZI Single-Cell Biology Program et al. CZ CELL×GENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.30.563174 (2023).
Most widely used platform for sharing single-cell data in an interactive fashion, offering efficient interactive visualization, data querying and storing in a unified format.
54. HCA Data Portal. Mapping the human body at the cellular level. https://data.humancellatlas.org/
55. Kulhankova, L. et al. Single-cell transcriptome sequencing allows genetic separation, characterization and identification of individuals in multi-person biological mixtures. *Commun. Biol.* **6**, 201 (2023).
56. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* https://doi.org/10.1038/s41592-023-02035-2 (2023).
57. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
Introduced the Cell Ontology, a now widely used resource and ongoing effort to standardize and centralize cell-type nomenclature. Single-cell data platforms such as CELLxGENE use this ontology to ensure standardized cell-type nomenclature across datasets.
58. Xu, C. et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* **186**, 5876–5891 (2023).
59. Cell Annotation Platform. https://celltype.info/
60. Cheng, C., Chen, W., Jin, H. & Chen, X. A review of single-cell RNA-seq annotation, integration, and cell–cell communication. *Cells* **12**, 1970 (2023).
61. Clarke, Z. A. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.* **16**, 2749–2764 (2021).
62. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
63. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
64. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.* **39**, 1095–1102 (2021).
65. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
66. Zhu, Y., Wang, L., Yin, Y. & Yang, E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Sci. Rep.* **7**, 5435 (2017).
67. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
Most widely used data integration method for integrated atlases. scVI is based on a neural network that simultaneously performs batch correction and dimensionality reduction. It is popular due to its high performance and scalability to large datasets.
68. Wang, Y. et al. Automated single-cell omics end-to-end framework with data-driven batch inference. *Cell Syst.* https://doi.org/10.1016/j.cels.2024.09.003 (2024).
69. Lütge, A. et al. CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life Sci. Alliance* **4**, e202001004 (2021).
70. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
71. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
72. He, P. et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell* **185**, 4841–4860 (2022).
73. Ranjan, B. et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat. Commun.* **12**, 5849 (2021).
74. Tyler, S. R., Lozano-Ojalvo, D., Guccione, E. & Schadt, E. E. Anti-correlated feature selection prevents false discovery of subpopulations in scRNAseq. *Nat. Commun.* **15**, 699 (2024).
75. Yang, P., Huang, H. & Liu, C. Feature selection revisited in the single-cell era. *Genome Biol.* **22**, 321 (2021).
76. Xu, Y. et al. CellBRF: a feature selection method for single-cell clustering using cell balance and random forest. *Bioinformatics* **39**, i368–i376 (2023).
77. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446–456 (2021).
78. Eltager, M. et al. Benchmarking variational AutoEncoders on cancer transcriptomics data. *PLoS ONE* **18**, e0292126 (2023).
79. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
80. Elmentaite, R. et al. Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).
81. Song, Y., Miao, Z., Brazma, A. & Papatheodorou, I. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat. Commun.* **14**, 6495 (2023).
82. Hrovatin, K. et al. Integrating single-cell RNA-seq datasets with substantial batch effects. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.03.565463 (2024).
83. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
One of the most widely used methods for mapping new data to an existing reference atlas, called scArches. It enables extending pretrained neural network-based integration models, to perform batch effect correction of new datasets with respect to an existing reference.
84. Kang, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
85. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
Paper that first introduced one of the most popular query-to-reference mapping methods (initially presented as an extension of a multimodal data integration method). The method is now part of the Seurat data analysis platform as a general query-to-reference method called 'Azimuth'.
86. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
87. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2023).
88. Wang, H., Leskovec, J. & Regev, A. Metric mirages in cell embeddings. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.02.587824 (2024).
89. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151 (2020).

90. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
91. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
92. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
93. Bonnycastle, L. L. et al. Single-cell transcriptomics from human pancreatic islets: sample preparation matters. *Biol. Methods Protoc.* **5**, bpz019 (2020).
94. Massoni-Badosa, R. et al. Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol.* **21**, 112 (2020).
95. Basile, G. et al. Using single-nucleus RNA-sequencing to interrogate transcriptomic profiles of archived human pancreatic islets. *Genome Med.* **13**, 128 (2021).
96. Maan, H. et al. Characterizing the impacts of dataset imbalance on single-cell data integration. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-023-02097-9 (2024).
97. Lotfollahi, M. et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell Biol.* **25**, 337–350 (2023).
98. Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun.* **12**, 2799 (2021).
99. Domcke, S. & Shendure, J. A reference cell tree will serve science better than a reference cell atlas. *Cell* **186**, 1103–1114 (2023).
100. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).
101. Sarkans, U. et al. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2018).
102. Speir, M. L. et al. UCSC cell browser: visualize your single-cell data. *Bioinformatics* **37**, 4578–4580 (2021).
103. Yin, D. et al. Scope+: an open source generalizable architecture for single-cell atlases at sample and cell levels. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.03.518997 (2023).
104. Single Cell Portal. https://singlecell.broadinstitute.org/single_cell/
105. Keller, M. S. et al. Vitessce: integrative visualization of multimodal and spatially resolved single-cell data. *Nat. Methods* https://doi.org/10.1038/s41592-024-02436-x (2024).
106. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).
107. Amezquita, R. A. et al. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
108. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
109. European Organization for Nuclear Research & OpenAIRE. *Zenodo* https://doi.org/10.25495/7GXK-RD71 (2013).
110. Hugging Face. The AI community building the future. https://huggingface.co/
111. ArchMap. https://www.archmap.bio/#/
112. Azimuth. https://azimuth.hubmapconsortium.org/
113. Osumi-Sutherland, D. et al. Cell type ontologies of the Human Cell Atlas. *Nat. Cell Biol.* **23**, 1129–1135 (2021).
114. Michielsen, L. et al. Single-cell reference mapping to construct and extend cell-type hierarchies. *NAR Genom. Bioinform.* **5**, lqad070 (2023).
115. Lindeboom, R. G. H., Regev, A. & Teichmann, S. A. Towards a human cell atlas: taking notes from the past. *Trends Genet.* **37**, 625–630 (2021).
116. Zhang, Z. et al. scMoMaT jointly performs single cell mosaic integration and multi-modal bio-marker detection. *Nat. Commun.* **14**, 384 (2023).
117. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01767-y (2023).
118. Quake, S. R. A decade of molecular cell atlases. *Trends Genet.* **38**, 805–810 (2022).
119. Haniffa, M. et al. A roadmap for the human developmental cell atlas. *Nature* **597**, 196–205 (2021).
120. Bock, C. et al. The organoid cell atlas. *Nat. Biotechnol.* **39**, 13–17 (2021).
121. Tarashansky, A. J. et al. Mapping single-cell atlases throughout metazoa unravels cell type evolution. *Elife* **10**, e66747 (2021).
122. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
123. van Gurp, L. et al. Generation of human islet cell type-specific identity genesets. *Nat. Commun.* **13**, 2020 (2022).
124. Kuemmerle, L. B. et al. Probe set selection for targeted spatial transcriptomics. *Nat. Methods* **21**, 2260–2270 (2024).
125. Pullin, J. M. & McCarthy, D. J. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol.* **25**, 56 (2024).
126. Song, Q., Ruffalo, M. & Bar-Joseph, Z. Using single cell atlas data to reconstruct regulatory networks. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkad053 (2023).
127. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
128. Du, J. et al. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* **20**, 82 (2019).
129. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01940-3 (2023).
130. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* **40**, 1467–1477 (2022).
131. Badia-I-Mompel, P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
132. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
133. Hansen, J. et al. A reference tissue atlas for the human kidney. *Sci. Adv.* **8**, eabn4965 (2022).
134. Massoni-Badosa, R. et al. An atlas of cells in the human tonsil. *Immunity* **57**, 379–399 (2024).
135. Jiang, S. et al. Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Rep.* **42**, 112210 (2023).
136. Wang, L. et al. A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets. *Nat. Cancer* **3**, 1534–1552 (2022).
137. Wang, Y.-Y. et al. CeDR Atlas: a knowledgebase of cellular drug response. *Nucleic Acids Res.* **50**, D1164–D1171 (2022).
138. COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916–938 (2022).
139. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
140. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).

141. Boyeau, P. et al. An empirical Bayes method for differential expression analysis of single cells with deep generative models. *Proc. Natl Acad. Sci. USA* **120**, e2209124120 (2023).
142. Law, C. W. et al. A guide to creating design matrices for gene expression experiments. *F1000Res.* **9**, 1444 (2020).
143. Toro-Domínguez, D. et al. A survey of gene expression meta-analysis: methods and applications. *Brief. Bioinform.* **22**, 1694–1705 (2021).
144. Schmid, K. T. et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun.* **12**, 6625 (2021).
145. Zilbauer, M. et al. A roadmap for the Human Gut Cell Atlas. *Nat. Rev. Gastroenterol. Hepatol.* https://doi.org/10.1038/s41575-023-00784-1 (2023).
146. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
147. Open Problems in Single-Cell Analysis Consortium. Open problems in single-cell analysis. https://openproblems.bio/ (2022).
148. Koch, B., Denton, E., Hanna, A. & Foster, J. G. Reduced, reused and recycled: the life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).
149. Lance, C. et al. Multimodal single cell data integration challenge: results and lessons learned. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.11.487796 (2022).
150. Luecken, M. D. et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2022).
151. Luecken, M. et al. Defining and benchmarking open problems in single-cell analysis. Preprint at *Res Sq.* https://doi.org/10.21203/rs.3.rs-4181617/v1 (2024).
Describes open problems in the single-cell computational field, many of which also exist in the atlas field. The paper defines specific tasks that require future benchmarking and offers curated benchmarking datasets.
152. Dann, E. et al. Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01523-7 (2023).
153. Lotfollahi M., Hao Y., Theis F. J., Satija R. The future of rapid and automated single-cell data analysis using reference mapping. *Cell* **187**, 2343–2358 (2024).
154. Athaya, T., Ripan, R. C., Li, X. & Hu, H. Multimodal deep learning approaches for single-cell multi-omics data integration. *Brief. Bioinform.* **24**, bbad313 (2023).
155. He, P. et al. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760–767 (2020).
156. Heimberg, G. et al. A cell atlas foundation model for scalable search of similar human cells. Preprint at *bioRxiv* https://doi.org/10.1038/s41586-024-08411-y (2024).
157. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
158. Engelmann, J. et al. Uncertainty quantification for atlas-level cell type transfer. Preprint at https://arxiv.org/abs/2211.03793 (2022).
159. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
160. Lang, N. J. et al. Ex vivo tissue perturbations coupled to single-cell RNA-seq reveal multilineage cell circuit dynamics in human lung fibrogenesis. *Sci. Transl. Med.* **15**, eadh0908 (2023).
161. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).
162. Wang, J. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
163. Liu, L., Li, W., Wong, K. -C., Yang, F. & Yao, J. A pre-trained large language model for translating single-cell transcriptome to proteome. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.04.547619 (2023).
164. Lotfollahi, M., Litinetskaya, A. & Theis, F. J. Multigrate: single-cell multi-omic data integration. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.16.484643 (2022).
165. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
166. Zhang, M. et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).
167. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Res.* **7**, 1740 (2018).
168. Im, Y. & Kim, Y. A comprehensive overview of RNA deconvolution methods and their application. *Mol. Cells* **46**, 99–105 (2023).
169. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).
170. Li, H. et al. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat. Commun.* **14**, 1548 (2023).
171. Chen, J. et al. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Brief. Bioinform.* **23**, bbac245 (2022).
172. Frishberg, A. et al. Cell composition analysis of bulk genomics using single-cell data. *Nat. Methods* **16**, 327–332 (2019).
173. Liao, J. et al. De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution. *Nat. Commun.* **13**, 6498 (2022).
174. Yazar, S. et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
175. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
176. Kanemaru, K. et al. Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–810 (2023).
177. Hu, J. et al. Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Comput. Struct. Biotechnol. J.* **19**, 3829–3841 (2021).
178. Szałata, A. et al. Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
179. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
180. Nolet, C. et al. Accelerating single-cell genomic analysis with GPUs. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.26.493607 (2022).
181. Xu, Y., Ahn, J. & Zanini, F. Fast and lightweight cell atlas approximations across organs and organisms. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.03.573994 (2024).
182. Fleck, J. S., Camp, J. G. & Treutlein, B. What is a cell type? *Science* **381**, 733–734 (2023).
183. Sina Booeshaghi, A., Galvez-Merchán, Á. & Pachter, L. Algorithms for a Commons Cell Atlas. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.23.586413 (2024).
184. Galvez-Merchán, Á., Sina Booeshaghi, A. & Pachter, L. A human commons cell atlas reveals cell type specificity for OAS1 isoforms. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.23.586412 (2024).
185. Evolution at the cellular level. *Nat. Ecol. Evol.* **7**, 1155–1156 (2023).
186. Jorstad, N. L. et al. Comparative transcriptomics reveals human-specific cortical features. *Science* **382**, eade9516 (2023).

187. Tabula Sapiens Consortiumet al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
188. Brennan, M. A. & Rosenthal, A. Z. Single-cell RNA sequencing elucidates the structure and organization of microbial communities. *Front. Microbiol.* **12**, 713128 (2021).
189. Alacid, E. & Richards, T. A. A cell-cell atlas approach for understanding symbiotic interactions between microbes. *Curr. Opin. Microbiol.* **64**, 47–59 (2021).
190. Rhee, S. Y., Birnbaum, K. D. & Ehrhardt, D. W. Towards building a plant cell atlas. *Trends Plant Sci.* **24**, 303–310 (2019).
191. Polychronidou, M. et al. Single-cell biology: what does the future hold? *Mol. Syst. Biol.* **19**, e11799 (2023).
192. Plant Cell Atlas Consortiumet al. Vision, challenges and opportunities for a Plant Cell Atlas. *Elife* **10**, e66877 (2021). **Describes the framework for building a Plant Cell Atlas to advance understanding of plant physiology, development and environmental responses. It is one of the first initiatives promoting atlases outside fields related to human biology and medicine.**
193. Gruber, C. Figshare - credit for all your research. https://figshare.com/
194. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).

Acknowledgements

This publication is part of the HCA (www.humancellatlas.org/publications/). We thank C. Xu and A. M. Cujba for providing comments on the manuscript and A. C. Villani for providing detailed information on the CAP. This work was supported by the Joachim Herz Stiftung via Add-on Fellowships for Interdisciplinary Life Science (to K.H.), the Helmholtz Association under the joint research school 'Munich School for Data Science' (to K.H. and V.A.S.), the Chan Zuckerberg Initiative via grant CZIF2022-007488 - HCA Data Ecosystem (to M.D.L., F.J.T., S.A.T. and P.H.), the LLC Seed Network via grant CZF2019-002438 - Lung Cell Atlas 1.0 (to F.J.T.), the Helmholtz Association and Helmholtz Munich (to F.J.T.), the RESPIRE4 Marie Sklodowska-Curie fellowship via grant agreement 847462 (to A.J.O.), St Edmund's College of University of Cambridge (to P.H.) and the European Union via ERC DeepCell-101054957 and BetaRegeneration-101054564 (to F.J.T.). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

K.H., L.S., M.D.L. and F.J.T. conceived the project. K.H., L.S., M.D.L. and G.H. wrote the manuscript with the support of other authors. V.A.S. collected information about existing single-cell datasets and L.S., M.S. and K.H. collected information about published atlases. V.A.S. performed the analysis and wrote the sections on methods. K.H., V.A.S. and L.S. prepared the figures. M.D.L. and F.J.T. supervised the work. All authors revised the manuscript.

Competing interests

G.H. and H.W. are employees of Genentech whose views are their own and do not represent those of Genentech, Roche or affiliates. M.D.L. contracted for the Chan Zuckerberg Initiative, consults for CatalYm and received speaker fees from Pfizer and Janssen Pharmaceuticals. S.A.T. has consulted for or been a member of scientific advisory boards at Qiagen, Sanofi, GlaxoSmithKline and ForeSite Labs. S.A.T. is a cofounder and an equity holder of TransitionBio and EnsoCell and a SAB member of Element Biosciences and an independent non-executive director on the 10X Genomics board. S.A.T. is a part-time employee at GlaxoSmithKline. F.J.T. consults for Immunai, Singularity Bio B.V., CytoReason, Cellarity and Curie Bio Operations and has an ownership interest in Dermagnostix and Cellarity. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02532-y.

Correspondence and requests for materials should be addressed to Fabian J. Theis or Malte D. Luecken.

Peer review information *Nature Methods* thanks Junyue Cao and Zizhen Yao for their contribution to the peer review of this work. Primary handling editor: Lin Tang, in collaboration with the *Nature Methods* team.