Genome Biology

REVIEW

Open Access

Insights, opportunities, and challenges provided by large cell atlases

Martin Hemberg[1,2*†], Federico Marini[3,4*†], Shila Ghazanfar[5,6,7*†], Ahmad Al Ajami[8,9,10], Najla Abassi[3,4], Benedict Anchang[11], Bérénice A. Benayoun[12,13], Yue Cao[5,6,7,14], Ken Chen[15], Yesid Cuesta-Astroz[16,17], Zachary DeBruine[18], Calliope A. Dendrou[19], Iwijn De Vlaminck[20], Katharina Imkeller[8,9,10], Ilya Korsunsky[2,21,22], Alex R. Lederer[23], Jessica Jingyi Li[24,35], Pieter Meysman[25], Clint L. Miller[26], Kerry A. Mullan[25], Uwe Ohler[27], Pratibha Panwar[5,6,7], Nikolaos Patikas[1,2], Jonas Schuck[8,9,10], Jacqueline H. Y. Siu[19], Timothy J. Triche Jr.[28,29], Alex Tsankov[30], Sander W. van der Laan[26,31], Masanao Yajima[32], Jean Yang[5,6,7,14], Fabio Zanini[33] and Ivana Jelic[34*]

†Martin Hemberg, Federico Marini, and Shila Ghazanfar contributed equally to this work.

†All authors participated in the series of meetings organized by CZI Single-Cell Biology program meetings.

*Correspondence: mhemberg@bwh.harvard.edu; marinif@uni-mainz.de; shila.ghazanfar@sydney.edu.au; ijelic@chanzuckerberg.com

[1] The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, Boston, USA
[3] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Mainz, Mainz, Germany
[5] School of Mathematics and Statistics, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia
[34] Chan Zuckerberg Initiative, Redwood City, USA
Full list of author information is available at the end of the article

Abstract

The field of single-cell biology is growing rapidly, generating large amounts of data from a variety of species, disease conditions, tissues, and organs. Coordinated efforts such as CZI CELLxGENE, HuBMAP, Broad Institute Single Cell Portal, and DISCO allow researchers to access large volumes of curated datasets, including more than just scRNA-seq data. These resources have created an opportunity to build and expand the computational biology ecosystem to develop tools necessary for data reuse and for extracting novel biological insights. We highlight achievements made so far, areas where further development is needed, and specific challenges that need to be overcome.

Introduction

Technological advances have enabled generation and collection of large volumes of data at the single-cell resolution [1]. For the most part, this is done by individual research groups, and to make these datasets more useful to the community, they need to be assembled into cell atlases. In the context of single-cell technologies, an atlas is a large collection of datasets that have been curated and made accessible through a web portal. In addition to making it easier to find datasets, atlases provide a coherent pipeline for data ingestion and processing, ensuring that datasets can be combined and leveraged to provide novel biological insights.

Today, there are thousands of single-cell datasets available, and building an atlas is a resource-intensive endeavor, requiring a large team of biologists and data scientists. Moreover, substantial infrastructure is needed, and to be useful to the community, it

must be sustained and updated over time. Hence, cell atlases are backed by large organizations or supported by large-scale projects (Table 1).

Together, we have been involved in the Chan Zuckerberg Initiative's effort to expand the ecosystem of computational methods that can support and exploit cell atlases in the context of the Data Insights program. Here, we present our shared experiences, and we discuss some of the issues involved in building and using a cell atlas (Fig. 1). We then go on to explore the possibilities enabled by cell atlases as well as the challenges that the community faces going forward.

Data ingestion, access, and representation

A central goal for any scientific resource is to make sure that it adheres to the FAIR principles [15], i.e., ensuring that data is findable, accessible, interoperable, and reusable. By serving as central repositories, cell atlases make it easier to find and access data. By making sure that data is uniformly processed and adheres to standard formats, it also becomes interoperable and reusable. Although straightforward in principle, the scale and complexity of a cell atlas make it difficult to achieve these goals.

Data pre-processing

For sequencing data to be useful, one must have access to the underlying reads, typically stored in fastq format [16]. In addition to the raw data, the various levels of processed data and metadata must also be ingested. The first step for the cell atlas is to carry out quality controls to ensure the integrity of the data. Thus, preprocessing is a key step that is often poorly documented and difficult to reproduce due to the use of different versions of software packages. Although preprocessing can improve the internal consistency, it does not prevent the emergence of discrepancies within and across atlases. A particular concern for cell atlases is batch effects, technical artifacts that emerge due to differences in how the data was obtained and processed. Although batch effects can be reduced, they cannot be eliminated altogether. Fortunately, it is possible to detect and correct batch effects post hoc, provided that detailed information about the processing is available. As no repository will be able to span all conditions, populations, organisms, cells, and modalities of interest, maintaining this possibility is a key requirement for enabling meta-analyses.

Data accessibility, interoperability, and reusability

By providing a portal where users can search for data, cell atlases greatly facilitate finding datasets. Depending on the use case, different levels of processing will be desired. Only providing the raw data is not sufficient, and various levels of processing will be required by most users. However, tools for indexing, metadata standardization, and cross-cohort queries are in their infancy, and this limits the ability of users to identify suitable datasets [2, 17]. Although it is possible to find and access individual datasets through a web browser, anyone interested in analyzing a large number of datasets needs to have both programming skills and sufficient computational resources. This creates a barrier for many users, and an important area of research is to develop computational tools to facilitate access to large cell atlases. Another key aspect is to provide APIs for those users who are developing code for accessing the cell atlas. This includes adhering

Table 1 Summary statistics for selected single cell atlases

Name	# cells	# donors	Year started	Organization	# species	URL	Reference
CZ CELLxGENE Discover	112.8 M	5 k	2022 (launch)	CZI	7	https://cellxgene.cziscience.com/	https://doi.org/10.1101/2023.10.30.563174 [2]
Single Cell Portal	57.6 M	Not reported	2020 (launch)	Broad Institute	18	https://singlecell.broadinstitute.org/single_cell	https://doi.org/10.1101/2023.07.13.548886 [3]
Single Cell Expression Atlas	13.5 M	Not reported	2018 (launch)	EMBL-EBI	21	https://www.ebi.ac.uk/gxa/sc/home	https://doi.org/10.1093/nar/gkab1030 [4]
Human BioMolecular Atlas Program (HuBMAP)	Not reported	214	2018 (launch)	NIH	1	https://hubmapconsortium.org/	https://doi.org/10.1038/s41556-023–01194-w [5]
Human Cell Atlas (HCA)	65.4 M	9.6 k	2016 (launch)	HCA	1	https://data.humancellatlas.org/	https://doi.org/10.7554/eLife.27041 [6]
Allen Brain Cell (ABC) Atlas	4.0 M	Not reported	2023 (publication)	Allen Institute	1	https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas	https://doi.org/10.1038/s41586-023–06812-z [7]
Tumor Immune Single Cell Hub 2 (TISCH2)	6.3 M	Not reported	2023 (publication)	Shanghai Putuo District People's Hospital	2	http://tisch.comp-genomics.org/	https://doi.org/10.1093/nar/gkac959 [8]
Deeply Integrated Single-Cell Omics (DISCO)	125.6 M	Not reported	2022 (publication)	Singapore Immunology Network	1	https://www.immunesinglecell.org	https://doi.org/10.1093/nar/gkab1020 [9]
Single Cell Atlas	200 M	125	2024 (publication)	SCA Consortium	1	https://www.singlecellatlas.org	https://doi.org/10.1186/s13059-024–03246-2 [10]
Asian Immune Diversity Atlas (AIDA)	1.3 M	625	2023 (launch)	HCA-Asia	1	https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508	https://doi.org/10.1016/j.cell.2025.02.017 [11]

Table 1 (continued)

Name	# cells	# donors	Year started	Organization	# species	URL	Reference
Curated Cancer Cell Atlas (3CA)	5.6 M	Mix of patient samples, cancer cell lines, organoids and mouse models	2025 (publication)	Weizmann Institute of Science	Mix of patient samples, cancer cell lines, organoids and mouse models	https://www.weizmann.ac.il/sites/3CA/	https://doi.org/10.1038/s43018-025–00957-8 [12]
single-cell multimodal omics (scMMO)-atlas	3.2 M	Not reported	2025 (publication)	Chinese Academy of Sciences	2	https://www.biosino.org/scMMO-atlas/	https://doi.org/10.1093/nar/gkae821 [13]
PlaqView	1.7 M	157 patients and 7 different mice models	2022	UVA & UMC Utrecht	2	https://www.plaqview.com	https://doi.org/10.3389/fcvm.2022.969421 [14]

A selection of cell atlases currently available. The number of cells corresponds to the approximate number of cells with transcriptomics readout at the time of writing

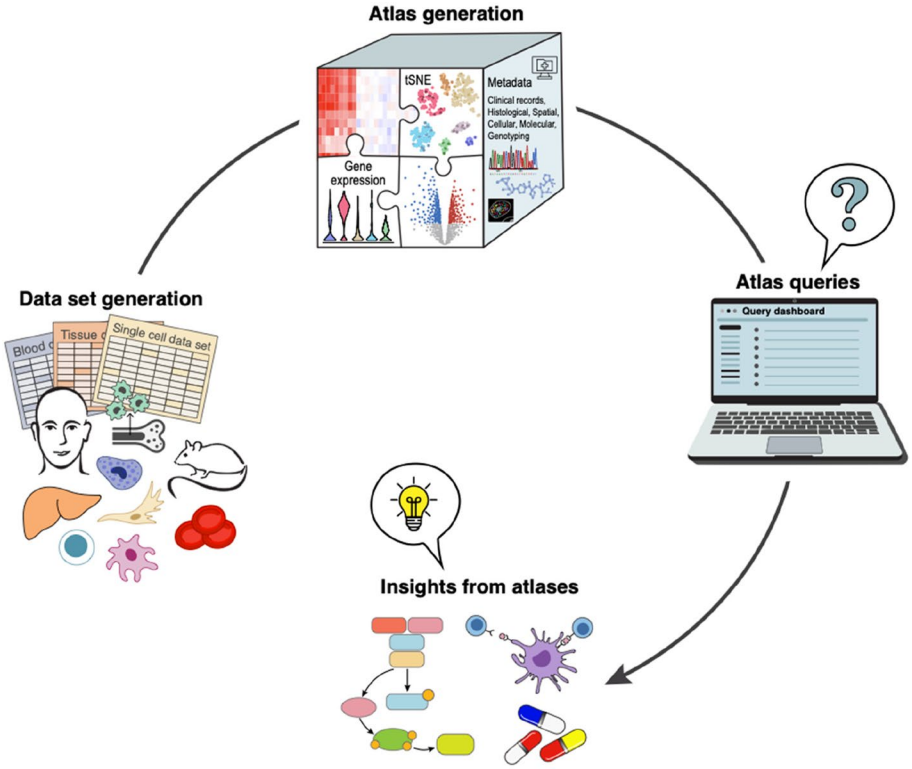

Fig. 1 Cell atlases ingest data from a wide range of labs based on specific criteria, e.g., species, disease, and tissue. Data is processed in a coherent manner and made available through a portal. The portal enables a wide range of queries to either download or interrogate multiple datasets. On their own or in combination with additional experiments, these queries can result in new findings

to standardized file formats as well as catering for multiple programming languages. At the time of writing, both R and Python are widely used, and cell atlases need to support both to be useful. As with many other aspects of a cell atlas, these resources need to be updated over time as the computational ecosystem and use cases evolve.

Metadata and ontologies

Metadata is crucial for researchers interested in re-analyzing existing datasets. Complete and well-curated metadata can transform a cell atlas from being a static reference to a dynamic hypothesis-generating tool, enabling a variety of stratified analyses. For example, metadata capturing timepoints post-infection enables reconstruction of disease trajectories. On the other hand, missing or incomplete metadata could mislead data interpretation (e.g., in a human study, effects that are actually due to donor sex or age would erroneously be attributed to treatment). This issue has long been recognized, and in the past, there have been community efforts such as MIAME [18] to set common standards. For scRNAseq, metadata can be divided into three categories: sample, gene, and cell. Sample metadata includes information about donor, time of collection, storage, experimental processing, computational processing, etc. Gene metadata is relatively straightforward, at least for model organisms, where various annotations are mature and stable. The most important aspect of cell metadata is its annotation, and this requires

mapping data to a cell type ontology [19]. An example of a metadata scheme developed for single-cell analysis is matrix and metadata standards (MAMS [20]). Although reporting and adhering to technical standards is important, it is essential to couple this to the establishment of a culture where data generators recognize their responsibility in providing complete metadata.

Ontologies allow formal and structured operations to be carried out, and they are thus essential to contextualize the resource and facilitate interpretations. In particular, they enable automated processing and application of ML/AI methods. The Cell Ontology [21] provides a standardized vocabulary to annotate cell types and states, which is vital for ensuring interoperability across datasets. Cell type annotation is a central step in the biological interpretation, yet it remains one of the most time-consuming tasks during the analysis. The traditional approach is to first identify marker genes for each cluster and then use the literature to determine which is the corresponding cell type or state. Several computational tools seek to assist with this process by comparing to previously annotated datasets [22, 23], and this has been highlighted as one of the main use cases of cell atlases. However, when aggregating datasets, it is often the case that they have not been annotated consistently, and assigning a consistent set of cell labels remains a major challenge. No universally accepted definition of cell types exists; however, as our knowledge of cell biology and identities is ever improving, consequently, ontologies must remain flexible to accommodate the diversity of cell types, states, and conditions. In this regard, we think that tools for harmonizing and standardizing annotations, such as [23] and [24], as well as initiatives such as the HuBMAP Common Coordinate Framework, will play an essential role in improving data consistency and enhancing the utility of atlases across contexts.

Cross-species comparisons can be particularly challenging due to differences not only in nomenclature but also in function. These analyses can be assisted by retaining gene/feature-level metadata and leveraging orthology-based information, and tools such as SAMap [25] can bridge gaps between different phyla by robustly reconstructing the latent manifolds. The integration of disease data into atlases also presents its own set of difficulties as disease-associated cells may acquire distinct cell states, requiring expanded annotations in the metadata. Moreover, when querying a "normal" atlas with disease-specific data, researchers must account for the potential lack of representation or mismatch in cell types [26]. For instance, cancer cells often recapitulate developmental gene programs, meaning these cells may map to normal developmental stages rather than the typical "diseased" cell types.

Extracting the most out of a cell atlas

The immediate use of cell atlases is to provide a global overview of cell types and cell states for a given tissue, disease, organism, or condition. An inventory of the building blocks is of great scientific value, and once generated, it serves as a springboard to address further biological questions. The challenge for the research community is that there is a breadth of needs for accessing a cell atlas. For some researchers, it may be enough to access a web server that displays gene expression and cell clusters, while bespoke analysis access may require downloading cell atlases to a local computer or server.

Data representation and subsampling

The typical workflow when using a cell atlas requires a user to first identify and download the relevant datasets. Given their size, already exceeding 1 TB using standard data structures, this can require significant bandwidth and be prohibitive to many research labs without high-memory computing resources. For most users, working with this data requires out-of-core processing, high-performance computing, and significant effort in data wrangling [27]. Consequently, there is an urgent need for methods for handling streaming data as well as lossless compression algorithms for single-cell data that significantly reduce memory footprint without compromising computational performance. New data structures for R/C++/Python can reduce memory footprint by up to tenfold over standard sparse matrices with minimal cost to compute. This has been achieved through substantial efforts in handling memory limitations and enabling processing of large datasets through the adoption of disk-backed or pyramidal data formats. Examples include Zarr, Parquet, and TileDB.

One algorithmic approach for dealing with large datasets is to subsample. Subsampling can ease computation over diverse datasets and help reduce bias of highly represented signals, but may also compromise the unprecedented modeling power that comes with a dataset of this size. Issues include racial and gender bias in samples [28], over-representation of specific cell types [29], opportunistic collection of rare samples [30], but the best approach depends on the scientific goals of the study. Simple random subsampling does not address signal representation and can miss rare subpopulations [31, 32], and it is desirable to balance the tradeoff between representing the true proportions of cell types and full extent of cell identities of rare cell types [33]. As cell atlases become more diverse across tissues and patient donors, subsampling will become more attractive. However, such summaries may impair our ability to tease out subtle and biologically relevant signals that only come with the massive statistical power offered by large sample sizes [34]. We also envision that latent space representations will be highly useful for providing compact representation, but further research is required to understand their accuracy and limitations. One promising line of work is the "biosketching" approach [32], which enables efficient and structure-preserving summarization of large-scale single-cell data. Another complementary strategy involves the construction of metacells, as developed in a series of works aiming to create compact, less sparse representations that preserve essential transcriptional structure while reducing noise (e.g., [35, 36]). While these approaches may sacrifice single-cell resolution, they offer potential advantages in terms of interpretability, robustness, and computational tractability.

Data integration and meta-analysis

To relate cell atlases with each other or with additional single-cell datasets, the research community relies on data integration tools [37]. These tools aim to identify joint low-dimensional representations of all the input data, with which further downstream analysis can be performed, e.g., joint clustering, cell type classification, and differential abundance testing. In modern single-cell analysis, this data integration is often key for harmonizing distinct datasets, and it serves as an initial step for meta-analysis of single-cell datasets. Appropriate meta-analyses will not only have to consider corrected cell type labels, but also other statistical factors that remain challenging to deal with.

Issues such as confounding factors, nested structure of single cells measured within biological samples (i.e., repeated measures), and understanding hidden sources of variability all may be relevant. Ideally, cells are curated with sample-level information such as donor identity, sex, age, tissue, organism, developmental stage, technology, and disease. Some of these confounders are technical, and performing data integration over them will remove noise and increase salience of biological signals [38]. Others are biologically driven, and thus data integration would enable researchers to compare analogous cell states across tissue, diseases, and organisms. A useful data integration algorithm must account for all of these sources of information and allow users to retain or remove specific variation relevant to their analyses. For instance, a comparison of T cells across tissues may emphasize tissue-specific differences (i.e., do not harmonize over tissue) to explain differential heritability, while others look to nominate for shared effector phenotypes across diseases (i.e., do harmonize over tissue) for basket clinical trials [39]. These scenarios highlight that data integration is not a static tool to be applied prior to further analysis, but rather needs to be adaptable to the research question at hand. This introduces a new computational challenge, as current integration algorithms are designed with the idea that they are run once per analysis, and therefore are not necessarily optimized to an online dynamic query setting. Successful data integration algorithms must address these emerging complexities, and they must (1) scale to 10,000 s of confounder levels (corresponding to the number of donors), (2) account for all sources of technical and biological variation in a way that lets users select which to account for, (3) perform consistently across a wide variety of cell atlas queries, (4) be fast and flexible enough to integrate diverse queries on the fly, (5) provide views of their impact on data distortion and signal degradation in a manner that is easy for the user to interpret. Numerous methods for carrying out batch integration have been proposed in recent years, resulting in substantial progress in terms of performance, as demonstrated by independent benchmarks [37, 40, 41]. However, several challenges remain, and existing methods often struggle in more complex scenarios, e.g., involving different species [42], imbalanced datasets [43], or very large numbers of cells. As quantitative metrics of batch integration provide an incomplete view, evaluation of the five criteria above will also require careful considerations based on the biological interpretations.

Building cell atlases in context

An important use of cell atlases, beyond defining cell types and corresponding markers, is to explore how cellular and transcriptional landscapes are impacted by specific disease/functional decline (e.g., disease states, reduced function), physiological/biological factors (e.g., age, sex/gender, ethnic/genetic background), or treatments (e.g., response to a drug) [44–46]. Indeed, important insights can be achieved by analyzing cell atlases in a context-aware fashion, both comparing how biologically relevant inputs can lead to changes in (i) cell composition [47–49] and (ii) cell-type specific gene expression. Importantly, even for single-cell atlases, biological replicates should consist of samples procured from independent biological sources/individuals, and not just independent cells from the same sample [50–52]. Thus, a key feature of context-aware cell atlases is the inclusion of sufficient independent biological samples across conditions to account for inter-individual variability. This requires sufficient numbers of true biological replicates,

similar to bulk approaches [53]. For the interpretation of context-dependent cell atlases, it is crucial to consider the need for approaches that limit the high false positive rates in single-cell differential analyses (e.g., considering the potential of pseudobulking approaches per identified cell type/state to avoid underestimation of true biological variability [50, 51]). As an example, in the study of menopause and its potential molecular drivers, a context-aware atlas should include sufficient samples covering both pre-menopausal and post-menopausal states, in tissues most relevant to the condition (e.g., ovary, pituitary gland, hypothalamus) [54].

Benchmarking and development of novel methods

There is currently a rapidly expanding ecosystem of computational tools in the single-cell field, and for most problems, there is more than one method available. To help researchers decide what tool to choose, benchmarking studies are essential, and multiple benchmarking papers are published every month. Several challenges exist in the current single-cell benchmarking field, and guidelines on best practices are needed. It needs to be clear what the evaluation metrics are, and although many details will depend on the specific topic, there are overarching trends [55]. One of the main challenges when comparing methods is that for most problems, we do not have an independent ground truth. Hence, evaluating the performance will involve some degree of subjectivity. One way of circumventing this challenge is to use simulations to create synthetic data. However, most synthetic datasets are unable to capture the full spectrum of complexities found in real datasets, and more work is required to build on recent developments [56]. Comparisons using real datasets require curation, a process that can be time-consuming and requires substantial domain knowledge. Moreover, it may further entrench any value of certain algorithms through the process of performing the curation (e.g., clustering algorithms). As such, there is tremendous value in datasets that can be considered as a "gold standard" by virtue of orthogonal means. Carefully curated cell atlases can serve an important role here by being commonly used for benchmarking specific analytical tasks.

Today, most analysis tools and strategies are designed within the context of a smaller number of datasets or total cells. Many methods that are commonly used may not scale well to tens of millions of cells and thousands of conditions, and consequently, there is a need to increase computational and algorithmic efficiency. This is likely to involve various types of approximations and lossy compression to achieve the desired speed-up and reduction in memory footprint. An example is the use of strategies like mini-batch to speed up the estimation of k-means clustering [57] without compromising the quality of results.

Alongside improvements to existing analytical strategies, a wider reformulation of these methods to address new biological questions is needed to fully leverage datasets spanning tissues, species, and organismal age. One such example of an established method that requires novel frameworks to be applicable to new types of data is RNA velocity. A manifold-constrained and biologically tailored velocity, as opposed to general-purpose tools, could be designed to statistically compare estimates in the case–control setting [58, 59]. With such a modular method, a diseased or otherwise perturbed sample could be used to investigate whether subtle disruptions of the RNA velocity vectors indicate an effect of a particular perturbation.

Recently, there have been tremendous advances in AI, in particular in applications related to natural language processing, protein structure prediction, and image processing. What these methods have in common is that they rely on large datasets for training, and consequently cell atlases will help their advance. Several groups have developed foundation models leveraging the large-scale data collected via cell atlas efforts, e.g., Geneformer [60], scGPT [61], scFoundation [62], scBERT [63], CellFM [64], UCE [65], and atlas approximations [66]. Foundation models learn generalizable representations of cell type and state from gene expression profiles, and they can be used to annotate new datasets, project them into shared latent spaces, infer missing modalities, and simulate responses to genetic or pharmacological perturbations. Despite these advances, widespread practical use remains limited. Current challenges include technical barriers to applying these models in user-friendly interfaces, dealing with memory and computational infrastructure requirements, limited interpretability and explainability of model predictions and representations, and the lack of widespread stress-testing of models on noisy, rare and disease-specific single cell datasets to improve trust. Still, the field is advancing rapidly, with several models already being applied to biologically meaningful tasks such as cross-species comparison [66, 67] and spatial pathology integration [68]. As tools become more accessible and validated, we anticipate that AI will increasingly serve as a bridge between cell atlases and translational insight.

Using cell atlases for biomedical research

Perhaps the most important application underpinning efforts to build cell atlases is the notion that they can help accelerate biomedical research to help manage and cure disease [69, 70]. Below, we discuss some of the areas where cell atlases will provide key resources.

From large-scale genome-wide association studies (GWAS), thousands of genetic loci have been identified that infer risk of disease or influence human traits. While these studies have yielded great and unexpected insights into complex and common diseases, they also reveal a yawning knowledge gap. For instance, for 50% of the risk loci identified for coronary artery disease (CAD) it is unclear which gene(s) and which cell(s), and therefore which molecular and cellular pathways may be involved [71]. From studying CAD-associated loci, it is clear that gene regulatory effects are context-dependent, and that genetic effects can be condition-specific in terms of effect, direction, and magnitude [72]. In other words, biological sex, environmental factors (e.g., smoking), and disease influence genotypic effects and change cellular gene expression and responses. The analysis of cell type and condition-specific genetic effects on cellular molecular traits quantitative trait locus (molQTL) analysis will provide critical insights. These efforts could enable cell type and cell state-specific colocalization of molQTL and GWAS signals to interpret the regulatory mechanisms for complex diseases and traits. Cell atlases that combine genetic information with cell molecular profiles will be a key resource in unraveling such complex regulatory effects.

Cell atlases can also facilitate therapeutic target discovery, e.g., by predicting disease-relevant cell states by identifying gene signatures along a trajectory from healthy to disease [73]. For example, muscle cells downregulate their contractile markers and become mesenchymal stem-like cells before adopting specialized cell states [74]. There are now

automated pipelines (e.g., scDrug, Drug2Cell [75, 76]) that take as input a cell-by-gene matrix of protein-coding genes and leverage the full compendium of drug-gene interactions (e.g., DGIdb [77]), cell perturbations (e.g., LINCS L1000 [78]), FDA-approved molecules and biologics (e.g., DrugBank [79]), or active ligands (e.g., ChEMBL [80]) to prioritize potential drug target genes. Importantly, atlases could also be used to predict drug responses or unwanted side effects (e.g., scDR [81]) by querying identified targets in public databases (e.g., SIDER [82]). Such a combination of cell atlas and therapeutic databases would enable the combination of genetic epidemiology, in particular, causal inference through Mendelian randomization, with single-cell biology, resulting in effective identification of druggable targets or surrogate markers of disease.

The large collections of data presented in cell atlases require appropriate tools and frameworks that enable efficient exploration. Only in this way can they fulfill their potential in assisting a wide spectrum of researchers to generate novel hypotheses, faithfully representing the results, and communicating the findings with the community. Several platforms and interfaces that aim to simplify the extraction of insight from such datasets have proliferated over the last decade, including, for example, the CELLxGENE tool [2], the Bioconductor iSEE package [83], Vitessce [84], and the browsers included in the Broad Single Cell Portal or the Single Cell Expression Atlas [85]. To illustrate how a typical user might interact with a cell atlas, a researcher interested in the expression of a fibroblast-associated gene (e.g., *COL1A2*) in tendons can filter by tissue, select relevant cell types, and visualize expression levels across conditions (healthy vs. after acute injury, [86]). The spectrum of operations covered by such tools enables a powerful, in-depth exploration, possibly blending different views and representations of these large corpora of data, and linking out to other existing databases or relevant resources. For a more systematic guide to best practices in interacting with integrated atlases, we refer readers to the work of [87].

Beyond atlases of dissociated single-cell transcriptomes

Single-cell RNA-seq was the first high-throughput method that allows for a high-plex characterization of individual cells, and thus it has been the most widely used approach [1, 88]. However, there are numerous other single-cell technologies under active development, and we foresee that over the coming years cell atlases will see an influx of other modalities [89]. These include TCR and BCR sequencing, ATAC-seq, and long-read sequencing. Although this is likely to be hugely beneficial to researchers, it also involves several different challenges. This starts with the organizations supporting the cell atlas, which must develop standards and protocols for how to process and curate other modalities [90]. Ensuring that different modalities can be combined for joint analyses is key, but it will present some major challenges, e.g., developing pre-processing pipelines, ontologies, and determining what metadata to include.

Several assays have been developed for measuring other aspects of the cellular state in single cells, e.g., DNA methylation, accessible chromatin (ATAC-seq), and transcription factor binding (scCUT&Tag), and an active area of research is to apply them to the same cell for multiomics profiling. This will provide numerous opportunities, e.g., by combining ATAC-seq and RNA-seq data, we are likely to improve our ability to infer gene regulatory networks [91]. However, integrating such data across tissues, donors, and

platforms at the cell atlas scale presents substantial challenges. As multi-omics datasets become more complex and heterogeneous, future integration frameworks will need to account for missing modalities, differing noise characteristics, and scale. Recent reviews have highlighted methods based on shared latent spaces and graph-based integration as promising approaches for atlas-scale applications [92], while recent benchmarking efforts have examined the performance and scalability to sufficiently large (atlas-scale) datasets [93].

Perhaps the most important direction of new technologies is toward spatial methods, primarily for transcriptomics and proteomics, but other modalities are likely to follow. Spatial data brings numerous challenges along with great potential for additional insights [94]. One challenge is in visualizing the data, and here, a user should be able to seamlessly toggle between gene expression space (typically a UMAP) and physical space. This representation is relatively straightforward for individual datasets, and indeed is implemented in cell atlas interfaces such as CxG, but for multiple samples it becomes much more challenging. For multiple samples, it may be more useful to map cells/spots to a common coordinate system, either informed by relative landmarks [95] or by merging across multiple samples [96]. There is also a need for further algorithmic development of methods, tools, and community standards for mining spatial data. Mining spatial data typically involves finding subcellular patterns in cells associated with disease, type, and outcomes [97], as well as spatial patterns from histopathological image data that can then be used across archival histology data without the companion omics layers [98]. As technologies evolve and are able to profile larger tissue sections at subcellular resolution, a particular challenge will be to develop methods that can bridge molecular, cellular, and tissue-level patterns. Given the success of machine learning in image analysis, spatial omics is well positioned to benefit from cross-pollination with computer vision and AI techniques.

Cell atlases outreach

At the moment, cell atlases are being built by scientists for other scientists [99]. However, given the potential implications to wider society and the substantial amounts of resources, much of which is coming from public funding, it is important that cell atlases can also cater to other audiences [99, 100]. Beyond the core constituency of academic biomedical researchers, potential users include clinicians and researchers in biotech and pharmaceutical industries. However, we believe that the ambition should be to make the resource at least somewhat accessible to the general public, including patients, teachers, and students of all ages [69]. Public-facing presentations of cell atlas resources should emphasize aspects that are relatable, visually intuitive, and grounded in relevance to human health. For example, simplified representations of how cells function across organs, or how they change in common diseases, can be powerful tools for public engagement. Interactive visualizations, curriculum-aligned resources for educators, and narratives that link cellular biology to real-world medical advances (e.g., cardiovascular disease, cancer, infection, or aging) are particularly valuable for broad audiences. This has the potential of helping to educate the public on the advances and benefits of biomedical research, involving citizen scientists, and inspiring the next generation of scientists to ensure that others will be able to build on the work. To maximize the accessibility

and impact of these efforts, collaborations with science communication professionals can help ensure that messaging is accurate, inclusive, and engaging for non-specialist audiences.

Conclusions and outlook

Here, we have outlined some of the challenges and opportunities brought along by cell atlases. We foresee that over the coming years, the existing atlases will continue to grow and that additional, more specialized collections will emerge. Having multiple atlases is likely to be beneficial to the field as different policies for curation, representation, interaction, and use cases. One analogy is gene annotations, where resources such as Ensembl, Refseq, and Gencode continue to be used in parallel. Depending on the specific need, one of these overlapping and complementary resources will be the most useful. In parallel, we expect substantial advances in computational methods that can effectively work with atlas-scale datasets to extract new insights.

Cell atlases have been enabled by technological advances, and we envision that continued innovation will govern how cell atlases evolve. As costs fall, we also expect that atlases will be broadened. To realize the potential not just for biomedical research, but for other aspects of biology, we need better coverage of human populations across the age spectra and for multiple disease states. Moreover, additional species are needed. Most likely, within the next few years, advancements in the fields of proteomics and metabolomics will enable the profiling of large numbers of single cells via these modalities, unlocking more accurate modeling of metabolism, signaling, and cell–cell communication.

In addition to their biomedical applications, cell atlases are proving to be invaluable resources for fundamental research in developmental biology, comparative genomics, and evolutionary biology. Developmental atlases, such as those of the human fetus [101], human brain [102], zebrafish [103], fruit fly [104], and mouse [105], provide rich datasets for understanding cell fate decisions and lineage specification, both in humans and in an evolutionary context ("evo-devo") [106]. Comparative atlases across species, including non-model mammals, birds, and reptiles [107], as well as plant atlases [108, 109], enable insights into conserved and divergent cellular programs. Resources like SPEED [110] and the Malaria Cell Atlas [111] further extend atlas applications into ecology and parasite biology. Together, these resources demonstrate that cell atlases are not only tools for biomedical discovery but are also essential for addressing fundamental questions in cell and developmental biology, evolution, and organismal diversity.

The availability of large-scale data resources contributes to the democratization of science. In fact, we have reached a point where many projects are run using only public data. Single-cell data is information dense, and there is little chance that the lab that generated the data has the capacity to make all the discoveries that are possible, in particular those that are only possible by combining with other datasets. Enabling the efficient use and reuse of complex and multiple datasets allows our scientific community to increase the pace of scientific discoveries.

In summary, we have described the development and utilization of cell atlases, which are comprehensive maps of cell types and states generated through the integration of large volumes of single-cell data. We detailed challenges and opportunities in building,

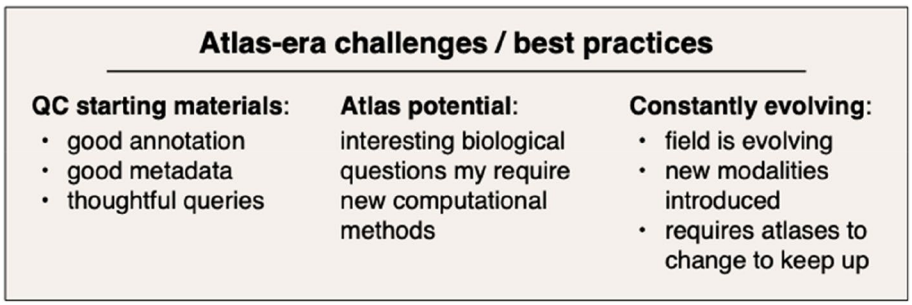

Fig. 2 Key challenges, opportunities, and issues regarding cell atlases today

maintaining, and utilizing these atlases (Fig. 2), emphasizing the importance of data standardization, accessibility, and computational tools for extracting meaningful biological insights, ultimately aiming to facilitate cross-tissue, cross-condition, and cross-species studies in the field of single-cell biology.

Acknowledgements
We would like to thank Leslie Gaffney for assistance with the figures. This work is funded by the Chan Zuckerberg Initiative.

Peer review information
George Inglis and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
Conceptualization: Ivana Jelic Writing: Martin Hemberg, Federico Marini, Shila Ghazanfar, with input from all other authors.

Funding
This work is funded by the Chan Zuckerberg Initiative.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, Boston, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Mainz, Mainz, Germany. [4]Research Center for Immunotherapy (FZI), Mainz, Germany. [5]School of Mathematics and Statistics, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia. [6]Sydney Precision Data Science Centre, University of Sydney, Sydney, NSW 2006, Australia. [7]Charles Perkins Centre, University of Sydney, Sydney, NSW 2006, Australia. [8]Neurological Institute/Edinger Institute, Goethe University, University Hospital Frankfurt, Frankfurt Am Main, Germany. [9]Goethe University, Frankfurt Cancer Institute, Frankfurt Am Main, Germany. [10]University Cancer Center (UCT), Frankfurt Am Main, Germany. [11]National Institute of Environmental Health Sciences, Durham, USA. [12]Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA 90089, USA. [13]Cancer Biology Department, USC Keck School of Medicine, Los Angeles, CA 90089, USA. [14]Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR, China. [15]The University of Texas MD Anderson Cancer Center, Houston, USA. [16]Escuela de Microbiología, Universidad de Antioquia, Ciudad Universitaria Calle 67 No 12 53-108, Medellín, Colombia. [17]Instituto Colombiano de Medicina Tropical, Universidad CES, 055413 Sabaneta, Colombia. [18]School of Computing, Grand Valley State University, Allendale, MI 49401, USA. [19]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK. [20]Meinig School of Biomedical Engineering, Cornell University, Ithaca, USA. [21]Brigham and Women's Hospital, Divisions of Genetics and Rheumatology, Boston, MA, USA. [22]Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA, USA. [23]Laboratory of Brain Development and Biological Data Science, School of Life Sciences, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. [24]Department of Statistics

and Data Science, Department of Biostatistics, Department of Computational Medicine, and, Department of Human Genetics, University of California, Los Angeles, CA, USA. [25]Adrem Data Lab, Department of Computer Science, University of Antwerp, Antwerp, Belgium. [26]Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA. [27]Max Delbruck Center for Molecular Medicine, Berlin, Germany. [28]Department of Epigenetics, Van Andel Institute, Grand Rapids, MI, USA. [29]Department of Pediatrics and Human Development, Michigan State University College of Human Medicine, East Lansing, MI, USA. [30]Icahn School of Medicine at Mount Sinai, New York, USA. [31]Central Diagnostic Laboratory, Division Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands. [32]Boston University, Boston, USA. [33]School of Clinical Medicine, UNSW Sydney, Sydney, NSW 2052, Australia. [34]Chan Zuckerberg Initiative, Redwood City, USA. [35]Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA.

Received: 14 June 2024 Accepted: 3 September 2025
Published online: 20 October 2025

References

1. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database. 2020. Available from: https://doi.org/10.1093/database/baaa073.
2. CZI Cell Science Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Res. 2025;53(D1):D886-900.
3. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. bioRxiv. 2023. Available from: https://doi.org/10.1101/2023.07.13.548886.
4. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Res. 2022;50(D1):D129–40.
5. Jain S, Pei L, Spraggins JM, Angelo M, Carson JP, Gehlenborg N, et al. Advances and prospects for the human biomolecular atlas program (HuBMAP). Nat Cell Biol. 2023;25(8):1089–100.
6. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017. https://doi.org/10.7554/eLife.27041.
7. Yao Z, van Velthoven CTJ, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624(7991):317–32.
8. Han Y, Wang Y, Dong X, Sun D, Liu Z, Yue J, et al. TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. Nucleic Acids Res. 2023;51(D1):D1425–31.
9. Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, et al. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Res. 2022;50(D1):D596-602.
10. Pan L, Parini P, Tremmel R, Loscalzo J, Lauschke VM, Maron BA, et al. Single Cell Atlas: a single-cell multi-omics human cell encyclopedia. Genome Biol. 2024;25(1):104.
11. Asian diversity in human immune cells. Cell. 2025;188(8):2288-306.e24.
12. Tyler M, Gavish A, Barbolin C, Tschernichovsky R, Hoefflin R, Mints M, et al. The curated cancer cell atlas provides a comprehensive characterization of tumors at single-cell resolution. Nat Cancer. 2025. https://doi.org/10.1038/s43018-025-00957-8.
13. Cheng W, Yin C, Yu S, Chen X, Hong N, Jin W. scMMO-atlas: a single cell multimodal omics atlas and portal for exploring fine cell heterogeneity and cell dynamics. Nucleic Acids Res. 2025;53(D1):D1186–94.
14. Ma WF, Turner AW, Gancayco C, Wong D, Song Y, Mosquera JV, et al. PlaqView 2.0: A comprehensive web portal for cardiovascular single-cell genomics. Front Cardiovasc Med. 2022;8(9):969421.
15. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;15(3):160018.
16. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger FASTQ file format for sequences with quality scores, and the solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71.
17. Fischer DS, Dony L, König M, Moeed A, Zappia L, Heumos L, et al. Sfaira accelerates data and model reuse in single cell genomics. Genome Biol. 2021;22(1):248.
18. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29(4):365–71.
19. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7(1):44.
20. Sarfraz I, Wang Y, Shastry A, Teh WK, Sokolov A, Herb BR, et al. MAMS: matrix and analysis metadata standards to facilitate harmonization and reproducibility of single-cell data. Genome Biol. 2024;25(1):1–15.
21. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, et al. Cell type ontologies of the human cell atlas. Nat Cell Biol. 2021;23(11):1129–35.
22. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29.
23. Xu C, Prete M, Webb S, Jardine L, Stewart BJ, Hoo R, et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. Cell. 2023;186(26):5876-91.e20.
24. Bian H, Chen Y, Wei L, Zhang X. uHAF: a unified hierarchical annotation framework for cell type standardization and harmonization. Bioinformatics. 2025. https://doi.org/10.1093/bioinformatics/btaf149.
25. Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. Elife. 2021;10. Available from: https://doi.org/10.7554/eLife.66747.
26. Dann E, Cujba AM, Oliver AJ, Meyer KB, Teichmann SA, Marioni JC. Precise identification of cell states altered in disease using healthy single-cell references. Nat Genet. 2023;55(11):1998–2008.

27. Ruiter S, Wolfgang S, Tunnell M, Triche T Jr, Carrier E, DeBruine Z. Value-Compressed Sparse Column (VCSC): Sparse matrix storage for redundant data. 2023. Available from: https://arxiv.org/abs/2309.04355.
28. Vaidya A, Chen RJ, Williamson DFK, Song AH, Jaume G, Yang Y, et al. Demographic bias in misdiagnosis by computational pathology models. Nat Med. 2024;30(4):1174–90.
29. Truong DD, Lamhamedi-Cherradi SE, Porter RW, Krishnan S, Swaminathan J, Gibson A, et al. Dissociation protocols used for sarcoma tissues bias the transcriptome observed in single-cell and single-nucleus RNA sequencing. BMC Cancer. 2023;23(1):488.
30. Haniffa M, Taylor D, Linnarsson S, Aronow BJ, Bader GD, Barker RA, et al. A roadmap for the human developmental cell atlas. Nature. 2021;597(7875):196–205.
31. Song D, Xi NM, Li JJ, Wang L. scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. Bioinformatics. 2022;38(11):3126–7.
32. Hie B, Cho H, DeMeo B, Bryson B, Berger B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. Cell Syst. 2019;8(6):483-93.e7.
33. Liang S, Willis J, Dou J, Mohanty V, Huang Y, Vilar E, et al. Sensei: how many samples to tell a change in cell type abundance? BMC Bioinformatics. 2022;23(1):1–22.
34. Schmid KT, Höllbacher B, Cruceanu C, Böttcher A, Lickert H, Binder EB, et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. Nat Commun. 2021;12(1):1–18.
35. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20(1):1–19.
36. Bilous, Mariia, Léonard Hérault, Aurélie AG Gabriel, Matei Teleman, and David Gfeller. Building and analyzing metacells in single-cell genomics data. Molecular Systems Biology. 2024; Available from: https://www.embopress.org/doi/10.1038/s44320-024-00045-6.
37. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41–50.
38. Emmanúel Antonsson S, Melsted P. Batch correction methods used in single cell RNA-sequencing analyses are often poorly calibrated. bioRxiv. 2024. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.03.19.585562.
39. Mullan KA, Ha M, Valkiers S, de Vrij N, Ogunjimi B, Laukens K, et al. T cell receptor–centric perspective to multimodal single-cell data analysis. Science Advances. 2024. Available from: https://www.science.org/doi/10.1126/sciadv.adr3196. Cited 2025 Jun 25.
40. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2018;16(1):43–9.
41. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16(12):1289–96.
42. Hrovatin K, Moinfar AA, Zappia L, Lapuerta AT, Lengerich B, Kellis M, et al. Integrating single-cell RNA-seq datasets with substantial batch effects. bioRxiv. 2024. Available from: https://www.biorxiv.org/content/10.1101/2023.11.03.565463v2.abstract. Cited 2025 Jun 13. p. 2023.11.03.565463.
43. Maan H, Zhang L, Yu C, Geuenich MJ, Campbell KR, Wang B. Characterizing the impacts of dataset imbalance on single-cell data integration. Nat Biotechnol. 2024;42(12):1899–908.
44. Thomas T, Rich-Griffin C, Pohin M, Friedrich M, Aschenbrenner D, Pakpoor J, et al. A longitudinal single-cell therapeutic atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. bioRxiv. 2023. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.05.05.539635.
45. Ren X, Wen W, Fan X, Hou W, Su B, Cai P, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell. 2021;184(7):1895-913.e19.
46. Chu Y, Dai E, Li Y, Han G, Pei G, Ingram DR, et al. Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. Nat Med. 2023;29(6):1550–62.
47. Phipson B, Sim CB, Porrello ER, Hewitt AW, Powell J, Oshlack A. propeller: testing for differences in cell type proportions in single cell data. Bioinformatics. 2022;38(20):4720–6.
48. Cao Y, Lin Y, Ormerod JT, Yang P, Yang JYH, Lo KK. scDC: single cell differential composition analysis. BMC Bioinformatics. 2019;20(Suppl 19):721.
49. Miller SA, Policastro RA, Sriramkumar S, Lai T, Huntington TD, Ladaika CA, et al. LSD1 and Aberrant DNA Methylation Mediate Persistence of Enteroendocrine Progenitors That Support -Mutant Colorectal Cancer. Cancer Res. 2021;81(14):3791–805.
50. Crowell HL, Soneson C, Germain PL, Calini D, Collin L, Raposo C, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11(1):6077.
51. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12(1):5692.
52. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40(2):245–53.
53. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22(6):839–51.
54. Singh PP, Benayoun BA. Considerations for reproducible omics in aging research. Nat Aging. 2023;3(8):921–30.
55. Cao Y, Yu L, Torkel M, Kim S, Lin Y, Yang P, et al. The current landscape and emerging challenges of benchmarking single-cell methods. bioRxiv. 2023. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.12.19.572303.
56. Song D, Wang Q, Yan G, Liu T, Sun T, Li JJ. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat Biotechnol. 2024;42(2):247–52.
57. Hicks SC, Liu R, Ni Y, Purdom E, Risso D. Mbkmeans: fast clustering for single cell data using mini-batch k-means. PLoS Comput Biol. 2021;17(1):e1008625.
58. Aivazidis A, Memi F, Kleshchevnikov V, Er S, Clarke B, Stegle O, et al. Cell 2fate infers RNA velocity modules to improve cell fate prediction. Nat Methods. 2025;22(4):698–707.

59. Lederer AR, Leonardi M, Talamanca L, et al. Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. Nat Methods. 2024;21:2271–86. https://doi.org/10.1038/s41592-024-02471-8.
60. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24.
61. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. ScGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024. https://doi.org/10.1038/s41592-024-02201-0.
62. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21(8):1481–91.
63. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nature Machine Intelligence. 2022;4(10):852–66.
64. Zeng Y, Xie J, Shangguan N, Wei Z, Li W, Su Y, et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. Nat Commun. 2025;16(1):1–17.
65. Rosen Y, Roohani Y, Agrawal A, Samotorcan L, Tabula Sapiens Consortium, Quake SR, et al. Universal cell embeddings: A foundation model for cell biology. bioRxiv. 2023. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.11.28.568918.
66. Xu Y, Gatt C, Kaymak E, Kikuchi K, Bourguignon T, Zanini F. Deep exploration of transcriptomic cellular identities over evolutionary time. bioRxiv. 2025. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.02.19.639005.
67. Yang X, Liu G, Feng G, Bu D, Wang P, Jiang J, et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. Cell Res. 2024;34(12):830–45.
68. Huang T, Liu T, Babadi M, Ying R, Jin W. STPath: A generative foundation model for integrating spatial transcriptomics and whole slide images. bioRxiv. 2025. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.04.19.649665.
69. Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the human cell atlas on medicine. Nat Med. 2022;28(12):2486–96.
70. Rood JE, Wynne S, Robson L, Hupalowska A, Randell J, Teichmann SA, et al. The Human Cell Atlas from a cell census to a unified foundation model. Nature. 2024;637(8048):1065–71.
71. Chen Z, Schunkert H. Genetics of coronary artery disease in the post-GWAS era. J Intern Med. 2021;290(5):980–92.
72. Aherrahrou R, Lue D, Perry RN, Aberra YT, Khan MD, Soh JY, et al. Genetic Regulation of SMC Gene Expression and Splicing Predict Causal CAD Genes. Circ Res. 2023;132(3):323–38.
73. Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov. 2023;22(6):496–520.
74. Grootaert MOJ, Bennett MR. Vascular smooth muscle cells in atherosclerosis: time for a re-assessment. Cardiovasc Res. 2021;117(11):2326–39.
75. Hsieh CY, Wen JH, Lin SM, Tseng TY, Huang JH, Huang HC, et al. scDrug: From single-cell RNA-seq to drug response prediction. Comput Struct Biotechnol J. 2023;21:150–7.
76. Kanemaru K, Cranley J, Muraro D, Miranda AMA, Ho SY, Wilbrey-Clark A, et al. Spatially resolved multiomics of human cardiac niches. Nature. 2023;619(7971):801–10.
77. Cannon M, Stevenson J, Stahl K, Basu R, Coffman A, Kiwala S, et al. DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. Nucleic Acids Res. 2024;52(D1):D1227-35.
78. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437-52.e17.
79. Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265-75.
80. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D1180–92.
81. Lei W, Yuan M, Long M, Zhang T, Huang YE, Liu H, et al. scDR: predicting drug response at single-cell resolution. Genes. 2023;14(2):268. https://doi.org/10.3390/genes14020268.
82. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075–9.
83. Rue-Albrecht K, Marini F, Soneson C, Lun ATL. iSEE: Interactive SummarizedExperiment Explorer. F1000Res. 2018;7:741.
84. Keller MS, Gold I, McCallum C, Manz T, Kharchenko PV, Gehlenborg N. Vitessce: integrative visualization of multimodal and spatially resolved single-cell data. Nat Methods. 2025;22(1):63–7.
85. Moreno P, Huang N, Manning JR, Mohammed S, Solovyev A, Polanski K, et al. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. Nat Methods. 2021;18(4):327–8.
86. Mimpen JY, Baldwin MJ, Paul C, Ramos-Mucci L, Kurjan A, Cohen CJ, et al. Exploring cellular changes in ruptured human quadriceps tendons at single-cell resolution. J Physiol. 2025. Cited 2025 Jun 13; Available from: https://onlinelibrary.wiley.com/doi/abs/10.1113/JP287812.
87. Hrovatin K, Sikkema L, Shitov VA, Heimberg G, Shulman M, Oliver AJ, et al. Considerations for building and using integrated single-cell atlases. Nat Methods. 2025;22(1):41–57.
88. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599–604.
89. Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. Nat Methods. 2020;17(1):11–4.
90. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24(8):550–72.
91. Zhang K, Zemke NR, Armand EJ, Ren B. A fast, scalable and versatile tool for analysis of single-cell omics data. Nat Methods. 2024;8:1–11.
92. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24(8):494–515.

93. Liu C, Ding S, Kim JH, Long S, Xiao D, Ghazanfar S, et al. Multi-task benchmarking of single-cell multimodal omics integration methods. bioRxiv. 2025. Cited 2025 Jun 13. p. 2024.09.15.613149. Available from: https://www.biorxiv.org/content/10.1101/2024.09.15.613149v2.abstract.
94. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med. 2022;14(1):1–18.
95. Yayon N, Kedlian VR, Boehme L, Suo C, Wachter BT, Beuschel RT, et al. A spatial human thymus cell atlas mapped to a continuous tissue axis. Nature. 2024;635(8039):708–18.
96. Zeira R, Land M, Strzalkowski A, Raphael BJ. Alignment and integration of spatial transcriptomics data. Nat Methods. 2022;19(5):567–75.
97. Dries R, Chen J, del Rossi N, Khan MM, Sistig A, Yuan GC. Advances in spatial transcriptomic data analysis. Genome Res. 2021;31(10):1706–18.
98. Wang C, Chan AS, Fu X, Ghazanfar S, Kim J, Patrick E, et al. Benchmarking the translational potential of spatial gene expression prediction from histology. Nat Commun. 2025;16(1):1–17.
99. Quake SR. A decade of molecular cell atlases. Trends Genet. 2022;38(8):805–10.
100. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. 2019;574(7777):187–92.
101. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science. 2020. https://doi.org/10.1126/science.aba7721.
102. Chen X, Huang Y, Huang L, Huang Z, Hao ZZ, Xu L, et al. A brain cell atlas integrating single-cell transcriptomes across human brain regions. Nat Med. 2024;30(9):2679–91.
103. Lange M, Granados A, VijayKumar S, Bragantini J, Ancheta S, Kim YJ, et al. A multimodal zebrafish developmental atlas reveals the state-transition dynamics of late-vertebrate pluripotent axial progenitors. Cell. 2024;187(23):6742-59.e17.
104. Calderon D, Blecher-Gonen R, Huang X, Secchia S, Kentro J, Daza RM, et al. The continuum of embryonic development at single-cell resolution. Science. 2022 Aug 5;377(6606):eabn5800.
105. Imaz-Rosshandler I, Rode C, Guibentif C, Harland LTG, Ton MLN, Dhapola P, et al. Tracking early mammalian organogenesis - prediction and validation of differentiation trajectories at whole organism scale. Development. 2024. https://doi.org/10.1242/dev.201867.
106. Wei J, Liu P, Liu F, Jiang A, Qiao J, Pu Z, et al. EDomics: a comprehensive and comparative multi-omics database for animal evo-devo. Nucleic Acids Res. 2023;51(D1):D913–23.
107. Chen D, Sun J, Zhu J, Ding X, Lan T, Wang X, et al. Single cell atlas for 11 non-model mammals, reptiles and birds. Nat Commun. 2021;12(1):1–17.
108. He Z, Luo Y, Zhou X, Zhu T, Lan Y, Chen D. scPlantDB: a comprehensive database for exploring cell types and markers of plant cell atlases. Nucleic Acids Res. 2024;52(D1):D1629–38.
109. Vong GYW, McCarthy K, Claydon W, Davis SJ, Redmond EJ, Ezer D. AraLeTA: An Arabidopsis leaf expression atlas across diurnal and developmental scales. Plant Physiol. 2024;195(3):1941–53.
110. Chen Y, Zhang X, Peng X, Jin Y, Ding P, Xiao J, et al. SPEED: Single-cell Pan-species atlas in the light of Ecology and Evolution for Development and Diseases. Nucleic Acids Res. 2022;51(D1):D1150–9.
111. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K, et al. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. Science. 2019. Available from: https://www.science.org/doi/10.1126/science.aaw2619. Cited 2025 Jun 13.

Publisher's Note