scBaseCamp: an AI agent-curated, uniformly processed, and continually expanding single cell data repository

Nicholas D. Youngblut[*,1], Christopher Carpenter[*,1], Jaanak Prashar[1,2], Chiara Ricci-Tam[1], Rajesh Ilango[1], Noam Teyssier[1], Silvana Konermann[‡,1,2], Patrick D. Hsu[‡,1,3], Alexander Dobin[‡,1], David P. Burke[‡,1], Hani Goodarzi[‡,†,1,4], Yusuf H. Roohani[‡,†,1,2]

[1]Arc Institute; [2]Stanford University; [3]University of California, Berkeley; [4]University of California, San Francisco

Abstract

Building a virtual model of the cell is an emerging frontier at the intersection of artificial intelligence and biology, aided by the rapid growth of single-cell RNA sequencing data. By aggregating gene expression profiles from millions of cells across hundreds of studies, single cell atlases have provided a foundation for training AI-driven models of the cell. However, reliance on datasets with pre-processed counts limits the size and diversity of these repositories and constrains downstream model training to data curated for divergent purposes. This introduces analytical variability due to differences in the choice of alignment tools, genome references, and counting strategies. Here, we introduce scBaseCamp, a continuously updated single-cell RNA-seq database that leverages an AI agent-driven hierarchical workflow to automate discovery, metadata extraction, and standardized data processing. Built by directly mining and processing all publicly accessible 10X Genomics single-cell RNA sequencing reads, scBaseCamp is currently the largest public repository of single-cell data, comprising over 230 million cells spanning 21 organisms and 72 tissues. Using studies comprised of both single cell and single nucleus sequencing data, we demonstrate that uniform processing across datasets helps mitigate analytical artifacts introduced by inconsistent data processing choices. This standardized approach lays the groundwork for more accurate virtual cell models and serves as a foundation for a wide range of biological and biomedical applications.

*These authors contributed equally to this work.
‡These authors jointly supervised this work.
†Corresponding authors: Y.R. (yusuf.roohani@arcinstitute.org), H.G. (hani.goodarzi@arcinstitute.org)

1. Introduction

The ability to precisely measure the transcriptomic state of individual cells has transformed the study of cell biology. By uncovering the heterogeneous states of cells as they engage in diverse processes and functions across species and tissue contexts, single-cell RNA sequencing has revealed the finer details of cell identity and behavior that were previously inaccessible through bulk approaches. These discoveries have significantly influenced various fields, from developmental biology to cancer research (Rood et al., 2022; Eraslan et al., 2022; Melms et al., 2021). More importantly, the growing scale of these single-cell datasets has driven efforts to build *in silico* models of the cell—artificial intelligence (AI) models designed to capture context-dependent cellular function and behavior and predict cellular responses to perturbations (Lopez et al., 2018; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Rosen et al., 2023). In many ways, building a "Virtual Cell" model has emerged as a major frontier in the application of AI to biology (Bunne et al., 2024).

Over the past decade, interest in integrating single-cell datasets across institutions and research laboratories has grown significantly. Notable efforts, such as the Human Cell Atlas (Rozenblatt-Rosen et al., 2017) and the CZ CELLxGENE dataset (CZI Cell Science Program et al., 2025), have made substantial strides in expanding the availability of curated single-cell RNA sequencing (scRNA-seq) datasets. These efforts have proven instrumental in advancing our understanding of cell identity, differentiation trajectories, and disease mechanisms, while also providing valuable training data for AI-driven modeling of cellular states. However, these initiatives primarily rely on contributed datasets, which is only a subset of the publicly available data that is accessible through the NIH-hosted Sequence Read Archive (SRA)—the largest repository of raw single-cell sequencing data (Leinonen et al., 2011). While this current approach allows for deeper, expert-level dataset curation and annotation, it also limits the scale and diversity of data available for analysis—particularly for AI models, which typically do not rely on cell labeling. This underscores the need for a new approach to single-cell genomics data curation—one that operates at the scale required for AI model training and is free from the constraints of manual dataset annotation.

Another challenge with existing repositories of single cell data is that aggregation of datasets from diverse sources introduces analytical variability due to differences in alignment tools, reference genomes, and read counting strategies. In the bulk RNA-seq domain, large-scale reanalysis efforts like the Recount initiative have previously demonstrated the power of standardized pipelines in minimizing these analytical batch effects (Frazee et al., 2011). By uniformly reprocessing bulk RNA sequencing data, Recount provided researchers with a resource where biological variation was not impacted by inconsistencies in data processing pipelines. Drawing on these lessons, we recognized the need for a similar data repository for single-cell genomics—one that spans a wide range of species and tissues while adhering to a consistent processing standard. Such a resource should enable more robust cross-study comparisons, facilitate meta-analyses, and better support AI-driven research aimed at modeling cell behavior across diverse biological contexts.

Here, we present scBaseCamp, a single-cell genomics data repository that represents the largest-scale reprocessing effort to date. An effort that will continue to expand as new data becomes available on SRA. scBaseCamp was built by leveraging an AI-driven agent to automate repository identification and metadata unification, enabling continuous discovery, annotation, and standardized processing of raw single-cell RNA-seq data. As a result, scBaseCamp not only provides a harmonized, large-scale resource for AI-driven modeling and integrative meta-analyses but also remains dynamic, growing alongside the ever-expanding landscape of publicly available single-cell data.

2. Results

2.1. scBaseCamp, a large, diverse and actively expanding single-cell data repository

scBaseCamp is the first comprehensive single cell database built by directly mining all publicly accessible 10X Genomics scRNA-seq data from the Sequence Read Archive (SRA) and applying a standardized processing pipeline to improve data harmonization. Using an AI-driven agent, `SRAgent`, we systematically and continually identify repositories, unify metadata across diverse sources, and facilitate the discovery, annotation, and reprocessing of raw single-cell RNA-seq data.

To date, our `SRAgent` has identified 63,892 SRA experiments (i.e. SRX entries), of which 43,587 were labeled as 10X Genomics sequencing libraries. With the inclusion of 6,059 additional samples currently part

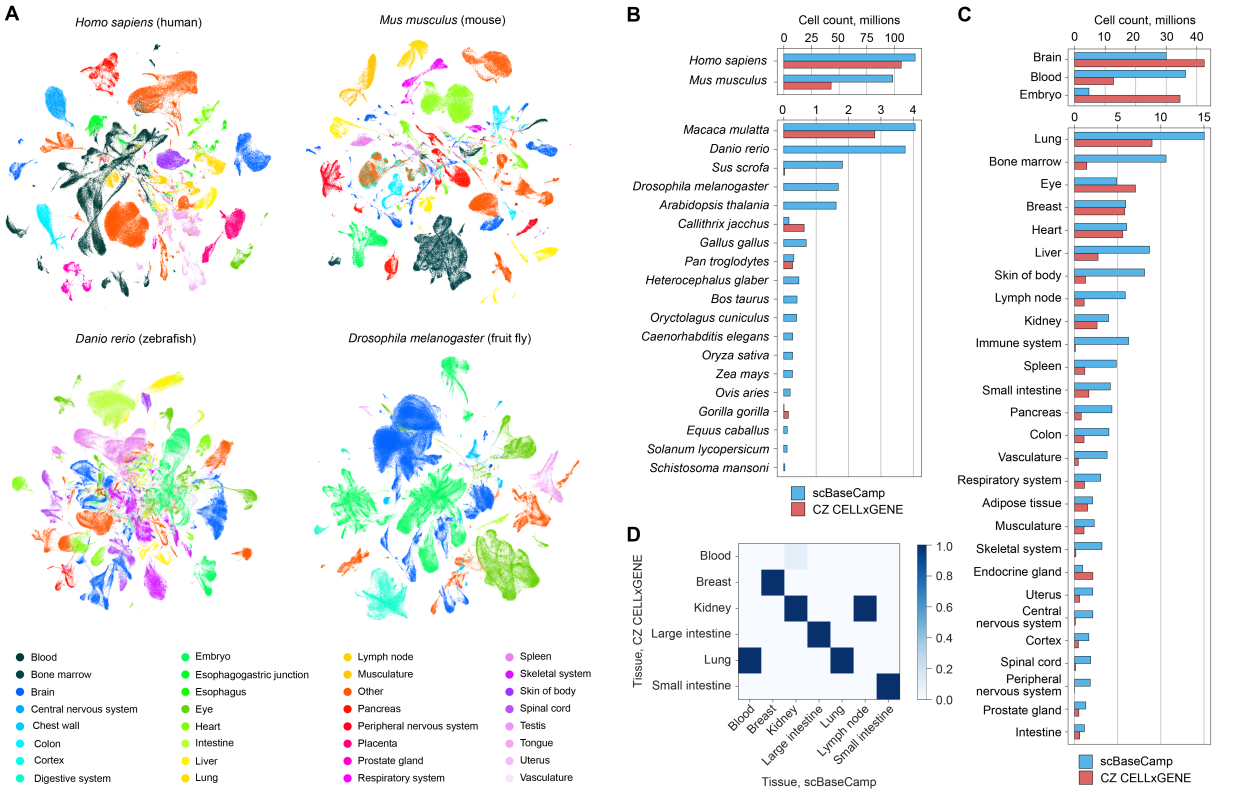

Figure 1 | **scBaseCamp: The largest publicly available repository for single-cell gene expression datasets across species and tissues.** (**A**) UMAP visualization of a random sampling of cells from scBaseCamp, colored by tissue type. Each panel represents a different species: *Homo sapiens* (human; N=243, 807), *Mus musculus* (mouse; N=249, 008), *Danio rerio* (zebrafish; N=501, 041), and *Drosophila melanogaster* (fruit fly; N=500, 877). (**B**) Comparison of cell distributions across species in scBaseCamp versus CZ CELLxGENE, highlighting a broader species representation in scBaseCamp. (**C**) Comparison of cell distributions across the top 30 tissues (all mammals) in scBaseCamp and CZ CELLxGENE (human and mouse), illustrating the increased tissue diversity and representation in scBaseCamp. (**D**) Confusion matrix comparing `SRAgent`'s automated tissue annotations in scBaseCamp with tissue labels from CZ CELLxGENE, demonstrating high concordance in most cases.

of CZ CELLxGENE (CZI Cell Science Program et al., 2025), we have identified a total of 49,646 SRX entries at the time of this writing. Thus far, we have reprocessed 30,387 entries.

Currently, scBaseCamp comprises over 230 million cells, with an average of 7,614 unique molecular identifiers (UMIs) per cell (**Figure 1A**). When compared to other large repositories of single cell data, such as CZ CELLxGENE (107 million cells) and Human Cell Atlas (65 million cells), scBaseCamp is already the largest collection of publicly available single-cell datasets at the time of this writing (**Figure S1A**). With data from 21 organisms and 72 tissues, scBaseCamp provides a significantly broader range of experimental contexts compared to the current largest repository, CZ CELLxGENE (**Figure 1B,C** and **Figure S1B,C**).

In addition to dataset identification, `SRAgent` also attempts to extract key metadata for each SRX, including 10X chemistry, cell vs nuclei suspension type, and associated diseases and tissues. For instance, when comparing `SRAgent`'s automated tissue annotations to those in CZ CELLxGENE, we found that in most cases, the agent accurately extracted the correct tissue labels (**Figure 1D**). This observation aligns with recent studies demonstrating the effectiveness of large language models in cell type annotation (Kazmi et al., 2025; Hou and Ji, 2024; Chen and Zou, 2024; Xiao et al., 2024). While AI models of cell state do not rely on these labels during training, `SRAgent`'s ability to perform reliable tissue labeling enhances confidence in its data curation

and annotation capabilities.

2.2. Automated single-cell data discovery and annotation using `SRAgent`

To systematically identify and integrate single-cell RNA-sequencing datasets, we developed `SRAgent`, a hierarchical agentic workflow built with LangGraph around large language models (LLMs) and specialized tools for querying the Sequence Read Archive (SRA). Specifically, `SRAgent` employs a hierarchical workflow of ReAct agents (Yao et al., 2022) that asynchronously access `eSearch`, `eSummary`, `eFetch`, `eLink`, NCBI HTML scraping, SRA BigQuery, `sra-stat`, and `fastq-dump`. This workflow continuously mines publicly accessible 10X Genomics datasets, retrieves key metadata (e.g., organism, tissue, disease, perturbation), and stores these annotations in a relational database (**Figure 2A**). This automated approach enables rapid discovery of new studies while ensuring metadata curation remains consistent and scalable.

`SRAgent` is deployed on GCP Cloud Run, utilizing 2 CPUs and 2 GB of memory per job. To avoid exceeding NCBI API rate limits, jobs are triggered every 1-5 minutes, processing 3-5 datasets per run, with a peak rate of up to 300 datasets per hour. In total, `SRAgent` has processed 63,892 datasets, of which 43,587 were identified as 10X Genomics sequencing libraries.

As part of this publication, we have made the `SRAgent` code publicly available. The research community can access and utilize `SRAgent` to further expand single-cell dataset discovery and integration efforts.

2.3. Standardized data re-processing using `scRecounter`

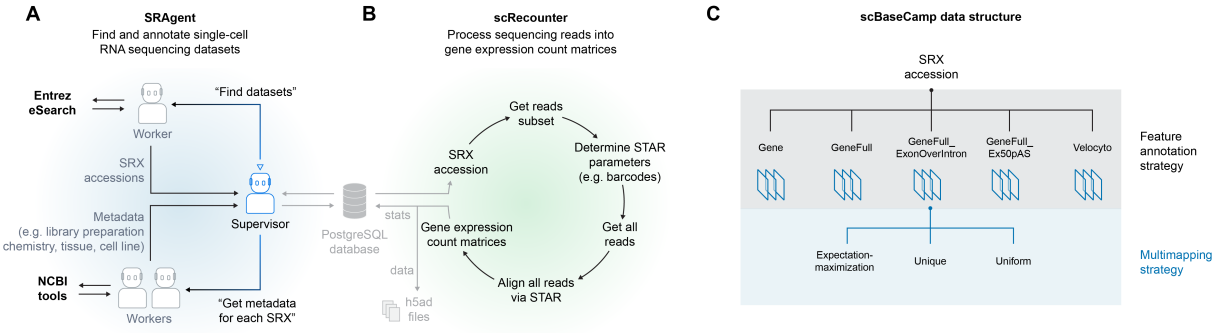

Figure 2 | **AI-driven curation and standardized processing of single-cell RNA-seq data in scBaseCamp.** (**A**) `SRAgent` workflow: A hierarchical AI-driven pipeline for automated dataset discovery and metadata curation from the Sequence Read Archive (SRA). `SRAgent` systematically queries NCBI tools (e.g., `eSearch`, `eFetch`) to identify 10X Genomics datasets, retrieve metadata (e.g., tissue type, library preparation chemistry), and store structured annotations in a GCP SQL database. (**B**) `scRecounter` pipeline: A `Nextflow`-based workflow for processing raw single-cell sequencing reads into gene expression count matrices. `scRecounter` downloads and aligns sequencing reads using `STARsolo`, automatically detects optimal barcode parameters, and generates harmonized expression matrices stored in h5ad format. Process tracking is managed via a PostgreSQL database hosted on GCP. (**C**) `scRecounter` uses multiple feature annotation and multimapping strategies to generate a variety of cell×gene count tables. Users can choose the option that best suits their application. We currently provide access to all variations in scBaseCamp.

To process the 10X Genomics datasets identified by `SRAgent`, we developed `scRecounter`, a custom `Nextflow` (Di Tommaso et al., 2017) pipeline designed for efficient and scalable single-cell data reprocessing. `scRecounter` runs on GCP Cloud Run, with jobs submitted to GCP Batch. To stay within the GCP resource quota limits, the pipeline processes a maximum of three datasets per run, with new runs triggered every three minutes.

Since a specific 10X Genomics chemistry version cannot be reliably identified from SRA metadata, we developed a simple algorithm for its automatic identification. For each identified dataset, 1 million paired-end reads are first downloaded via `fastq-dump`, which are then used to determine the chemistry and proper

`STARsolo` (Kaminow et al., 2021) parameters (10X Genomics cell barcode version, UMI length, and strandedness) based on the number of valid barcodes. Then, the entire datasets are downloaded using `fasterq-dump` and mapped with `STARsolo` to generate cell×gene count tables. To date, a total of 7.7×10^{12} reads have been processed with the `scRecounter` pipeline.

For gene annotations, we used human and mouse reference genome from the 10X Genomics repository. `STAR` references for other species were generated using the following workflow: for each species, a widely used ENSEMBL genome assembly was selected and downloaded (Frölich et al., 2023). As with human and mouse annotations, we applied a custom script to filter each species' GTF file, retaining only relevant biotypes (e.g., "protein-coding" and "lncRNA") based on their clade (e.g., mammals, birds, fungi). This filtering ensures that only genes measurable by single-cell RNA-seq are included and maintains consistency with 10X Genomics annotations. For `STAR` reference generation, we adjusted the `genomeSAindexNbases` parameter according to genome size, following recommendations from the `STAR` developers (**Figure 2B**) (Dobin et al., 2013). These annotations and references are available for download through our portal.

In analyzing single-cell RNA-seq data, researchers often make choices regarding gene models and counting strategies, as there is no universal approach that fits all applications. Depending on the specific analysis—whether focused on standard gene expression quantification, alternative splicing, or transcriptional dynamics—different gene models and counting strategies may be more appropriate. With `scRecounter`, our goal is to provide a comprehensive set of possibilities, enabling researchers to select the most suitable approach for their own study. To achieve this, `scRecounter` offers multiple feature annotation strategies, including Gene, GeneFull, GeneFull_Ex50pAS, GeneFull_ExonOverIntron, and Velocyto (La Manno et al., 2018) (see Methods). These models vary in their treatment of exonic and intronic regions, allowing users to choose between standard gene counts (GeneFull), expanded annotations that capture antisense transcription (GeneFull_Ex50pAS), or intron-aware quantification designed for RNA velocity analysis (Velocyto) (**Figure 2C**).

Additionally, `scRecounter` is flexible in terms of multimapping strategies, which determine how multimapping reads are handled. These include expectation-maximization (EM), which probabilistically assigns multimapped reads, as well as unique and uniform strategies, which either retain only uniquely mapped reads or distribute multimapped reads evenly (**Figure 2C**).

By incorporating these diverse options, `scRecounter` ensures that scBaseCamp remains a versatile resource, accommodating a broad range of single-cell RNA-seq applications while minimizing technical biases. As part of this publication, we have also made the `scRecounter` code available.

2.4. Uniform data re-processing reduces the contribution of analytical confounders

A key advantage of a uniform processing pipeline for single-cell transcriptomic data is its ability to minimize technical variability introduced by differences in analytical pipelines. To assess this effect, we used silhouette scoring (Büttner et al., 2019), a metric that quantifies how well individual data points cluster based on a given categorical variable. The silhouette score ranges from -1 to 1, where higher values indicate that cells are more cohesively grouped within the same category and well-separated from other categories, while lower values suggest overlapping or poorly defined clusters. We applied this approach to evaluate how various categorical metadata variables influence gene expression patterns in scBaseCamp. Some factors primarily reflect biological variation (e.g., tissue type), while others capture a mix of technical and biological variation (e.g., sample suspension type, library preparation chemistry, or sample ID). Our goal was to quantify the extent to which groupings based on each factor shape the dataset's structure and contribute to observed variation in gene expression. However, given the prohibitively large dataset size, we instead analyzed repeatedly sampled subsets of 500,000 cells and recorded the silhouette scores for each. As expected, in scBaseCamp, technical factors such as library preparation chemistry and sample suspension type (single-cell vs. single-nucleus sequencing) exhibited comparable or lower silhouette scores than biologically meaningful categories like tissue type (**Figure 3A**).

To further illustrate this point, we performed the same analysis on CZ CELLxGENE, which is the largest public repository of single-cell RNA sequencing datasets and is composed of pre-processed datasets that are not uniformly processed. We found that tissue type, which is a major biological source of variation, exhibited significantly higher silhouette scores in scBaseCamp than in CZ CELLxGENE. To assess the contribution of the more technical factors beyond the biological signal capture by tissue type, we normalized Silhouette scores

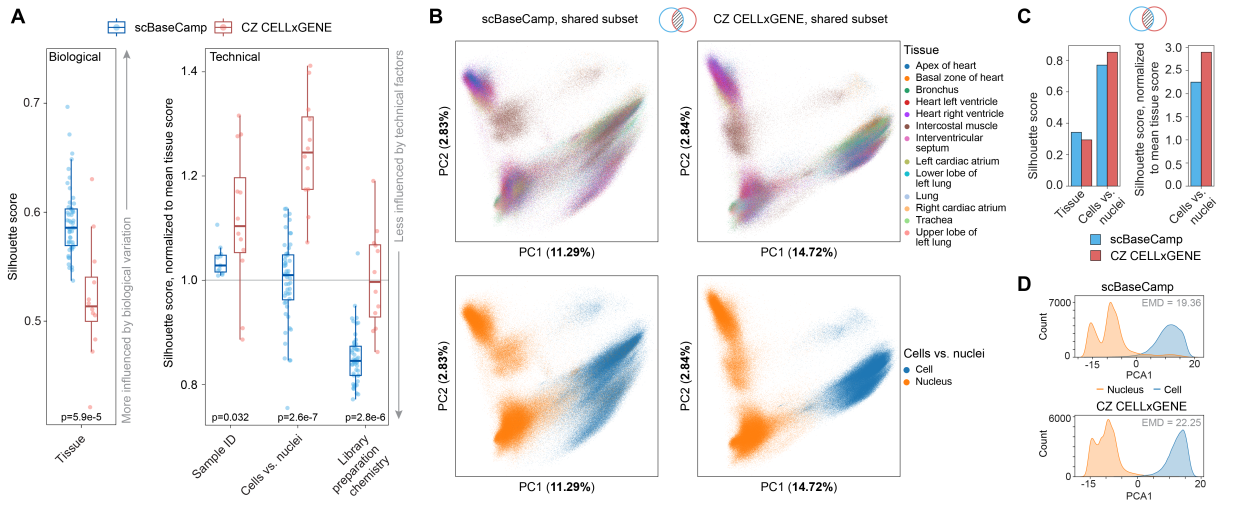

Figure 3 | **Comparing the impact of technical and biological factors on clustering in scBaseCamp and CZ CELLxGENE.** (**A**) Silhouette scores computed across random subsets of 250,000 cells from scBaseCamp (blue) and CZ CELLxGENE (red) for different metadata factors. Higher scores indicate a greater influence of the corresponding factor on data clustering. For technical factors, scores are normalized to the mean tissue score for each dataset. CZ CELLxGENE subsets were drawn from 20 individual human datasets and downsampled to 250,000 cells in two steps, and scBaseCamp subsets were randomly sampled from human and mouse accessions. The resulting objects were filtered to retain genes present in at least 10 cells and cells with at least 30 detected genes. All datasets were sum-normalized, log-transformed, and processed using PCA before calculating silhouette scores (`scib_metrics package`). These results demonstrate that, given the contribution of tissue type, technical factors such as sample ID, library preparation chemistry, and sample suspension type contribute less to clustering in scBaseCamp compared to CZ CELLxGENE, highlighting how scBaseCamp's uniform processing pipeline reduces analytical variability while preserving biological structure. (**B**) Principal Component Analysis (PCA) of a shared subset of cells present in both scBaseCamp and CZ CELLxGENE that contain both single cell and single nuclei data (N=566, 224). Left: PCA plots of the shared subset in scBaseCamp, colored by tissue type (top) and suspension type (bottom). Right: PCA plots of the same subset in CZ CELLxGENE, showing a greater separation between single-cell and single-nucleus samples along the first two principal components. (**C**) Silhouette scores in each dataset, as a function of tissue type, and sample suspension type (i.e. cells vs nulcei), confirming that sample suspension type is a stronger driver of variation in CZ CELLxGENE than in scBaseCamp, and tissue type is better preserved in scBaseCamp. (**D**) Distribution of cells and nuclei projected onto the first PC for scBaseCamp and CZ CELLxGENE. We observed a more pronounced separation between cells and nuclei in CZ CELLxGENE as measured by Earth Mover's Distance (EMD).

from other factors to average silhouette score from the tissue variable. Across all technical variables, CZ CELLxGENE consistently exhibited higher silhouette scores than scBaseCamp, indicating a greater influence of technical variation on its datasets. For example, sample ID (largely equivalent to the SRA study or SRP ID in scBaseCamp; **Figure S2**) behaves differently in the two datasets: in scBaseCamp, sample ID scores are similar to tissue type, while in CZ CELLxGENE, they are significantly higher, suggesting that study-specific processing introduces stronger batch effects (**Figure 3A**).

This effect is even more pronounced for sample suspension type, which differentiates single-cell from single-nucleus sequencing. In CZ CELLxGENE, silhouette scores for suspension type are substantially higher. We reasoned that this may likely be due to the common practice of counting single-nucleus data using both exonic and intronic reads (i.e. pre-mRNA annotations), while single-cell data typically includes only exonic reads. While neither approach is inherently incorrect, this inconsistency could introduce strong batch effects when integrating both data types. scBaseCamp mitigates this issue by retaining and reporting both exonic and intronic read counts, resulting in better integration of single-nucleus and single-cell datasets.

To further examine this possibility, we focused on four shared datasets, listed in Table S1, that contain

both single-cell and single nucleus-sequencing data. We performed Principal Component Analysis (PCA) and visualized the datasets in two-dimensional PC space (**Figure 3B**). In the 2-dimensional PC space, tissue type showed a higher silhouette score in scBaseCamp than in CZ CELLxGENE, and conversely, sample suspension type (i.e. cells vs nuclei) scored lower in scBaseCamp (**Figure 3C**). Given the inherent differences in sample preparation between these two profiling strategies, unsurprisingly, PC1 mainly separated single cells from single nuclei. However, we should further highlight two key observations. First, the variance explained by PC1 decreased in scBaseCamp compared to CZ CELLxGENE, dropping by approximately 4%—from 15% in CZ CELLxGENE to 11% in scBaseCamp. Second, as we have shown in **Figure 3D**, cells and nuclei were better separated on PC1 in CZ CELLxGENE than in scBaseCamp. Consistent with a reduced separation between cells and nuclei in scBaseCamp, the Earth Mover's Distance (EMD) was higher for CZ CELLxGENE for PC1 of this dataset. A higher EMD in this context indicates that PC1 more strongly separates gene expression profiles by technical variation rather than meaningful biological differences, suggesting that data processing choices amplify the differences between cells and nuclei in CZ CELLxGENE.

Examining the source publications of these four integrated datasets, we found that the single nucleus data in this subset was indeed processed with pre-mRNA references and the single cells were processed with a standard exonic reference, as is common practice (**Table S1**). Correcting this technical artifact in scBaseCamp resulted in better separation of tissue types along the first two PCs compared to CZ CELLxGENE, as reflected in their respective silhouette scores (**Figure 3B**). Together, these findings suggest that in CZ CELLxGENE version of this integrated dataset, more than 4% of variation in gene expression can be attributed to differences in data analysis processing than biological differences between single-cell and single-nucleus sequencing.

Beyond harmonization, uniform data processing in scBaseCamp—including consistent treatment of exonic and intronic reads across samples—enables additional downstream analyses, such as RNA velocity, to better model transcriptional dynamics and cell state transitions. As mentioned in the previous section, `scRecounter` provides cell×gene count tables for spliced and unspliced transcripts from Velocyto as well.

2.5. Planned expansion of scBaseCamp

To date, scBaseCamp has focused on publicly accessible 10X Genomics datasets available through SRA, with plans to extend to additional single-cell genomics technologies and data sources. Currently, scBaseCamp includes 81 collections—comprising 18 million cells—that overlap with CZ CELLxGENE, which itself contains over 60 million 10X Genomics cells. Some of these datasets originate from sources beyond SRA; for instance, we are in the process of incorporating 7 million cells from the NeMO Archive. However, accessing raw data for a significant number of additional cells requires access to protected datasets, a process that cannot be automated at scale and instead necessitates submission of manual requests. Despite this challenge, we have begun engaging with authors for access and we anticipate that this gap in scBaseCamp will gradually close. Looking ahead, we will also expand `scRecounter` beyond 10X Genomics data to support additional single-cell genomic platforms, broadening the scope of scBaseCamp even further.

3. Discussion

In this study, we present scBaseCamp, a continuously updated and uniformly processed multi-species single-cell RNA-seq data repository that is built through directly mining all publicly accessible 10X Genomics datasets. Our motivation stems from the need to create a large, high-quality resource for training computational models of cell state and behavior. Existing single-cell databases rely on manually curated datasets, which limits their size and diversity, failing to leverage the vast reserves of publicly available transcriptomic data. Moreover, integrating individual datasets within these repositories remains a major challenge due to variation in how each of these datasets were processed.

By systematically mining, annotating and processing raw single-cell transcriptomic reads directly from SRA, scBaseCamp is designed to be a large, harmonized repository of single-cell datasets with minimal analytical confounders. Inspired by the Recount initiative which demonstrated the power of uniformly processed bulk RNA-seq data, scBaseCamp aims to reduce technical artifacts across studies ensuring that the representations of cellular state learned by AI models are based on biologically meaningful variation. Additionally, scBaseCamp far exceeds other single cell databases in the diversity of species and tissue contexts that are represented, reflecting the breadth of phenotypic information that it contains.

scBaseCamp is the first large biological data repository curated by an AI agent. Our automated workflow, powered by the `SRAgent` AI system and a `Nextflow`-based processing pipeline, ensures consistent metadata curation and analytical treatment across all datasets. This agentic workflow has the advantage of being entirely automated, easily scalable and capable of continuously updating as new data becomes available. Unlike many existing resources, scBaseCamp reduces data processing assumptions by providing multiple gene count options, including both exonic and intronic counts. This flexibility allows researchers to tailor their analyses while supporting downstream applications such as RNA velocity (La Manno et al., 2018; Lange et al., 2022), which relies on intronic reads to infer cell state transitions.

Despite these advantages, there are several future directions and challenges to consider. First, while our agentic workflow automates metadata extraction, some aspects, e.g. cell type annotation, may require human expertise, community-sourced curation, or additional tools for mapping individual cells onto annotated references (Lopez et al., 2018; Lotfollahi et al., 2022; Domínguez Conde et al., 2022; Heimberg et al., 2024). Other cell-level annotation such as the applied perturbation or donor information are also not accessible through SRA-derived metadata and must be incorporated manually. Second, in the current release, we have focused on libraries created using the 10X Genomics platform, and sequenced on established Illumina sequencers. As alternative library preparation chemistries become more commonplace and sequencing technologies continue to evolve, we must adapt scBaseCamp to maintain compatibility with new modalities, including various chemistries, multi-omic measurements, and spatial transcriptomics.

Ultimately, this unified, continuously updated resource provides a foundation for AI-driven efforts aiming to build integrative models of cell behavior across health and disease. By providing the largest publicly available single-cell data repository that is also uniformly processed, we hope to reduce the barriers for researchers to perform large-scale computational model training and integrative analysis.

4. Data and Code Availability

4.1. Data availability

All `STARsolo` count matrices are located on Google Cloud Storage at `gs://arc-ctc-scbasecamp/2025-02-25`. Documentation on accessing the data can be found at `https://github.com/ArcInstitute/arc-virtual-cell-atlas`.

4.2. Code availability

`SRAgent` and `scRecounter` are available on GitHub at `https://github.com/ArcInstitute/SRAgent` and `https://github.com/ArcInstitute/scRecounter`, respectively. Code used for data analysis is available at `https://github.com/ArcInstitute/scBaseCamp_analysis`.

5. Methods

5.1. Identification of 10X chemistry

Once `SRAgent` predicted that an SRX contained 10X Genomics single-cell data, the dataset was processed through the `scRecounter` pipeline to determine the specific assay chemistry using a `Cell Ranger`-inspired automated detection algorithm. For this, the first 1 million reads from each SRX were processed using `STARsolo`, testing barcode sets corresponding to 10X 5', 10X 3' v2, 10X 3' v3, and 10X Multiome GEX. 10X 3' v1 was not included. The barcode file yielding the highest percentage of valid cell barcodes was selected. Datasets with fewer than 30% valid barcodes were removed at this stage.

For datasets identified as 10X 3' v2 or 10X 5', which share the same whitelist, an additional alignment was performed on the reverse strand. The final strand assignment was determined by selecting the strand with the higher gene alignment percentage, provided that it was at least twice that of the opposite strand. This approach replicates the strand detection strategy used by `Cell Ranger`, ensuring accurate chemistry identification.

5.2. Data processing

After identifying the correct 10X chemistry, we then aligned and counted all SRRs within the SRX accession as one library, in a single `STARsolo` run, using the following parameters:

```
--soloType CB_UMI_Simple
--clipAdapterType CellRanger4
--outFilterScoreMin 30
--soloCBmatchWLtype 1MM_multi_Nbase_psuedocounts
--soloCellFilter EmptyDrops_CR
--soloUMIfiltering MultiGeneUMI_CR
--soloUMIdedup 1MM_CR
--soloFeatures Gene GeneFull GeneFull_ExonOverIntron GeneFull_Ex50pAS Velocyto
--soloMultiMappers EM Uniform
--outSAMtype None
--soloBarcodeReadLength 0
```

scRecounter leverages `STARsolo`'s `--soloFeatures` parameter, enabling the generation of various count matrices tailored to specific analytical needs:

- **Gene**: Counts reads mapping entirely to exonic regions of annotated transcripts, capturing fully spliced transcripts. This procedure represents traditional gene expression analyses and was the default in earlier `CellRanger` versions.
- **GeneFull**: Counts reads overlapping entire gene locus, i.e. including exonic and intronic regions. This captures both unspliced (primary) and spliced transcripts and providing a more comprehensive view of gene activity.
- **GeneFull_ExonOverIntron**: Counts reads overlapping exonic and intronic regions, but assigns higher priority to exonic overlaps. This option helps to resolve reads that map to overlapping genes.
- **GeneFull_Ex50pAS**: Similar to the above option, but with more sophisticated priorities scheme, which prioritizes partial and antisense exonic overlap over intronice reads.
- **Velocyto**: Generates separate count matrices for spliced, unspliced, and ambiguous reads, following the rules from (La Manno et al., 2018). This enables RNA velocity analyses to infer dynamic cellular processes.

5.3. Reprocessing datasets in CZ CELLxGENE

We obtained accession codes corresponding to CZ CELLxGENE collections and re-processed the corresponding SRXs using the method outlined above. Constituent SRXs from each collection were then assembled into one `h5ad` object per collection, using *Unique GeneFullEx50pAS* counts.

We then mapped the observation metadata from each CZ CELLxGENE collection to our re-processed collections by matching cell barcodes. Specifically, within each collection, the cell barcodes follow the format `[ATCG]-[10x_well_id]`, and our `STARsolo` processed data follows the same format. Therefore, for each collection, we mapped the `10x_well_ids` in the CZI version of the collection to ours by assigning each identifier to the corresponding CZ CELLxGENE identifier that maximized the number of intersecting barcodes. By using this method, we were able to successfully map metadata from CZ CELLxGENE collections into scBaseCamp.

5.4. Evaluation of analytical factors in CZ CELLxGENE and scBaseCamp

In an effort to evaluate the contribution of suspension type to the data structure, we selected four human datasets shared between CZ CELLxGENE and scBaseCamp collections that contained both single-cell and single-nucleus data. We identified the matching cells in CZ CELLxGENE using the combination of `observation_joinid` and `cell_type` as unique cell identifiers. We excluded cells with fewer than 100 unique genes and genes observed in fewer than 40 cells. The resulting datasets were then sum-normalized and log-transformed. Principal component analysis (PCA) was performed using 50 principal components (PCs), with the top two PCs visualized using `scanpy`. Silhouette scores were computed using the `scib_metrics` package.

5.5. Dataset subsampling

For CZ CELLxGENE, each subsampled data contained samples from 20 projects, where all projects were solely 10X assays and labeled as 'primary_data' in CZI metadata. Each dataset was subsampled to 30% of the original before joining. Each samples was subsequently downsampled to 250, 000 cells as necessary. For scBaseCamp, we selected GeneFullEx50pAs counts from randomly chosen SRX accessions, and concatenated them together until the object reached 250, 000 cells. In both cases, half of the subsets were mouse, and half were human.

6. Acknowledgments

We especially thank Nianzhen Li for support and Jeremy Sullivan for managing infrastructure and compute resources. We also thank Brian Plosky, Joseph Caputo, Julia Kazaks, Arshia Nayebnazar. Finally, we thank the CZI CELLxGENE team for sharing SRA accession IDs to their datasets when available. S.K., P.D.H, and H.G. are Arc Core Investigators and acknowledge funding support from Arc Institute.

7. Author contributions

Y.R., A.D., and H.G. conceived the project. Y.R, H.G., A.D., P.D.H., S.K, and D.P.B. supervised the project. N.Y. conceived and developed `SRAgent`. C.C., R.I, and A.D. developed `scRecounter`. C.C. and J.P. performed data analysis, incorporation of CZ CELLxGENE projects, and comparisons between the two datasets. C.C. and C.R. visualized data. H.G., C.C., N.Y, and Y.R. wrote the document, incorporating comments from all authors.

8. Competing interests

D.P.B. acknowledges outside interest as a Google Advisor. H.G. acknowledges outside interest as a co-founder of Exai Bio, Vevo Therapeutics, and Therna Therapeutics, serves on the board of directors at Exai Bio, and is a scientific advisory board member for Verge Genomics and Deep Forest Biosciences. P.D.H. acknowledges outside interest as a co-founder of Terrain Biosciences, Stylus Medicine, and Spotlight Therapeutics, serves on the board of directors at Stylus Medicine, is a board observer at EvolutionaryScale and Terrain Biosciences, a scientific advisory board member at Arbor Biosciences and Veda Bio, and an advisor to NFDG, Varda Space, and Vial Health. All other authors declare no competing interests.

References

C. Bunne, Y. Roohani, Y. Rosen, A. Gupta, X. Zhang, M. Roed, T. Alexandrov, M. AlQuraishi, P. Brennan, D. B. Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, 16(1):43–49, Jan. 2019.

Y. Chen and J. Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nat. Biomed. Eng.*, pages 1–11, Dec. 2024.

H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

CZI Cell Science Program, S. Abdulla, B. Aevermann, P. Assis, S. Badajoz, S. M. Bell, E. Bezzi, B. Cakir, J. Chaffer, S. Chambers, J. M. Cherry, T. Chi, J. Chien, L. Dorman, P. Garcia-Nieto, N. Gloria, M. Hastie, D. Hegeman, J. Hilton, T. Huang, A. Infeld, A.-M. Istrate, I. Jelic, K. Katsuya, Y. J. Kim, K. Liang, M. Lin, M. Lombardo, B. Marshall, B. Martin, F. McDade, C. Megill, N. Patel, A. Predeus, B. Raymor, B. Robatmili, D. Rogers, E. Rutherford, D. Sadgat, A. Shin, C. Small, T. Smith, P. Sridharan, A. Tarashansky, N. Tavares, H. Thomas, A. Tolopko, M. Urisko, J. Yan, G. Yeretssian, J. Zamanian, A. Mani, J. Cool, and A. Carr. CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.*, 53(D1):D886–D900, Jan. 2025.

P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35(4):316–319, Apr. 2017.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. Howlett, O. Suchanek, K. Polanski, H. King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

G. Eraslan, E. Drokhlyansky, S. Anand, E. Fiskin, A. Subramanian, M. Slyper, J. Wang, N. Van Wittenberghe, J. M. Rouhana, J. Waldman, O. Ashenberg, M. Lek, D. Dionne, T. S. Win, M. S. Cuoco, O. Kuksenko, A. M. Tsankov, P. A. Branton, J. L. Marshall, A. Greka, G. Getz, A. V. Segrè, F. Aguet, O. Rozenblatt-Rosen, K. G. Ardlie, and A. Regev. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science*, 376(6594):eabl4290, May 2022.

A. C. Frazee, B. Langmead, and J. T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(1):449, Nov. 2011.

S. Frölich, M. van der Sande, T. Schäfers, and S. J. van Heeringen. Genomepy: Genes and genomes at your fingertips. *Bioinformatics*, 39(3), Mar. 2023.

M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

G. Heimberg, T. Kuo, D. J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, T. Biancalani, S. J. Turley, J. R. Rock, H. Corrada Bravo, J. Kaminker, J. A. Vander Heiden, and A. Regev. A cell atlas foundation model for scalable search of similar human cells. *Nature*, Nov. 2024.

W. Hou and Z. Ji. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods*, 21 (8):1462–1465, Aug. 2024.

B. Kaminow, D. Yunusov, and A. Dobin. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, May 2021.

A. Kazmi, D. Singh, S. Jatav, and S. Luthra. Beyond the hype: The complexity of automated cell type annotations with GPT-4. *bioRxiv*, page 2025.02.11.637659, Feb. 2025.

G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, Aug. 2018.

M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, D. Pe'er, and F. J. Theis. CellRank for directed single-cell fate mapping. *Nat. Methods*, 19(2):159–170, Feb. 2022.

R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.*, 39(Database issue):D19–21, Jan. 2011.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*, 40(1):121–130, 2022.

J. C. Melms, J. Biermann, H. Huang, Y. Wang, A. Nair, S. Tagore, I. Katsyv, A. F. Rendeiro, A. D. Amin, D. Schapiro, C. J. Frangieh, A. M. Luoma, A. Filliol, Y. Fang, H. Ravichandran, M. G. Clausi, G. A. Alba, M. Rogava, S. W. Chen, P. Ho, D. T. Montoro, A. E. Kornberg, A. S. Han, M. F. Bakhoum, N. Anandasabapathy, M. Suárez-Fariñas, S. F. Bakhoum, Y. Bram, A. Borczuk, X. V. Guo, J. H. Lefkowitch, C. Marboe, S. M. Lagana, A. Del Portillo, E. J. Tsai, E. Zorn, G. S. Markowitz, R. F. Schwabe, R. E. Schwartz, O. Elemento,

A. Saqi, H. Hibshoosh, J. Que, and B. Izar. A molecular single-cell lung atlas of lethal COVID-19. *Nature*, 595(7865):114–119, July 2021.

J. E. Rood, A. Maartens, A. Hupalowska, S. A. Teichmann, and A. Regev. Impact of the human cell atlas on medicine. *Nat. Med.*, 28(12):2486–2496, Dec. 2022.

Y. Rosen, Y. Roohani, A. Agarwal, L. Samotorčan, T. S. Consortium, S. R. Quake, and J. Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11, 2023.

O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, and S. A. Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, Oct. 2017.

C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616–624, 2023.

Y. Xiao, J. Liu, Y. Zheng, X. Xie, J. Hao, M. Li, R. Wang, F. Ni, Y. Li, J. Luo, S. Jiao, and J. Peng. CellAgent: An LLM-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*, page 2024.05.13.593861, May 2024.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv [cs.CL]*, Oct. 2022.

9. Supplementary figures

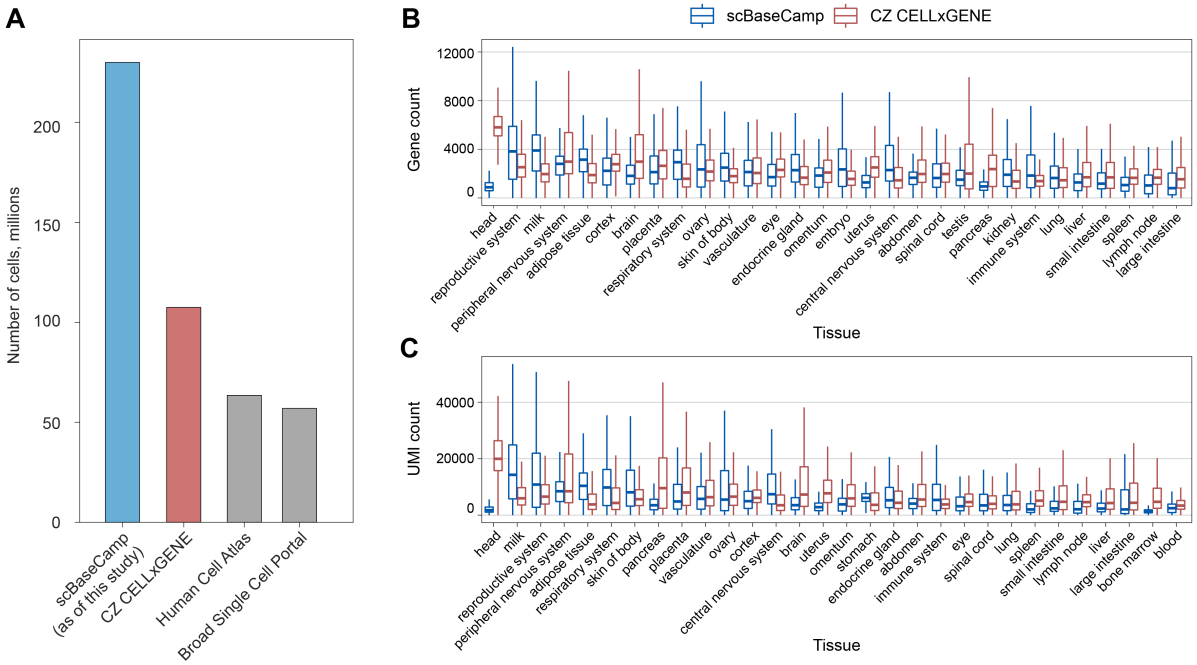

Figure S1 | **Size and count distribution comparison of scBaseCamp to other publicly accessible single-cell data repositories.** (**A**) Total number of unique cells in the largest publicly accessible single cell data repositories. (**B-C**) Boxplots depicting the distribution of gene and UMI counts per cell across the top 30 tissues for scBaseCamp (blue) and CZ CELLxGENE (red). The whiskers denote $1.5 \times IQR$. For scBaseCamp the data represents all mammals, and for CZ CELLxGENE human and mouse.

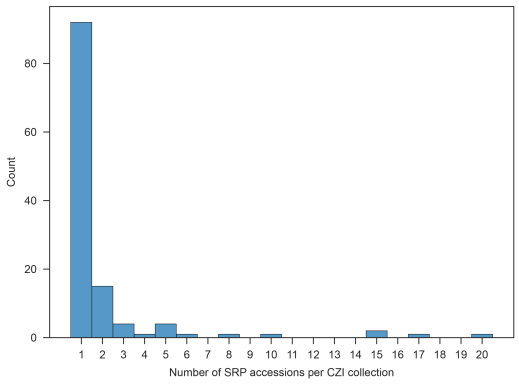

Figure S2 | **Histogram of the number of SRA studies (SRP) matching CZ CELLxGENE collection IDs.** Since in most cases, there was a single SRP associated with a given CZI collection, we opted to treat these two variables as comparable in our data visualization in Figure 3.

Dataset	Processing method	Cell prep	Publication
Integrated adult and foetal heart single-cell RNA sequencing	Demultiplexing, cell calling, alignment, and counts matrix generation were carried out using `Cell Ranger` version 6.1	Single cell, single nucleus (N=26, 972)	https://www.nature.com/articles/s44161-022-00183-w
A spatially resolved atlas of the human lung characterizes a gland-associated immune niche	Both types of libraries were mapped to an Ensembl 93-based reference (10X-provided GRCh38 reference, version 3.0.0). For nuclei samples, the reference was altered into a pre-mRNA reference as per 10X instructions	Single cell/single nucleus (N=142, 492)	https://doi.org/10.1038/s41588-022-01243-4
Cells of the adult human heart	Single-cell samples were mapped against the reference as it was provided. For single-nuclei samples, the reference for pre-mRNA was created using the 10X Genomics instructions.	Single cell/single nucleus (N=328, 595)	https://doi.org/10.1038/s41586-020-2797-4
Human skeletal muscle ageing atlas	Genomics skeletal muscle sequencing data were aligned and quantified using `Cell Ranger` (3.1.0) with GRCh38-3.0.0 human reference genome. Pre-mRNA version of reference genomes was used for alignment of nuclei datasets. `STARsolo` (STAR 2.7.3) '`-soloFeatures Gene GeneFull Velocyto`' was employed to separate spliced and unspliced counts, which were used to differentiate MF fragments.	Single cell, single nucleus (N=57, 581)	https://doi.org/10.1038/s43587-024-00613-3

Table S1 | **Overview of datasets, processing methods, and cell preparation techniques.** Shared human single cell data collections in scBaseCamp and CZ CELLxGENE containing both single-cell and single-nucleus sequencing data. The number of cells/nuclei in each dataset are shown in paranthesis in the "Cell prep" column. As tabulated here, for single cell nuclei, the authors had used pre-mRNA references for read mapping, and default `CellRanger mRNA` mapping for single cell RNA-seq datasets. This difference in the choice of analytical approach contributes to the separation of single-cell and single-nuclei studies.