The future of rapid and automated single-cell data analysis using reference mapping

Mohammad Lotfollahi[1,2*], Yuhan Hao[3,4*], Fabian J. Theis[1,2,5,+], Rahul Satija[3,4,+]

1 Institute of Computational Biology, Helmholtz Center Munich, Germany.
2 Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK.
3 Center for Genomics and Systems Biology, New York University, New York, NY, USA.
4 New York Genome Center, New York, NY, USA.
5 Department of Mathematics, Technical University of Munich, Germany.

*Equal contribution
[+]Correspondence: fabian.theis@helmholtz-muenchen.de, rsatija@nygenome.org

Summary

As the number of single-cell datasets continues to grow rapidly, workflows that map new data to well-curated reference atlases offer enormous promise for the biological community. In this perspective, we discuss key computational challenges and opportunities for single-cell reference mapping algorithms. We discuss how mapping algorithms will enable integration of diverse datasets across disease states, molecular modalities, genetic perturbations, and diverse species, and will eventually replace manual and laborious unsupervised clustering pipelines.

Introduction

Reference datasets (see **Glossary**) and mapping algorithms are transforming analytical workflows for single-cell sequencing datasets. This mirrors similar trends that resulted from the construction of the first human genome map[1]. reference-based analysis shifts data interpretation from an unsupervised to a supervised domain, enabling information accumulated from multiple prior experiments to help interpret new data. When analyzing genome sequence data, the existence of a reference map ensures that each new experiment does not require a re-assembly of the genome from the data itself, dramatically simplifying analytical workflows and reducing

requirements on read length and data quality. Similarly, for single-cell analysis, efficient reference-mapping workflows can replace manual, laborious, and subjective unsupervised clustering and labeling tasks with automated mapping and annotation.

The widespread use of reference mapping for genome sequence analysis also provides a roadmap for similar potential applications in single-cell sequencing. Human genome references enable the mapping of genomic data from millions of individuals, placing data into a standardized space that allows comparative analysis **(Fig1. a)**, and identifying genetic variants. Genome references can also serve as a scaffold for diverse data types and modalities, including epigenomic profiling technologies such as chromatin immunoprecipitation sequencing (ChIP-seq) and high□throughput chromosome conformation capture technique (Hi-C). Moreover, exploring differences across multi-species genomic references is a powerful evolutionary and comparative genomics technique. There are analogies for each of these applications for single-cell sequencing, underscoring the potential for reference-mapping algorithms to process multiple data types beyond just single-cell RNA-sequencing (scRNA-seq) such as spatial molecular profiles. However, genomic analysis has also revealed challenges for reference-mapping approaches that users of single-cell tools are beginning to encounter. These include technical challenges, including the ability to map genomes with structural variations to healthy references, and data-driven challenges, such as the necessity to update references based on newly available data continually.

For these reasons, there has been substantial interest in developing computational methods to assemble single-cell reference datasets and map new datasets onto them. Popular techniques encompass a diverse set of approaches, including statistical approaches based on dimensionality reduction, machine-learning based discretized classification techniques, and deep artificial neural networks[2–6]. These advances enabled construction of single-cell atlases for human organs like lung[7], tonsil[8], and brain[9], and enabled researchers to study diseases[10] and development[11] by integrating data from across multiple studies. These tools are being increasingly paired with collections of reference datasets that are already being assembled by multiple collaborative efforts, including the Human Cell Atlas and Human Biomolecular Atlas Project. While the most common analytical task is automated cell annotation (or ‘label transfer’), reference-mapping workflows can also transfer continuous data sources, including developmental trajectories and additional cellular modalities. Computational development continues in this area, with new methods improving robustness, accuracy, and scalability.

In this perspective, we start by reviewing exciting possibilities and pressing challenges for the field of single-cell reference mapping. We also explore the scope of broad applications that can

be encompassed by reference-mapping workflows, and the diverse set of query datasets that can be mapped. These include perturbed cells (reflecting disease states but also biochemical or drug-induced perturbations), cross-species and evolutionary analyses, and spatially-resolved datasets that contain information about cellular position and morphology. Lastly, we explore inherent challenges in constructing authoritative references in a dynamic field and argue for open source atlasing that can rapidly and reliably update references as new data becomes available.

Reference mapping workflows

The concept of mapping newly generated biological 'query' data to curated references is a powerful idea that predates single-cell genomics. For example, genomic reference assembly algorithms are computationally intensive, require high-quality long-read data, and typically involve extensive manual curation. In contrast, read-mapping algorithms are highly efficient, compatible with short-read data, and fully automated. The ability to map new query datasets to established references instead of performing *de novo* assembly and annotation for each dataset has a transformative impact, and substantially improved analytical data workflows. Moreover, genomic references can serve as a single scaffold to consistently interpret and compare data from multiple samples. This enables in-depth analysis of genetic variation, but also integrative analysis across a wide variety of functional genomics technologies, including ChIP-seq, Hi-C, and RNA-seq.

The recent growth of single-cell sequencing technologies has led to the emergence of high-quality single-cell tissue, organ or even whole-species 'atlases'[12–16]. As with genome assembly, the construction of these atlases is often laborious, computationally intensive, and requires manual curation and annotation. The reference data often contain multiple single-cell datasets across one or multiple modalities and metadata **(Fig. 1b)**, typically characterizing up to thousands of cell types and where they are found[12,17] Analytical strategies and challenges for single-cell data integration and reference construction have been reviewed before[18] and compared[19] . Briefly, single-cell reference datasets often consist of at least two components. The first component is a data transformation which projects data measurements into a low-dimensional space. This transformation can include multiple linear or non-linear steps and often aims to facilitate data integration by placing cells in similar biological states in similar positions, even if they originate from different datasets. Ideally, such a transformation should be able to integrate multiple data views across different non-overlapping features (e.g., gene, peaks) while

correcting for technical variations originating from different sequencing protocols and environments, also known as “batch effect”[19,20]. There has been substantial progress in developing methods in the overarching theme of data integration, solving either batch correction in one modality or multimodal integration, which we later discuss in detail. The second component of a reference is the manual assignment of metadata, typically a set of annotations provided for each cell in the dataset, which can optionally conform to established cell ontologies and exhibit a hierarchical structure.

Different analytical techniques for mapping new samples, or ‘query’ datasets, onto the reference also tend to follow a common strategy. First, the same data transformation that is learned when assembling the reference dataset, is applied to the query. This projects the query cells into a reference-defined space and effectively integrates the two datasets. Neighbor relationships can then be utilized to transfer discrete or continuous information onto query cells based on the most similar reference data points. While the accuracy of these techniques depends on the quality of the reference-defined transformation and annotations, supervised mapping offers substantial advantages compared to unsupervised analysis. These advantages include higher-quality annotations in particular with noisy and sparse query data, improved detection of rare or molecularly subtle cell states, compatibility with fully automated workflows that do not require manual steps or parameter tuning, and substantial improvements in speed and memory requirement[7]. Most importantly, reference mapping offers the ability to interpret query datasets without the need to recluster and reannotate **(Fig. 1c)**.

Reference mapping methods can be categorized according to the type of transformation learned to assemble the reference. The first group of approaches learns a data transformation using statistical approaches and their variants. For example, Seurat uses reference low-dimensional representation (such as single-modality Principal Component Analysis (PCA) or multimodal supervised PCA) projection and anchor-based integration to map query cells onto the reference[6,21] Similarly, Symphony[5] learns a low-dimensional transformation (e.g., principal component analysis) in which cells are softly assigned to clusters representing different cell states to build the reference model. Alternatively, scArches[3] exploits probabilistic neural networks[22–24] to learn a non-linear transformation of the data while correcting for technical effect between datasets. Once the initial reference transformation is learned, it is applied to query datasets to map them to the reference subject to the criteria outlined in the previous section.

Regardless of the specific computational method chosen to perform reference mapping, these workflows have the potential to support a wide array of applications. While mapping datasets from healthy individuals to healthy reference atlases is possible, interpreting diseased samples

using healthy atlases is highly desirable but accompanied by distinct challenges. Similarly, while most mapping approaches are tailored for scRNA-seq datasets, a key computational challenge is to map data types originating from diverse data types, including alternative modalities and spatial profiles. Finally, reference atlases may enable robust interpretation datasets from different species to explore evolutionary changes, or to samples originating from genetic perturbations to reconstruct molecular networks. We discuss these potential applications, and their associated challenges and opportunities, below **(Fig. 1c)**.

Identification of disease states by contextualizing disease within a healthy reference

Understanding disease pathology requires the identification and characterization of affected cell types. Single-cell datasets provide an unprecedented opportunity to study disease mechanisms by comparing disease cells to matched control samples to characterize cellular changes caused by the disease[25]. Users leverage metrics to assess compositional cell-type changes in combination with statistical methods to obtain genes and pathways altered by the perturbation to assess overall molecular changes[26]. Reference mapping methods have successfully been applied to map cells from COVID-19 patients from Bronchoalveolar lavage fluid (BALF) or Human peripheral blood mononuclear cells (PBMCs) [3,6,27] to healthy atlases based on cells from same or multiple tissues and detected disease-associated cell-types. Further, tumor-derived cells from patients with renal cell carcinoma (RCC) have been successfully mapped to healthy kidney atlas, revealing the separation of tumor-compartment cells from reference cells while immune/stromal compartments integrated into the reference[5].

If diseased samples display variance that is not evident in the healthy reference, robust detection of compositional or cell-state alterations in disease can be challenging[28]. Thus, mapping a disease dataset to a sufficiently large and diverse healthy reference atlas enables rapid identification of disease states. Successful mapping of disease queries should meet the following criteria: (1) conservation of the heterogeneity of healthy cell states in the reference, (2) integration of identical cell types in reference and query and (3) preservation of previously uncharacterized cell types and states emerging in disease datasets that are not present in the reference. This reflects the well-known trade-off of maintaining biological variation but reducing batch variation from standard data integration[19] in the context of mapping new samples. Despite advances in data integration, the automatic identification of disease states remains challenging. The robust recovery of defined disease populations needs an uncertainty metric to discriminate new cell states from existing references (**Fig. 1c**). For example, a simple K-nearest

neighbors (Knn) classifier trained with a distance metric could identify previously uncharacterized cell types and states using the uncertainty of transfer labels from reference to the query [3, 29]. In addition, HLCA authors identify disease-specific populations during SARSCoV-2 pneumonia and idiopathic pulmonary fibrosis. Similarly, Symphony exploits the query cells' cell-based or cluster base Mahalanobis distance to reference cells to detect unknown cell types or disease states. Finally, a recent approach[30] combines scArches with hierarchical classifiers[31] to learn and extend hierarchical representations of cell types. The hierarchical representation enables the identification of specific populations (e.g., disease) when adding the query population not fitting the existing hierarchy. These examples show that there is ample need for new methods or improvements toward more robust detection of disease states. Unsupervised disease-state identification using reference mapping is related to out-of-distribution (OOD) detection[28], an unsolved challenge in the machine-learning community. While methods leveraging deep generative models (DGMs) can build upon model likelihood to detect the OOD samples (i.e., novel states in the query), however they may assign a higher likelihood to OOD samples compared to in-distribution (reference data)[32]. As these and other methods improve, we note that disease-specific references can also be constructed. While the construction of multiple healthy and diseased references may be a laborious effort, this approach will provide contextual flexibility to map a wide variety of query datasets.

Once multiple samples have been mapped into a shared space, a suite of statistical methods have been developed to prioritize (i.e., rank) cell types according to the magnitude of responses to perturbations. Responses can be quantified either based on a change in the proportion of cell types in perturbed datasets, or alternately, based on the magnitude of gene expression changes within a cell type. For compositional changes, the methods MASC[33] and scCODA[34] identify compositional changes based on discrete cell-type clusters, while Milo[35] and MELD[36] use cell-type independent, continuous approaches to quantify compositional changes at the level of neighborhoods or single cells. These approaches have been used to analyze the compositional changes that occur upon disease perturbation, such as COVID-19 or liver cirrhosis, or aging. Complementing these tools are methods that focus on quantifying subpopulation-specific state transitions by comparing gene expression profiles from cell groups detected in multiple conditions. For example, Robinson and colleges introduce robust statistical tools for multi-sample comparison[37], while Augur [38], trains classifiers to identify the most responsive populations to perturbations in single-cell data.

Overall, these approaches demonstrate how scRNA-seq combined with reference mapping becomes a powerful tool to identify how complex populations respond to perturbation. Looking

ahead, we see potential in addressing *interpretability* of multi-sample comparisons. Disease perturbations in particular are unlikely to affect only a single cell subpopulation, and instead will represent complex responses that are both shared across cell types, and unique to particular cell states. Methods that can help to decompose these differential sources of variation and prioritize particular cell populations for downstream analysis, will be highly beneficial to the broader community.

Population-scale reference mapping

Reference mapping approaches also have important potential to analyze and explore large-scale variation across populations of samples. In the same way that large genetic databases of human variation, such as gnomAD, catalog and compare hundreds of thousands of samples after mapping into a consistent reference framework, single-cell reference mapping tools enable similar types of meta-analysis. An example is a meta-analysis[39] of 22 separate scRNA-seq studies of COVID-19 blood samples. These studies encompassed a total of more than 3 million cells varying in age, sex, and ethnicity, disease state, and disease severity. In order to facilitate robust comparisons, all samples were mapped to a single reference, facilitating the automated harmonization of cell type labels and metadata. Standardization facilitates the performance of large-scale meta-analysis, and in particular, the ability to identify reproducible COVID-induced changes in cell type composition across hundreds of different donors. As single-cell studies routinely present data not just from a large number of cells, but also from a large number of individuals, reference mapping is likely to play an essential role in interpreting these datasets. In addition to facilitating standardization of cell labels, reference mapping may also help to infer and classify disease state[40] and severity in query samples. For example, a class of machine learning algorithms called multi-instance learning (MIL) enables learning such mappings. The MIL algorithm allows learning a transformation for each sample (e.g., patient) and classifying it as a whole without knowing individual labels (disease affected or healthy). In addition, MIL methods can identify cell populations responsible for disease severity. Such applications will enable the automation of disease severity classification, facilitating diagnostics while helping potential (personalized) treatments by identifying disease-associated cell types for each patient and disease **(Fig. 2)**.

Construction of cellular perturbation atlases

Single-cell healthy atlases are increasingly available via consortia such as the human cell atlas[12]. However, human cell atlas -generated data is focused on healthy homeostatic conditions, and large-scale perturbation experiments here referred to as "perturbation atlases"[41] aimed toward drug discovery and regenerative medicine, represent a new frontier. The concept of mapping query datasets to perturbation atlases has been widely explored for bulk studies. In particular, large scale bulk molecular profiling technologies have been utilized to generate maps of molecular responses to thousands of perturbations, including genetic perturbations, small molecules, cytokines, and drugs. Databases such as Connectivity Map (CMap), LINCS 1000, and ChemPert assemble these perturbations into data, and can be used to interpret the broad sets of transcriptional signatures[42].

This conceptual framework has clear promise for perturbations measured at single-cell resolutions. Recently, the development of barcoding technologies [43,44] enables high-throughput characterization of the effect of small molecules [43], or CRISPR-Cas9/13-based single-gene or combinatorial genes perturbations [45,46]. These approaches are being increasingly applied to organoid systems and iPSC-derived models[47] , in-vivo models[48], and can even extend to genome-wide perturbation experiments[49].

As these approaches continue to develop, reference mapping can help connect these datasets to single-cell profiles from healthy and diseased samples, drawing connections between experimentally driven perturbations and naturally observed disease states. However, the explorative space of perturbations and their combinations is enormous and experimentally infeasible to test **(Fig. 1c).** This hinders the construction of comprehensive perturbation atlases similar to healthy counterparts. Data integration algorithms can reduce the sample sparsity in this scenario by allowing the integration of multiple sparse perturbation experiments into a more thorough atlas. While integration can potentially increase the discovery power, there is a trade-off between data integration and preserving biological variability, requiring careful metrics and assessment[19] An alternative approach involves machine-learning algorithms to 'impute' missing perturbations, using initial reference datasets to infer the effect of previously unseen perturbations on cellular behavior [41].

Initial approaches based on dynamical models[50,51] have been proposed to predict proliferation measurements or gene expression effects across many perturbations. However, dynamical approaches require prior knowledge about the regulatory system for model design, and often rely on time-resolved measurements, which are hard to obtain at the single-cell level. This results in parameter identifiability and fitting challenges. In contrast, linear approaches are

easier to fit but have limited generalization to unseen perturbations or modeling complex cell-type specific behaviors[52].

Deep learning methods have been developed to address these challenges to predict cellular behaviors. Variational autoencoders (VAE)[53] have been the main tool for learning low-dimensional latent representation from single-cell data. An example is scGen[52], a VAE combined with latent space vector arithmetics to predict single-cell response to disease and chemical perturbations across cell types and species. Following on from this work, the compositional perturbation autoencoder (CPA)[54] has been proposed to extend existing methods to predict combinatorial responses to drugs or genetic perturbations. CPA learns a cell representation as the composition of a basal state combined with learned representation for perturbations and covariates (e.g., cell type, patient, species). Finally, recent efforts have extended pre-existing methods to forecast the effects of previously unprofiled chemical perturbations or genetic deletions[55,56]. All of these methods propose a clear vision to predict molecular response to unseen perturbations, either individually or in combination, and reveal an exciting path forward to augmenting perturbation atlases. Going forward, combination of large-scale perturbation experiments with deep-learning based imputation and integration across multiple studies, potentially together with experimental augmentation via active learning approaches [57] will lead to the assembly of systematic perturbation atlases.

Single-cell data mapping across molecular modalities

While the techniques above focus on the mapping of scRNA-seq query datasets onto scRNA-seq reference atlases, the field of single cell genomics is rapidly transitioning to routinely profile alternative molecular modalities. In particular, there is substantial interest in profiling genomic features, such as chromatin accessibility[58,59], DNA-protein interaction maps[60,61], or chromosome contact interactions[62]. Creating a new reference dataset to enable mapping query datasets from each new modality would represent a crippling burden on the research community. Therefore, there is interest in exploring the potential for cross-modality mapping. One example would be to map scATAC-seq query datasets onto scRNA-seq defined reference atlases. If successful, these approaches would extend the widespread benefits of reference-mapping framework[63,64] to a diverse set of modalities and technologies extending beyond scRNA-seq. The fundamental challenge in cross-modality mapping is a lack of correspondence features that are measured in different datasets. For example, scATAC-seq datasets measure chromatin accessibility at

genomically defined regions, while scRNA-seq measures quantitative levels of gene expression. The lack of overlapping features between reference and query datasets invalidates the use of scRNA-seq reference mapping tools, and necessitates the development of new methods.

Feature Conversion

The first set of cross-modality mapping methods attempted to solve the issue of feature correspondence by converting one type of measurement into another (**Fig. 3a**). For example, the Cicero algorithm quantified the total accessibility of ATAC-seq peaks located within each gene-body and 2 kilobase upstream regions[65]. Noting that genes located in regions of open chromatin tend to be actively expressed, Cicero referred to these accessibility quantifications as 'gene activity' scores, which were a proxy for transcriptional output. Importantly, this feature conversion transforms the features measured from ATAC-seq into the same set of features measured by scRNA-seq, representing a first step towards integration.

After feature conversion, existing integration and mapping algorithms can be used to perform cross-modality alignment and mapping. For example, Seurat v3[21] utilizes canonical correlation analysis to identify a conserved biological subspace between a gene activity score matrix estimated from scATAC-seq, and scRNA-seq measurements. This conserved subspace enables the identification of cell-to-cell correspondences across datasets, termed anchors. These anchors enable the automated annotation of scATAC-seq profiles based on established transcriptomic reference maps of the mammalian brain. Similarly, the LIGER algorithm[66] utilizes non-negative matrix factorization to infer a set of linear latent factors that represent shared biological signals across modalities. While LIGER successfully mapped chromatin accessibility data, DNA methylation measurements tended to be inversely correlated with gene expression, which would also allow for the mapping of methylation query datasets. Chiefly, both Seurat v3 and LIGER methods utilize cross-modality mapping to explore relationships between a cell's regulatory landscape and its transcriptional output, leading to the inference of cell-type specific regulatory networks.

Building upon these advances, MultiMAP[67] uses a manifold learning method for the dimensionality reduction and integration of multiple datasets after feature conversion, generalizing the UMAP distance metrics to learn a single latent manifold where data from multiple modalities is evenly distributed. Another method, GLUE[68] , implements a variational autoencoder for adversarial alignment across modalities, guided by a prior-knowledge based 'guidance graph' which links individual genomic peaks to their associated genes. Both methods

MultiMAP and GLUE demonstrate the potential for 'tri-omics' integration, successfully integrating scATAC-seq, scRNA-seq, and DNA methylation profiles from different cells. The diversity of methods demonstrates the potential for feature conversion-based approaches to generate meaningful mappings. However, all these approaches rely on rigid and simplistic biological assumptions that are inherent to the conversion process. When these features correlation assumptions fail to hold true, the conversion method could transform uncertainties into errors. For example, while open chromatin is often associated with active transcription, this may not always be the case, particularly in developing systems where a 'lag' between dynamic changes in chromatin accessibility and transcriptional output have been well-documented[69–71].

Bridging with multi-omic datasets

An alternative approach to cross-modality mapping exploits the recent development of a suite of 'multi-omic' single cell technologies, where more than one molecular modality is simultaneously measured in single cells[72]. For example, CITE-seq[73] utilizes barcoded antibodies to jointly profile RNA and protein levels in single-cells, while the SHARE-seq[74], SNARE-seq[75], and 10x multiome technologies enable paired single-cell measurements of chromatin accessibility profiles and gene expression levels. While powerful, multi-omic profiling typically has a higher financial cost than the combination of two separate modalities. Beyond this, increased technical noise, and decreases in throughput[76] limits its widespread application. However, in cases where multi-omic profiles are available, a suite of computational methods can leverage these datasets to assist in cross-modality mapping (**Fig. 3a**).

For example, Seurat v5[71] accomplishes cross-modality mapping by utilizing a multi-omic dataset as a 'bridge'. Since the bridge dataset includes paired measurements of the modalities that are individually represented in the reference and query datasets, all reference and query cells can be accurately represented as weighted combinations of the bridge cells. This procedure effectively transforms datasets from different modalities into a common feature space, but without making any underlying biological assumptions. Similarly, the StabMap algorithm[77] constructs a mosaic data topology that connects reference, bridge, and query cells - and then performs cross-modality mapping by identifying shortest paths across this topology. Both bridge integration and StabMap demonstrate how a multi-omic bridge can substantially improve the accuracy of cross-modality integration compared to previous approaches based on feature conversion. Moreover, they demonstrate how utilizing diverse bridge datasets, including 10X

multiome (scRNA-seq+scATAC-seq)[78], Paired-Tag (scCUT&Tag+scRNA-seq)[79], and CITE-seq[73] bridge datasets can enable the mapping of a wide variety of query datasets to pre-existing scRNA-seq references.

In addition, a suite of deep learning tools also leverages multi-omic datasets to integrate datasets measuring different molecular modalities. For example, the BABEL algorithm[80] utilizes multi-omic data to learn a 'translation' that maps one data modality to another, based on an interoperable neural network model. Based on this model, BABEL can generate 'predicted' values for one modality based on measured values from another, and demonstrates the ability to translate across chromatin, RNA, and protein modalities. Recently a body of published work such as MultiVI[81], Cobolt[82] and CLUE[83] and also a preprint[84] all leverage multimodal variational autoencoders (MVAEs). MVAEs represent a recent advance in deep learning where individual neural networks initially perform individual modeling of separate datasets, but they are then subsequently projected into a uniform biological subspace. Since this subspace encompasses both unimodal and multi-modal cells, this approach effectively enables cross-modality mapping. Each technique highlights powerful features of the MVAE framework, including the tailoring of modality-specific noise models (MultiVI), the application of a hierarchical generative model (Cobolt), the application of cross-encoders to learn cross-modality representations (CLUE), and simultaneous correction of batch effects alongside cross-modality integration (Multigrate).

As the suite of cross-modality integration tools continues to mature, we anticipate a greater emphasis on computational tools to analyze and interpret their outputs. We also expect a more systematic comparison to evaluate these tools, applying diverse metrics that concentrate on different performance metrics. In particular, cross-modality integration enables a flexible experimental design where different modalities are collected in different experiments, but can then be analyzed together. This serves as an alternative to true multi-omic (i.e. simultaneous measurement) technologies, but may also allow for increased per-modality data quality and higher cellular throughput. For example, SCENIC+[85], learns gene regulatory networks from paired measurements of chromatin accessibility and gene expression, exploiting co-variation across both modalities in order to infer key transcriptional regulators and their target genes. Similarly, MultiVelo[86] integrates chromatin accessibility and gene expression to estimate chromatin switch and gene splicing states. This multimodal inference also allows researchers to study the dynamics between transcription factor expression and its binding sites accessibility. While originally developed for multi-omic measurements[85,86], these and similar approaches can also be applied to the results of cross-modality integration, which would greatly broaden the

scope of datasets that could be utilized to help identify relationships across molecular modalities.

Cross species mapping

Single-cell sequencing has now been able to scale molecular characterization of cells to entire organisms at the "whole-animal" scale, encompassing worm[87], fly[88], zebrafish[89], frog[90], mouse[14], and even human fetus[16]. These datasets not only enable a detailed characterization of cellular heterogeneity *within* organisms, but they also allow for a comparison of cell types and states *across* organisms. Comparative genomics represents an invaluable tool for the annotation and identification of human genomic elements, including both ultra-conserved[91] and rapidly evolving regions[92]. We anticipate single-cell analysis following a similar roadmap, and for cross-species analysis to substantially improve our evolutionary understanding of shared and unique cell states across species.

Even in the face of broad transcriptional differences, evolutionarily shared molecular patterns can facilitate the identification of homologous cell types across species. One such example is the discovery of a subset of evolutionarily conserved markers in pancreatic islets separated from humans and mice[93]. Despite representing only a fraction of the transcriptome, shared markers were sufficient to accurately align cell types cross-species via canonical correlation analysis[94]. Similar approaches have been used repeatedly to explore evolutionary cell type conservation in the mammalian brain. For example, cross-species mapping tools can be used to identify broad and surprising conservation across excitatory, inhibitory, and non-neuronal cell types between the human and mouse cortex[95]. The initial alignment step enabled a detailed exploration of cross-species differences in cell-type abundance, localization, as well as the identification of substantial differentially expressed gene modules. The results are comparable to a comprehensive atlas of the motor cortex in humans, mice, and marmosets[96].

Cross-species alignment also enables the prediction of cellular properties across species. One example is the characterization of van Economo neurons (vENs)[96], which exhibit a distinct morphology and maybe associated with neuropsychiatric conditions, but whose functional properties are poorly understood. By identifying a rare group of these cells in human scRNA-seq data and performing cross-species mapping, a multimodal cell census and atlas of the mammalian primary motor cortex identified strong homology to a particular subset of extra telencephalic (ET) excitatory neurons that project to subcerebral targets. These findings ultimately support the hypothesis that vENs project to subcortical targets, and point to particular partners with whom vENs may form circuits.

Cross-species mapping approaches can also help to identify distinct differences across species. One example is the alignment of neuronal scRNA-seq samples from turtles, lizards, and mammalian datasets[97]. Strikingly, it revealed clear (one-to-one) homology between broad GABAergic interneurons subsets across all amniotes, suggesting a deep conservation and shared evolutionary origin of these cell types. In contrast, glutamatergic neurons were detected in all species but lacked clear molecular homology, suggesting significant evolutionary diversification. Despite broad conservation, distinctions in the primate inhibitory interneuron repertoire were found when compared to other mammals, including an abundant striatal interneuron subgroup that exhibited no molecular homology with mice[98].

Usually, cross-species mapping depends on a reference from one species. The creation of a universal multi-species reference is a promising approach to enhance identification of genes that are functionally related and co-expressed across species. It may even allow to uncover potentially divergent functions along the evolution. For example, the SATURN algorithm[99] which incorporates the protein language model ESM2[100] integrates Aqueous Humor Outflow cell atlas[101,102] scRNA-seq data from five species (humans, cynomolgus macaques, rhesus macaques, mice, and pigs) into a shared low-dimensional embedding space based on gene expression and the protein structural similarity. One notable finding with SATURN was that human *Myoc* gene function is divergent from its orthologous genes in other species. Such universal references enable us to understand the relationship between gene sequences and functions across a vast array of species.

We anticipate continued improvement in cross-species mapping methods which remain challenging[19], particularly given lack of clear definition of homologous features across species, and the broader challenge of identifying biological homology amidst widespread evolutionary changes. Nonetheless, we expect that cross-species analyses at single-cell resolution will continue to inform our understanding of the function, uniqueness, and evolutionary origins of human cell types. In particular, cross-species comparisons of developmental processes[103], offer a powerful opportunity to compare developmental stages based on molecular profiles.

Moreover, cross-species alignment of alternative modalities, especially chromatin features measured at single-cell resolution through techniques like single-cell ATAC-seq and single-cell CUT&Tag, will set a new direction in genomic research. Utilizing genome liftover translating genomic coordinates from one species assembly to another, these chromatin features from diverse species can be harmonized into a unified genomic space, which enables subsequent cross-species alignment. The alignment of functional genomic modalities will represent a unique

approach to annotate and characterize regulatory elements that drive cellular state and diversification across species.

Path toward machine learning based open source atlasing

Above, we have outlined how reference mapping enables integration of perturbation, multimodal, patient cohorts and even cross-species data sets. A key question for the community is how references will be made, released, and iteratively updated. Moreover, there are multiple consortia such as the Human Cell Atlas[104], Human Biomolecular Atlas Program[13], LifeTime Initiative, and Chan Zuckerberg Initiative[105], all of whom aim to generate substantial datasets and release them openly to the community.

version control in software development, such as Git and GitHub, is a fitting analogy to this. Assigning version numbers to reference models (e.g., Human lung cell atlas v1.0.0) allows for clear tracking of changes and updates. When new data is generated, a "pull request" can be made, suggesting an updated version of the reference model (e.g., V1.1.0) along with the corresponding updated data. This mechanism facilitates collaborative review and integration of the updates by different working groups and researchers, ensuring accuracy and relevance. Moreover, with the proliferation of large-scale machine learning models[106–110] and open-source repositories like Hugging Face, the practical implementation of this version-controlled approach becomes feasible. Researchers can leverage and contribute to shared machine learning model repositories, promoting collaboration and democratization of the reference models.

While it is tempting to turn to the human genome project to explore community-based solutions, there are key differences between genomic references and single-cell references that drive unique challenges. Most importantly, the iterative releases of the human genome project cleanly built upon each other, with each subsequent reference adding new data primarily to fill in existing gaps. This has led to a stable, well-curated, and authoritative reference genome for the community. By contrast, single-cell references updates often refine, change, and add to previous versions. This reflects the highly dynamic nature of cells, their ability to take on a wide variety of states, and our incomplete understanding of their heterogeneity.

Addressing this challenge will require overcoming both logistical and computational challenges that do not yet have clear solutions. For example, multiple groups may initially release overlapping reference datasets for the same human tissues. Different biological communities will also explore different approaches for how and when to update references, for example based on a set timeline, or in response to the generation of new landmark datasets. In contrast to the standardization of the Human Genome Project, this will likely lead to a wide variety of

distinct references for the same set of human tissues because they were generated by different groups. The scientific community benefits from having a variety of options for testing and iterative refinement, and over-enforced standardization can limit the process of discovery. Yet, the current human cell atlas reference seems like a puzzle with many missing pieces. Instead of parallel efforts to profile similar organs or tissues, the idea of an open-source atlas can help guide experimental design toward identifying cells or tissues that haven't been profiled yet and should be prioritized. It is essential for both the computational and experimental communities to work together as part of smaller networks that focus on different organs to lay out such a plan.

It is tempting to enforce standardization and adoption of a unique community-accepted reference atlas for each human tissue to alleviate this concern. In principle, benchmarking approaches could be used to compare different reference atlases and select a “winner.” In practice, however, enforcing strict standardization is likely to be detrimental given the current early stage of the field. No single reference atlas is likely to be “correct”, and multiple groups who approach the same problem will likely produce references that have distinct or complementary strengths. The biological community benefits from a variety of options for testing and iterative refinement, and over-enforced standardization can limit the process of discovery. A middle-ground approach, where multiple groups can construct independent reference atlases, but elected institutions or leaders oversee eventual pooling of datasets and resources, may be an attractive approach. For example, multiple groups have released reference atlases of the Mouse Brain, but the Allen Brain Atlas and NIH BICCN have led efforts to bring these groups and datasets together to establish a more thorough and authoritative cell ontology. As the comprehensive scale of this reference atlas grows, groups that create new references should be encouraged (though not required) to contribute their data into this framework. However, deciding on a “winner” or merging datasets into one reference is not always ideal. The Human Cell Atlas and NIH LungMAP initiative for example have each constructed scRNA-seq atlases of the human lung. Both initiatives bring together a wide diversity of labs and expertise for data generation, integration, and annotation. Though their resulting atlases substantially overlap, the differences between them represent cutting-edge discoveries of molecular lung cell states that would be diminished by choosing only a single “winner.” Individual labs can map their datasets to both atlases, compare and contrast results, and provide feedback that will yield a more comprehensive and standardized reference atlas over time.

Advanced high-throughput sequencing technologies enable detailed investigation of previously uncharacterized tissues and species in millions of cells. However, the absence of well-established references for these novel biological entities complicates the analysis of these

large-scale datasets. Constructing a reference atlas from millions of cells becomes critical, especially when conventional methods fail due to high computational time and memory demands. This challenge can be aptly termed 'data compression'. There are three primary strategies for this purpose: 1) Aggregating homogenous cells to form meta cells; 2) Sketching representative cells from the entire dataset; and 3) Segmenting the entire dataset into manageable chunks. These strategies aim to preserve the inherent cellular heterogeneity while demanding minimal computational resources. Notably, these three types of approaches are not mutually exclusive but can be complementary for different tasks. Innovations like the single-cell large-language foundation models further broaden the horizons of reference atlas creation, diversifying its utility in downstream analyses.

Reference mapping approaches depend on the quality of reference building algorithms, leading to inherent limitations. For instance, scArches relies on conditional generative models and deep representation learning. These algorithms necessitate extensive training datasets encompassing various experimental protocols to model complex batch effects effectively. Without sufficient data, they may struggle to map query datasets, especially if the query data comes from different technologies or species not present in reference[3]. Addressing this challenge requires the development of more robust neural network architectures capable of generalizing well under low data conditions. On the other hand, non-deep learning algorithms (e.g. Seurat and Harmony) for reference mapping may not be data-hungry. Still, they may encounter scalability issues with tens of millions of datasets. Overcoming this hurdle involves down-sampling [111] or pseudo-bulking strategies [99,112], potentially introducing biases into the models. Finally, existing reference mapping algorithms primarily operate in a latent space instead of a corrected feature matrix. To enable downstream analysis using the corrected feature matrix calls, more robust reference-building algorithms that operate directly on the input space must be developed[113].

In addition to the computational hurdles, the effectiveness of transferring knowledge from the reference to the query is impacted by the quality of reference metadata, particularly cell type annotations. This is important in a scenario where one organ has multiple references, each annotated by different groups with distinct sets of annotations. Diversity in annotations makes choosing the most suitable atlas a challenge. Hence, a more systematic approach is crucial to establish a consensus annotation across similar references. A reference cell ontology [99] or frameworks similar to a "reference cell tree" [114] can aid in harmonizing and integrating diverse annotation sources into a cohesive set (tree). This integration mitigates the ad hoc nomenclature of cell types and states. Machine learning methods [30,115] can be employed to

construct and continually update these hierarchical references, assigning a tree to each organ. This principled and unified approach allows practitioners to systematically name and annotate cell types and states.

The here described examples and use cases highlight the broad potential for reference-mapping algorithms to transform the basic analytical pipelines by which users analyze, interpret, and explore single-cell data. Going forward, we envision that reference mapping will, slowly but surely, begin to replace unsupervised clustering and manual annotation workflows. In doing so, single-cell analysis will transition from an expert-centric and tedious pipeline to a rapid, accessible, and accurate procedure for beginners and experts alike.

Acknowledgements

We acknowledge members of the Satija and Theis labs for thoughtful discussion. M.L and acknowledges financial support from the Joachim Herz Stiftung via Add-on Fellowships for Interdisciplinary Life Science. FJT acknowledges support by the BMBF (01IS18036A and 01IS18036B), by the European Union's Horizon 2020 research and innovation program (grant 874656), by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI (ZT-I-PF-5-01), and sparse2big (ZT-I-0007). RS acknowledges support the Chan Zuckerberg Initiative (EOSS5-0000000381 to R.S., HCA-A-1704-01895), and the National Institutes of Health (OT2OD033760, 5RM1HG011014, 5R01HD096770).

Declaration of Interests

M.L. consults Santa Ana Bio, is a part-time employee at Relation Therapeutics, and owns interests in Relation Therapeutics. F.J.T. consults for Immunai Inc., CytoReason Ltd, Cellarity, Inc and Omniscope Ltd, and owns interests in Dermagnostix GmbH and Cellarity Inc. In the past 3 years, R.S. has worked as a consultant for Bristol-Myers Squibb, Regeneron and Kallyope and served as a scientific advisory board member for ImmunAI, Resolve Biosciences, Nanostring and the NYC Pandemic Response Lab. R.S. and Y.H. are co-founders and equity holders of Neptune Bio. As of August 1, 2023, Y.H. is an employee of Neptune Bio.

References

1. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
3. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
4. Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C. & Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat. Commun.* **11**, 1–13 (2020).
5. Kang, J. B. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
6. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
7. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
8. Massoni-Badosa, R. *et al.* An atlas of cells in the human tonsil. *Immunity* **57**, 379–399.e18 (2024).
9. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
10. Salcher, S. *et al.* High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503–1520.e8 (2022).
11. Herring, C. A. *et al.* Human prefrontal cortex gene regulatory dynamics from gestation to adulthood at single-cell resolution. *Cell* **185**, 4428–4447.e28 (2022).
12. Regev, A. *et al.* Science Forum: The Human Cell Atlas. *Elife* **6**, e27041 (2017).
13. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
14. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
15. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

16. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
17. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).
18. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
19. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
20. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* vol. 177 1888–1902.e21 Preprint at https://doi.org/10.1016/j.cell.2019.05.031 (2019).
22. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* vol. 15 1053–1058 Preprint at https://doi.org/10.1038/s41592-018-0229-2 (2018).
23. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
24. Lotfollahi, M., Naghipourfar, M., Theis, F. J. & Wolf, F. A. Conditional out-of-distribution generation for unpaired data using transfer VAE. *Bioinformatics* **36**, i610–i617 (2020).
25. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
26. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
27. Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell Biol.* **25**, 337–350 (2023).
28. Dann, E. *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* **55**, 1998–2008 (2023).
29. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *bioRxiv* 2022.03.10.483747 (2022) doi:10.1101/2022.03.10.483747.
30. Michielsen, L. *et al.* Single-cell reference mapping to construct and extend cell-type hierarchies. *NAR*

Genom Bioinform **5**, lqad070 (2023).

31. Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun.* **12**, 2799 (2021).
32. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv [cs.NE]* (2016).
33. Fonseka, C. Y. *et al.* Mixed-effects association of single cells identifies an expanded effector CD4+ T cell subset in rheumatoid arthritis. *Sci. Transl. Med.* **10**, (2018).
34. Buettner, M., Ostner, J., Mueller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 1–10 (2021).
35. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
36. Burkhardt, D. B. *et al.* Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.* **39**, 619–629 (2021).
37. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
38. Skinnider, M. A. *et al.* Cell type prioritization in single-cell data. *Nat. Biotechnol.* **39**, 30–34 (2021).
39. Tian, Y. *et al.* Single-cell immunology of SARS-CoV-2 infection. *Nat. Biotechnol.* **40**, 30–41 (2022).
40. De Donno, C. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *bioRxiv* 2022.11.28.517803 (2022) doi:10.1101/2022.11.28.517803.
41. Ji, Y., Lotfollahi, M., Alexander Wolf, F. & Theis, F. J. Machine learning for perturbational single-cell omics. *cels* **12**, 522–537 (2021).
42. Chen, H. *et al.* Drug target prediction through deep learning functional representation of gene signatures. *Nat. Commun.* **15**, 1853 (2024).
43. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
44. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* 1–8 (2021).
45. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell

phenotypes. *Science* **365**, 786–793 (2019).

46. Wessels, H.-H. *et al.* Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. *bioRxiv* 2022.02.02.478894 (2022) doi:10.1101/2022.02.02.478894.
47. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* (2022) doi:10.1038/s41586-022-05279-8.
48. Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, (2020).
49. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).
50. Fröhlich, F. *et al.* Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model. *Cell Systems* **7**, 567–579.e6 (12/2018).
51. Yuan, B. *et al.* CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy. *Cell Systems* vol. 12 128–140.e4 Preprint at https://doi.org/10.1016/j.cels.2020.11.013 (2021).
52. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
53. P Kingma, D. & Welling, M. *Auto-Encoding Variational Bayes*. (2014).
54. Lotfollahi, M. *et al.* Compositional perturbation autoencoder for single-cell response modeling. *bioRxiv* 2021.04.14.439903 (2021).
55. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
56. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01905-6.
57. Budd, S., Robinson, E. C. & Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* **71**, 102062 (2021).
58. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
59. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation.

Nature **523**, 486–490 (2015).

60. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
61. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1–10 (2019).
62. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
63. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
64. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
65. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).
66. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).
67. Jain, M. S. *et al.* MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol.* **22**, 346 (2021).
68. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01284-4.
69. Reed, K. S. M. *et al.* Temporal analysis suggests a reciprocal relationship between 3D chromatin structure and transcription. *Cell Rep.* **41**, 111567 (2022).
70. Wagh, K. *et al.* Dynamic switching of transcriptional regulators between two distinct low-mobility chromatin states. *Sci Adv* **9**, eade1122 (2023).
71. Hao, Y. *et al.* Dictionary learning for integrative, multimodal, and scalable single-cell analysis. Preprint at https://doi.org/10.1101/2022.02.24.481684.
72. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17**, 11–14 (2020).
73. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
74. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**,

1103–1116.e20 (2020).

75. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
76. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17**, 11–14 (2020).
77. Ghazanfar, S., Guibentif, C. & Marioni, J. C. Stabilized mosaic single-cell data integration using unshared features. *Nat. Biotechnol.* **42**, 284–292 (2024).
78. Luecken, M. *et al.* A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* **1**, (2021).
79. Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
80. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
81. Ashuach, T., Gabitto, M. I., Jordan, M. I. & Yosef, N. MultiVI: deep generative model for the integration of multi-modal data. Preprint at https://doi.org/10.1101/2021.08.20.457057.
82. Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol.* **22**, 351 (2021).
83. Tu, X., Cao, Z.-J., Chenrui, X., Mostafavi, S. & Gao, G. Cross-Linked Unified Embedding for cross-modality representation learning. *Adv. Neural Inf. Process. Syst.* **35**, 15942–15955 (2022).
84. Lotfollahi, M., Litinetskaya, A. & Theis, F. J. Multigrate: single-cell multi-omic data integration. *bioRxiv* 2022.03.16.484643 (2022) doi:10.1101/2022.03.16.484643.
85. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
86. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat. Biotechnol.* 1–12 (2022).
87. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

88. Li, H. *et al.* Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly. Preprint at https://doi.org/10.1101/2021.07.04.451050.
89. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
90. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
91. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
92. Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, e168 (2006).
93. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346–360.e4 (2016).
94. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
95. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
96. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
97. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
98. Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262–269 (2020).
99. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01716-9.
100. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
101. Rosen, Y. *et al.* Towards Universal Cell Embeddings: Integrating Single-cell RNA-seq Datasets across Species with SATURN. *bioRxiv* 2023.02.03.526939 (2023) doi:10.1101/2023.02.03.526939.
102. van Zyl, T. *et al.* Cell atlas of aqueous humor outflow pathways in eyes of humans and four model

species provides insight into glaucoma pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10339–10349 (2020).

103. Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* (2022) doi:10.1038/s41586-022-05547-7.

104. Regev, A. *et al.* The Human Cell Atlas White Paper. (2018) doi:10.48550/arXiv.1810.05192.

105. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

106. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

107. Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).

108. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

109. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* (2024) doi:10.1038/s41592-024-02201-0.

110. Birk, S. *et al.* Large-scale characterization of cell niches in spatial atlases using bio-inspired graph learning. *bioRxiv* 2024.02.21.581428 (2024) doi:10.1101/2024.02.21.581428.

111. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Syst* **8**, 483–493.e7 (2019).

112. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

113. Hausmann, F. *et al.* DISCERN: deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection. *Genome Biol.* **24**, 212 (2023).

114. Domcke, S. & Shendure, J. A reference cell tree will serve science better than a reference cell atlas. *Cell* **186**, 1103–1114 (2023).

115. Wang, S. *et al.* Leveraging the Cell Ontology to classify unseen cell types. *Nat. Commun.* **12**, 5556 (2021).

Figure Legends

Fig1. Automated analysis of single-cell data using reference mapping. (a) Mapping RNA or DNA short reads to the reference genome using reference mappers as an alternative for computationally expensive de novo reference assembly. **(b)** Assembly of a single-cell reference - similar to a reference genome - enables automated analysis of newly generated query datasets by mapping them into the reference using a reference mapping algorithm. **(c)** Applications of single-cell reference mapping are automated cell-type annotation of query data (first row), analyzing single-cell perturbations such as disease states or missing perturbations to be imputed in the query data (second row), imputing continuous information for the query data including spatial location for scRNAseq using a spatial atlas or chromatin accessibility for query data using a multimodal reference including scRNAseq and scATACseq (third row).

Fig2. Reference mapping at population scale. (a) The availability of Cohort-level single-cell references enables the assembly of resources composed of many samples (or patients) to learn heterogeneity across populations and cells (b). (c) Query samples are mapped to both cell and sample-level representations. (d) After mapping the new samples leveraging cell embedding and supervised analysis, the disease phenotype for query samples can be classified (e.g. type of the tumor type). (e) Sample-level representation can infer sample-sample similarity maps between reference and query directly linked to cell-level representation. The circle represents a group of donors in query with different cellular compositions, as reflected in the reference embedding.

Fig3. Single-cell data reference mapping across molecular modalities. **(a)** Two frameworks to build cross modality feature correspondence. Feature conversion: transforming one type of measurement into another. For example, ATAC-seq peaks within gene bodies can be converted into gene activity scores, the same set of features measured by scRNA-seq. Multi-omics bridge: leveraging multi-omic datasets to establish connections between different modalities. For example, bridging ATAC-seq peaks and RNA-seq genes using datasets that measure both ATAC peaks and gene expression. **(b)** Expanding RNA reference to other query modalities using single-cell multi-omics datasets. By using single-cell multi-omics technologies as molecular bridges, RNA references can be expanded to include additional modalities such as DNA methylation (DNA met), ATAC peaks, surface proteins, CUT&Tag (cleavage under targets and tagmentation) , and Spatial data. snmC2T-seq, single-nucleus methylCytosine, Chromatin accessibility and Transcriptome sequencing; SNARE-seq, single-nucleus chromatin accessibility and mRNA expression sequencing; ASAP-seq, ATAC with select antigen profiling by sequencing; CITE-seq, cellular indexing of transcriptomes and epitopes by sequencing; Paired-Tag, parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing; CUT&Tag-Pro, single-cell cleavage under targets and tagmentation with cell surface proteins ; Spatial-CUT&Tag, spatial cleavage under targets and tagmentation.

Text Box:

Glossary

Multimodal reference: A reference atlas that is built using more than one modality (for example. RNA and ATAC).

Multimodal omics: Technologies capable of capturing multiple data types from the same sample

Supervised vs unsupervised learning: in this context, data integration while leveraging cell-type labels in the reference and query dataset (supervised) compared to the scenario in which the method has no access to these labels (unsupervised).

Principal Component Analysis(PCA): a linear dimensionality reduction technique used to reduce the dimensionality of datasets.

Low-dimensional representation: in this context, the reduced dimensional space of a dataset after applying a transformation (e.g. PCA).

Hierarchical classifier: a machine learning model that organizes and categorizes data into multiple levels or layers of nested classes, allowing for a structured and granular classification approach.

Deep generative models: The class of artificial intelligence algorithms that use deep neural networks to learn and generate new data samples, exhibiting the ability to create novel and realistic outputs in diverse domains such as images, text, or audio.

Out of distribution detection: a machine learning task focused on identifying instances or data points that differ significantly from the patterns learned during training, helping models recognize and flag inputs lying outside the known distribution, thus enhancing robustness and reliability in real-world application.

Multi-instance learning: A machine learning paradigm where the training data is organized into bags, each containing multiple instances (examples). The model is tasked with making predictions at the bag level, and while the labels are provided for the bags, the specific instance-level labels within each bag are uncertain or unknown. This approach is often used in scenarios where only partial information about the labels is available, making it suitable for tasks like image classification, drug discovery, and anomaly detection.

Dynamical models: Mathematical representations that capture the time-dependent behavior and evolution of a system. These models describe how variables change over time based on a

set of differential equations or iterative rules, enabling the simulation and prediction of system dynamics in various fields such as physics, biology, economics, and engineering.

Variational autoencoders (VAEs): types of generative model in machine learning that combine elements of autoencoders and variational inference. VAEs aim to learn probabilistic mapping between the input data and a latent space, allowing for the generation of new data points. The encoder network maps input data to a probability distribution in the latent space, and the decoder network generates data from samples drawn from this distribution. VAEs are commonly used for tasks like generating novel data samples, data compression, and unsupervised learning.

Manifold learning: a set of techniques in machine learning and data analysis focused on capturing the underlying structure, or manifold, of high-dimensional data in a lower-dimensional space. The goal is to represent complex data in a way that preserves its underlying geometric relationships.

Latent space: in the context of machine learning, a lower-dimensional space in which the representations of data are learned and encoded. It is a crucial concept in different model architectures including both autoencoders and generative models (variational autoencoders and generative adversarial networks).

Reference atlas: extensively annotated and curated single cell data that show a comprehensive view of cellular heterogeneity of specific tissues or samples serving as a detailed map of cellular and molecular characteristics.

Label Transfer: Projecting labels from a well annotated reference atlas onto a newly generated query dataset.

Reference mapping for genome sequence: Aligning DNA/RNA sequencing short reads to a reference genome to get genomic identify of reads.

Single-cell reference mapping: Aligning genetic profiles, such as the transcriptome of individuals cells, to a reference atlas in order to obtain annotations at the single cell level.

Single-cell RNA-sequencing: A sequencing technique for profiling the gene expression profiles of individual cells.

Single-cell ATAC-sequencing: A sequencing technique used to profile the open chromatin regions within individual cells.

Cross-modality mapping: A specialized type of single-cell reference mapping in which the query and reference belong to two different modalities, such as mapping scATAC profiles onto a scRNA reference.

Canonical correlation analysis: A statistical method used to understand the relationships between two datasets, capturing shared variance and identifying correlated patterns.

Non-negative matrix factorization: An algorithm decomposing high-dimensional data into a lower-dimensional representation ensuring that all components of the decomposed matrices are non-negative.

Adversarial alignment: An algorithm that harmonizes datasets from different sources or platforms by reducing batch effects and other confounding variations.

Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq): a multimodal sequencing technique that enables the simultaneous measurement of protein and RNA in single cells.

SHARE-seq, SNARE-seq, and 10x multiple: Three different sequencing techniques that enable the simultaneous measurement of open chromatin regions and RNA in single cells.

Multi-omics bridge: A single-cell multi-omics dataset used in cross-modality mapping as a bridge between query and reference.

Paired-Tag: Parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing (Paired-Tag) is a sequencing technique that simultaneously profiles of different histone modifications and transcriptome in single cells.

Multimodal variational autoencoders: A type of variational autoencoder used to integrate multiple modalities of data and learn one single joint latent representation.

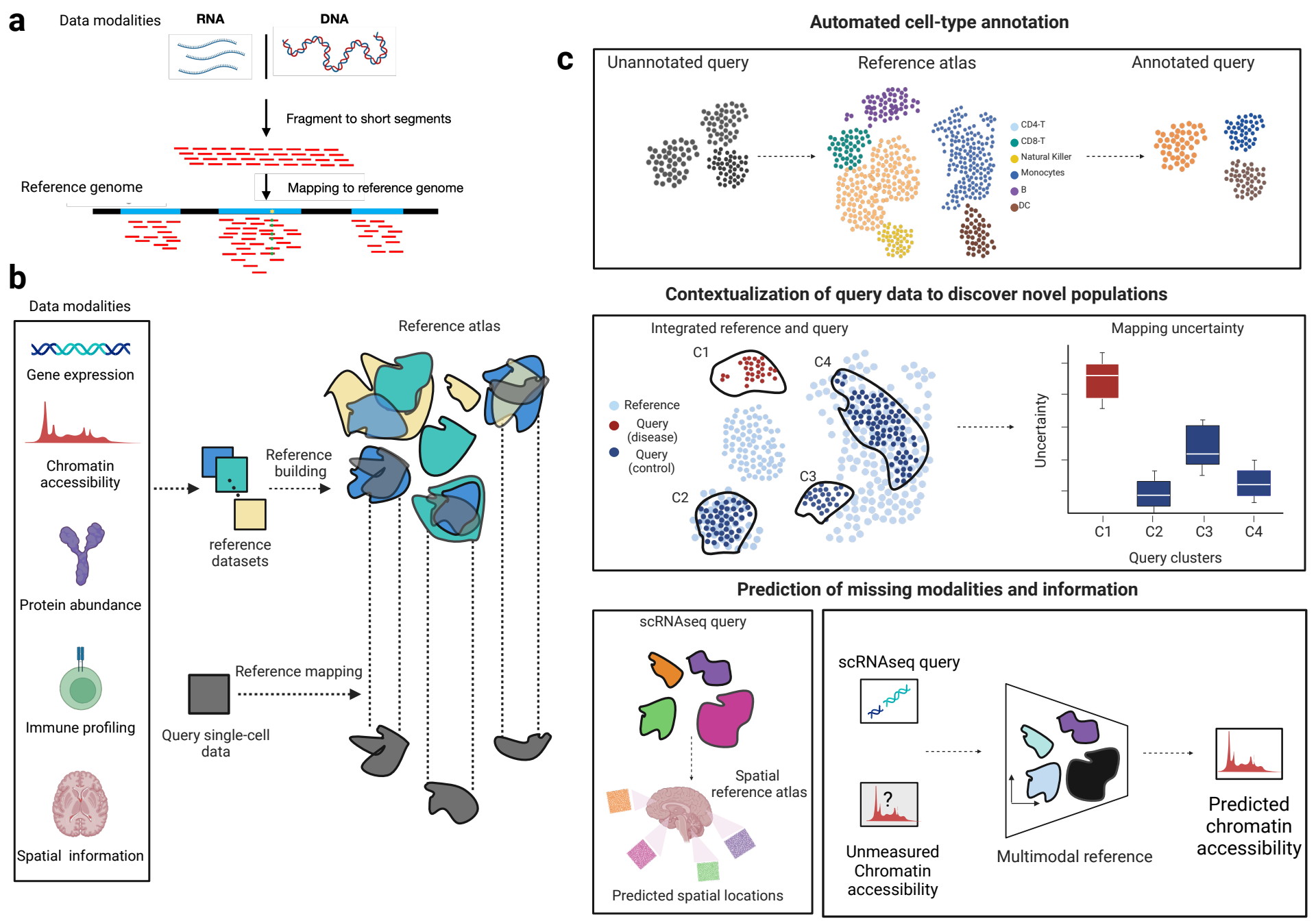
Representative cells sketching: An algorithm to sample a subset of cells from the entire data. The sampled cells are expected to effectively preserve cellular heterogeneity and gene expression covariance from the full dataset.

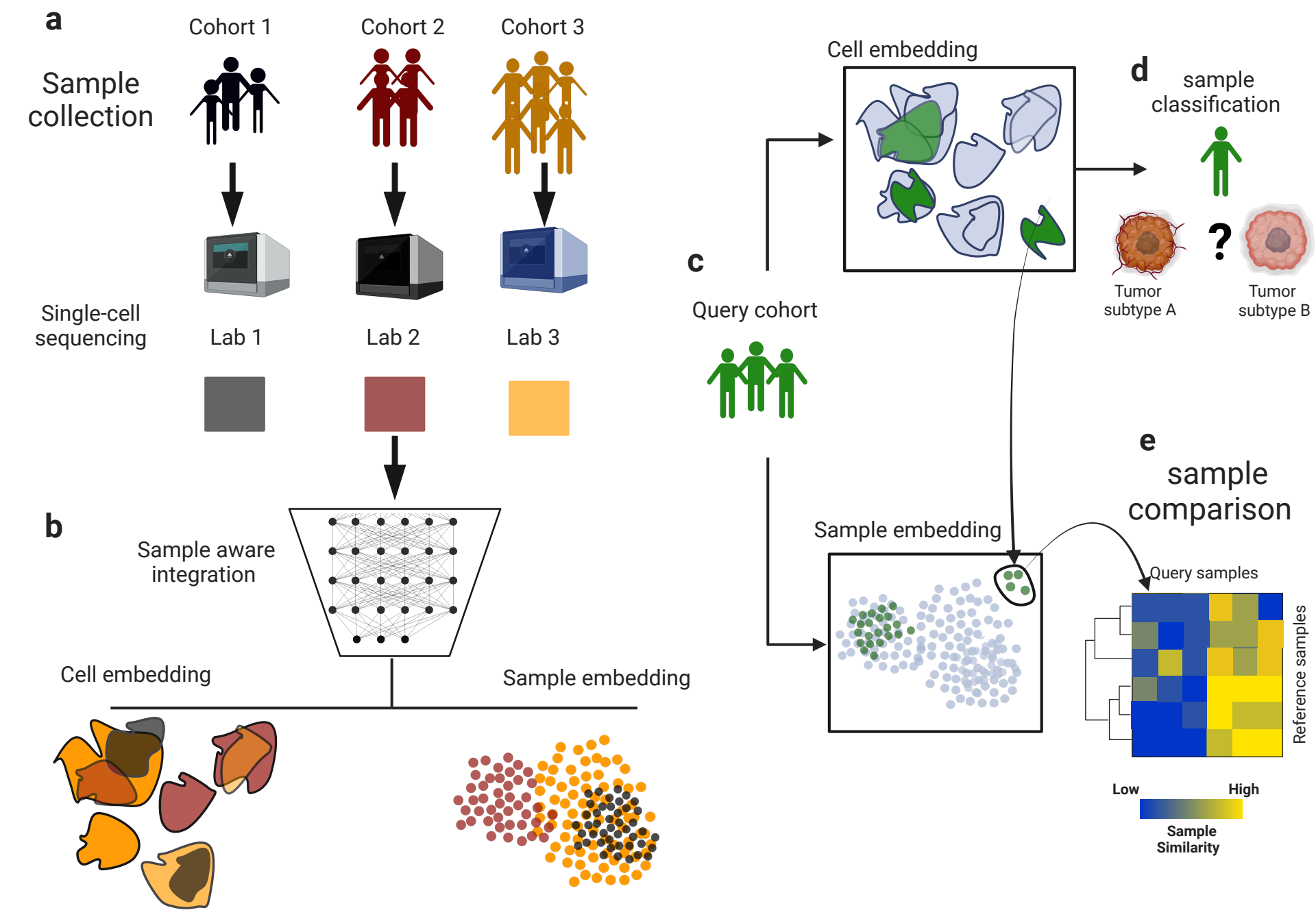
Metacell: A computational concept for grouping homogeneous cells based on the similarity of their genetic molecular to represent distinct cell types or states.

Cross species mapping: A type of single-cell reference mapping in which the query and reference are from two different species, used to understand conserved and diversified cell types and gene programs in terms of evolutionary relationships.

Protein language model Evolutionary Scale Modeling 2 (ESM2): a transformer-based language model designed to predict protein structure and function based on amino acid sequences.

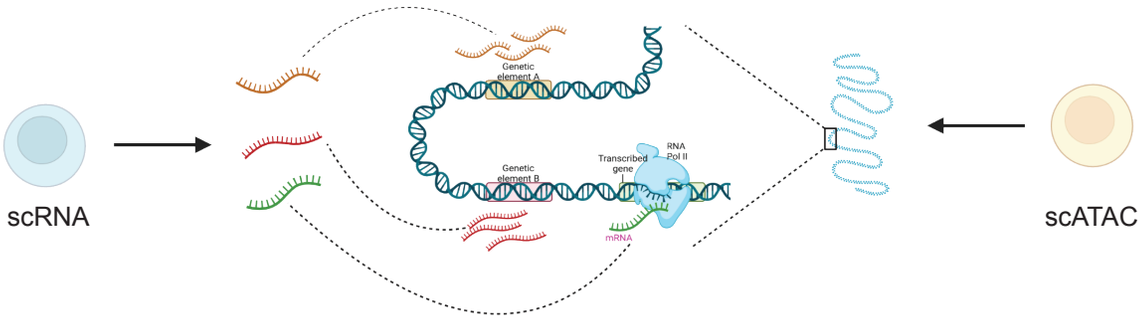
Genome liftover: Converting genomic coordinate information from one genome assembly to another, enabling the comparison of genomic data across different versions or different species of reference genomes.


a
Data modalities
RNA
DNA
Fragment to short segments
Reference genome
Mapping to reference genome
b
Data modalities
Gene expression
Chromatin accessibility
Protein abundance
Immune profiling
Spatial information
Reference atlas
Reference building
reference datasets
Reference mapping
Query single-cell data
c
Automated cell-type annotation
Unannotated query
Reference atlas
Annotated query
CD4-T
CD8-T
Natural Killer
Monocytes
B
DC
Contextualization of query data to discover novel populations
Integrated reference and query
Reference
Query (disease)
Query (control)
C1
C2
C3
C4
Mapping uncertainty
Uncertainty
Query clusters
Prediction of missing modalities and information
scRNAseq query
Spatial reference atlas
Predicted spatial locations
scRNAseq query
Unmeasured Chromatin accessibility
Multimodal reference
Predicted chromatin accessibility


a
Cohort 1
Cohort 2
Cohort 3
Sample collection
Single-cell sequencing
Lab 1
Lab 2
Lab 3
b
Sample aware integration
Cell embedding
Sample embedding
c
Query cohort
Cell embedding
d
sample classification
Tumor subtype A
Tumor subtype B
Sample embedding
e
sample comparison
Query samples
Reference samples
Low
High
Sample Similarity

a

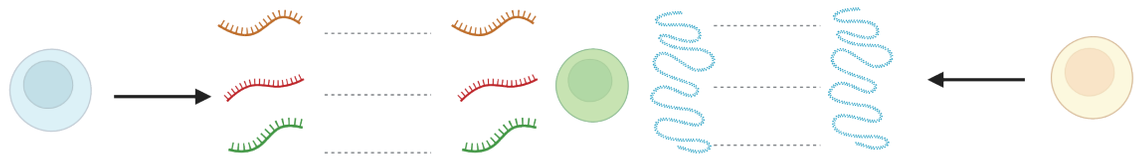
Cross-modality integration (Feature Conversion)



scRNA

scATAC

Cross-modality integration (Multi-omic bridge)



scRNA

multi-omics
RNA + ATAC

scATAC

b