nature methods

Perspective

https://doi.org/10.1038/s41592-024-02390-8

Small data methods in omics: the power of one

Received: 19 August 2023

Accepted: 24 July 2024

Published online: 22 August 2024

Check for updates

Kevin G. Johnston[1,2,8], Steven F. Grieco[2,3,8], Qing Nie[1,4] ✉, Fabian J. Theis[5,6,7] ✉ & Xiangmin Xu[2,3] ✉

Over the last decade, biology has begun utilizing 'big data' approaches, resulting in large, comprehensive atlases in modalities ranging from transcriptomics to neural connectomics. However, these approaches must be complemented and integrated with 'small data' approaches to efficiently utilize data from individual labs. Integration of smaller datasets with major reference atlases is critical to provide context to individual experiments, and approaches toward integration of large and small data have been a major focus in many fields in recent years. Here we discuss progress in integration of small data with consortium-sized atlases across multiple modalities, and its potential applications. We then examine promising future directions for utilizing the power of small data to maximize the information garnered from small-scale experiments. We envision that, in the near future, international consortia comprising many laboratories will work together to collaboratively build reference atlases and foundation models using small data methods.

Why 'small data' methods?

Large single-cell 'omics atlases are now almost routinely generated by consortia such as the BRAIN Initiative Cell Census Network (BICCN) and the Human Cell Atlas, and serve as references for smaller-scale studies performed by individual labs[1–3]. Catalyzed by jumps in single-cell RNA-sequencing technology, these 'big data' approaches have been instrumental in shaping the current renaissance in science by elucidating the cellular diversity of the body, region by region[4]. The colloquial 'big data' when referring to transcriptomics generally consists of many often multimodal high-dimensional data points, where data structures can be complex, and are frequently generated in a high-throughput manner in terms of volume and speed[5]. However, in recent years experts in data science and machine learning have announced the arrival of 'small data' methods[6–9], which focus on using small data efficiently by effectively contextualizing small datasets within large-scale reference atlases, and which are forecasted herein to be a major driver of discovery going forward.

Small data methods have the potential to substantially increase the insights that can be drawn from studies of any size, greatly improving cost efficiency in terms of time and money spent[10]. Methods such as 'transfer learning', which use machine learning models, often deep neural networks that are trained to generalize learned 'rules' across datasets, allow scientists to learn from reference atlases[11,12]. Smaller datasets can then be used to further train the model, and to ultimately update the reference data in an iterative process. This approach opens up possibilities for collaboration among hundreds or thousands of labs to build large, accurate reference atlases, which can be used for comparing analyses across brain regions, brain disorders, drug conditions and even species[13,14].

Problems integrating 'small data' with 'big data'

While consortia-produced, single-cell atlases are large, they are increasingly dwarfed in comparison with the combined transcriptomic assay output of individual labs[15]. This disparity will only grow as

[1]Department of Mathematics, University of California, Irvine, Irvine, CA, USA. [2]Department of Anatomy and Neurobiology, School of Medicine, University of California, Irvine, Irvine, CA, USA. [3]Center for Neural Circuit Mapping, University of California, Irvine, Irvine, CA, USA. [4]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, USA. [5]Helmholtz Center Munich—German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. [6]School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. [7]Department of Mathematics, Technical University of Munich, Munich, Germany. [8]These authors contributed equally: Kevin G. Johnston, Steven F. Grieco. ✉e-mail: qnie@uci.edu; fabian.theis@helmholtz-munich.de; xiangmix@hs.uci.edu

transcriptomic and epigenetic assays become increasingly standardized aspects of cellular interrogation across biology.

Individual transcriptomic assays typically interrogate individual brain regions with limited sample sizes. In this sense, single-cell atlases can be viewed as analogous to 'reference genomes', and individual assays as 'reads'. Within this analogy, integration of assays with atlases is effectively 'variant calling'. In practice, this 'variant calling' amounts to discriminatory analyses identifying biological differences between lab-produced assays and single-cell atlases, and this can range from cell-type proportion alterations to perturbation analysis in gene regulatory networks (GRNs). However, there are several issues impeding integrative analysis.

Batch effects

First, single-cell transcriptomic assays are typically produced from relatively few mice, with similar genetic backgrounds, and are prepared consistently across samples[16]. Since query data and reference data are often generated in different labs and by different scientists using different protocols, batch effects can dominate the quality of the joint embedding, leading to spurious results[17]. If a large-scale reference dataset has been generated in a single lab, it may not have had sufficient exposure to this type of variation to allow a robust mapping of novel data, in contrast to a cross-lab integrated atlas. Sophisticated data integration methods are often used to overcome these batch effects, and typically treat perturbations that affect most cells as batch effects. However, this can mask real biological differences between datasets[18].

Computational challenges

Second, integrative analysis co-embedding individual assays with atlases is still computationally difficult for many labs but is important when attempting to dissect individual cell types and states[19]. This is particularly true when analyzing rare cell types, as identification of higher-resolution distinct cell states contextualized within cell-type atlases frequently requires co-embedding approaches, requiring access to processed atlas data, and potentially the embedding model[20]. While computational algorithms for minimizing time and resource requirements exist, these algorithms are relatively new, their use is not widespread, and their use must be considered before creation of the atlas itself[20,21].

Data standardization

Third, while there are standardized requirements for deposition of raw (fasta and metadata level) single-cell RNA (scRNA) experimental data, no such requirement currently exists for processed data. While many authors do deposit processed data, and there have been substantial efforts toward developing databases for cross-conditional comparison of processed scRNA datasets, performing anything but the most basic comparisons (for example, cell-type querying, differential expression) is still time consuming and inefficient. Integrative pseudotime and GRN analysis is uncommonly used with more than two or three external datasets.

Coordination

Fourth, while consortia have made major efforts to enable accessibility to their cell atlases, this information flow travels only from consortia outward. That is, single-cell atlases are fundamentally non-collaborative beyond the consortia itself. Many single-cell datasets have been produced outside consortia, and single-cell atlases at present do not incorporate this wealth of information. Of course, consortia have valid reasons for not incorporating this information, ranging from resource allocation to data quality, and it is neither incentivized nor incumbent on them to include data from outside sources. The fact remains that at present, there is no concerted fieldwide collaborative effort to create integrative cell atlases in health and disease.

A variety of recent approaches have been attempted to ameliorate some of these issues. New integration methods explicitly model both technical and biological variation[22,23], and recent benchmarks indicate that top-of-the-line integration algorithms can accurately account for technical variation while retaining biological variation[18].

Numerous attempts have been made to construct single-cell datasets for public usage and comparison, across multiple fields of biology. Notable examples include PanglaoDB[24], EMBL's Single Cell Expression Atlas[25], the Broad Institute's Single Cell Portal[26] and CZI sciences CZ CELL×GENE Discover[27]. Smaller databases have been developed to investigate specific diseases such as cancer[28] and Alzheimer's disease[29]. One recent model-based example of note is scGPT, a pretrained transformed model trained utilizing 33 million cells[30]. Critically, the authors demonstrate that this model can be optimized to facilitate downstream applications such as cell-type annotation, integration, perturbation response prediction and GRN inference, notably outperforming competitors such as scBERT[31], Harmony[32] and GEARS[33] at these critical tasks.

However, each of these approaches has only partially fulfilled the promise of enabling integrative analysis of complex transcriptomic features (for example, lineage tracing analysis, GRN inference) across the entire field of biology. We posit that the primary reason for this is the lack of incentive and technical ability for individual researchers to integrate and share their own data on these platforms, requiring database authors to reanalyze, process and integrate individual datasets before incorporating them into databases that still typically only allow for simple cell-type and differential expression analysis via online portals. Enabling and incentivizing the individual researchers to integrate and share their own data in updateable, collaborative transcriptomic atlases is critical to development of the field of single-cell omics.

Constructing a robust and accessible updateable integrated atlas

Creation of an updateable integrable transcriptomic atlas requires a clear delineation of the technical difficulties attendant thereto (Fig. 1a). Such a project requires dedicated organization to standardize updates and model training. The primary issues are (1) standardization of RNA preprocessing, (2) choice of updateable atlas approach, (3) mechanisms for integration and validation and (4) computational resource allocation.

While log-normalized processing of RNA data is still common and generally effective[34], count matrix normalization has been an active area of research for almost two decades, and a variety of additional normalization techniques such as SCTransform[35] and scran's deconvolution method[36] have become increasingly popular. Unfortunately, many specialized methods are written and maintained in only one coding language (usually R or Python). Consideration must be given for accessibility for users of both language ecosystems, and integration across normalization techniques is nontrivial. Additionally, some methods (for example, the scVI framework[21]) do not require explicit normalization a priori. Additional preprocessing steps such as mitochondrial percentage thresholding[37] and doublet removal[38] are also critical to consider before integration. While it is possible that optimal preprocessing algorithms exist, the primary requirements for collaboration are consistency and ease of use.

Given a standard choice for computational normalization, a low-dimensional embedding algorithm must be chosen for integration purposes. Currently, neural network-based methods such as scANVI and scVI (a semisupervised expansion of scVI)[39] are perhaps the best choice, as they enable easy updates and cell-type querying, are shown to preserve biological variation while removing technical variation, can utilize partial cell-type labeling from current reference atlases, and have already been shown to work with updateable transfer learning-based models such as scArches[20]. SCANORAMA[40] and fastMNN[41] also offer potential options for integration and have shown improved integration results over scANVI in benchmark tests. However, these methods

https://doi.org/10.1038/s41592-024-02390-8

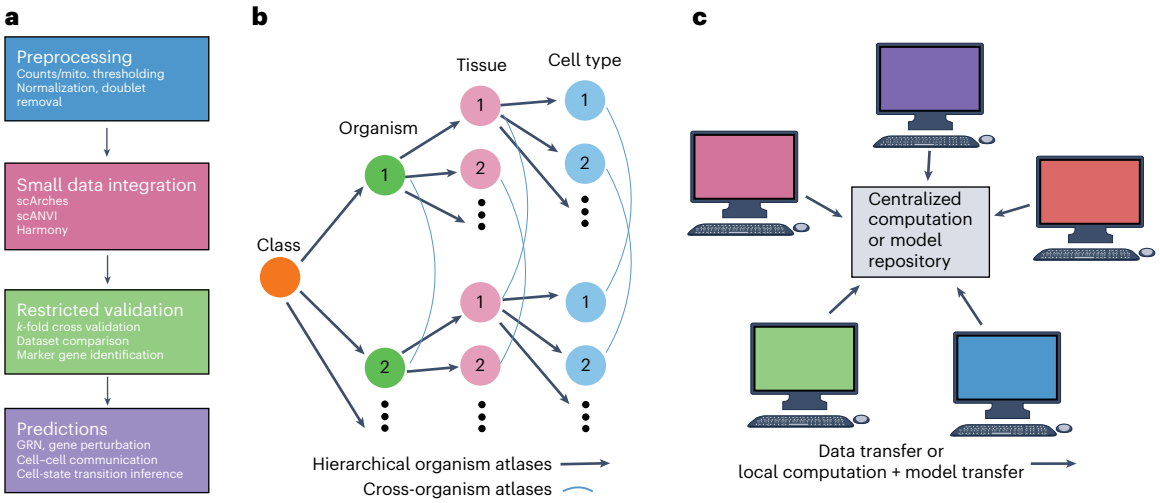

Fig. 1 | Constructing an updateable integrated cell atlas. a, Key points for construction of cell atlas. **b**, Graph model of hierarchical structure levels within classes (for example, taxonomical class or collection of model organism) indicated with black arrows. Blue lines indicate cross-organism integrative comparisons possible at each hierarchical level. **c**, Visualization of centralized versus distributed computational mechanisms.

utilize nearest-neighbor approaches, and it is not clear that they are scalable in terms of speed to collaborative atlases.

It is technically feasible to create an integrated atlas using the technologies discussed above, and indeed these techniques have been used many times to construct consortia-sized atlases. The critical point is that additional considerations regarding collaborative updating and accessibility are required to consider practical atlas updating and integration in a distributed framework.

Mechanisms for integration must maintain quality control and must ensure integration accuracy. A 'git-style'[42] version control approach, in which atlas updates are submitted for review before final validation and publication, is a possible mechanism. This would require submission of standardized quality-control metrics, among other possible factors. The updated atlas needs to be compared with the previous atlas to ensure previous cell-type resolution is retained. Additional potential forms of validation include *k*-fold cross-validation approaches[43], independent corroboration of new cell types within similar datasets, and identification of high-quality marker genes for new cell types. Critically, the whole atlas does not always need to be validated during updating, as most individual datasets are restricted to specific tissues and, in the case of cell sorting, cell types (Fig. 1a).

Also adapted from a git approach, atlas branches (Fig. 1b) provide a mechanism for in-depth analysis of cell subtypes, particularly important in the brain where over 5,000 replicable distinguishable cell types have been identified[16]. Additionally, this feature would enable comparison across genotype, disease condition and drug administration for individual cell types without compromising the primary cell atlas. In practice, hierarchical cross-condition atlases can be independently created at the organ and cell-type level, continuously updated as data from additional organisms and animal models are included, creating a continuously increasing computational resource for all scientists.

Finally, it is worth considering allocation of computational burden for atlas updating and integration (Fig. 1c). For ease of use and accessibility, one strategy would enable computation on remote servers maintained for this purpose. This would reduce the required number of data transfers (from user to atlas only), ensure equitable access across researchers, and substantially speed up analysis time frames, while expanding contextualization of the dataset. Additionally, more complex analyses (for example, pseudotime, GRN, computational perturbation) could be computed on these servers, enabling combined analysis across multiple datasets, drastically increasing the analysis power of individual datasets.

An alternative to centralized computation is a federated or distributed learning approach[44]. In this framework, individual models are trained locally by individual labs on their own data. A centralized model is then created via iterative updating based on the weights and losses computed on individual nodes. This is an extremely useful framework when data privacy is an issue, for example, when working with datasets in the PsychEncode database[45]. This framework can easily be adapted for iteratively updating centralized foundational models.

We also note the possibility of integrative frameworks with other omics modalities. In many ways, transcriptomics is the easiest modality to integrate across assays, due to the common feature set (genes). Integrative atlases of other omics modalities generally do not have this benefit. The assay for transposase-accessible chromatin with sequencing (ATAC-seq) for example, utilizes accessibility in genomic loci as its feature space, which is typically computed separately for each dataset[46]. Integration requires refinement of peak accessibility, and it is not yet clear which methods for creating integrated feature sets work best. However, it is known that uniform binning of the genome typically underperforms other feature selection methods, which implies difficulty in selecting a priori features that achieve optimal performance[47]. Additionally, these epigenetic assays are increasingly combined with transcriptomics (or other omics modalities), allowing RNA to serve as the 'bridge' for integrating these modalities, which at present may be a better approach than direct collaborative atlas creation with single-cell epigenomic assays.

Use cases for contextualization of 'small data' within 'big data' atlases

Integration (Fig. 2a) of individual transcriptomics assays with large datasets enables consistent interrogation of the same cell types across multiple studies. Computationally, enormous effort has been put into single-cell integration (Fig. 2b), enabling creation of consistent taxonomies across studies[18,32,39–41,48]. Herein we discuss three specific use cases for small data integration either with large-scale atlases, or with multiple smaller datasets: (1) computational perturbation analysis (Fig. 2c), (2) comparative GRN analysis (Fig. 2d) and (3) multispecies integration for translational medicine.

A case study in integration of mostly 'small data' to create large datasets, is the scPerturb database[49]. This database incorporates 44 individual datasets containing scRNA and epigenetic screens after (typically CRISPR) induced gene perturbations, primarily from cancer cell lines. scPerturb provides uniform cell-type annotation, RNA counts

https://doi.org/10.1038/s41592-024-02390-8

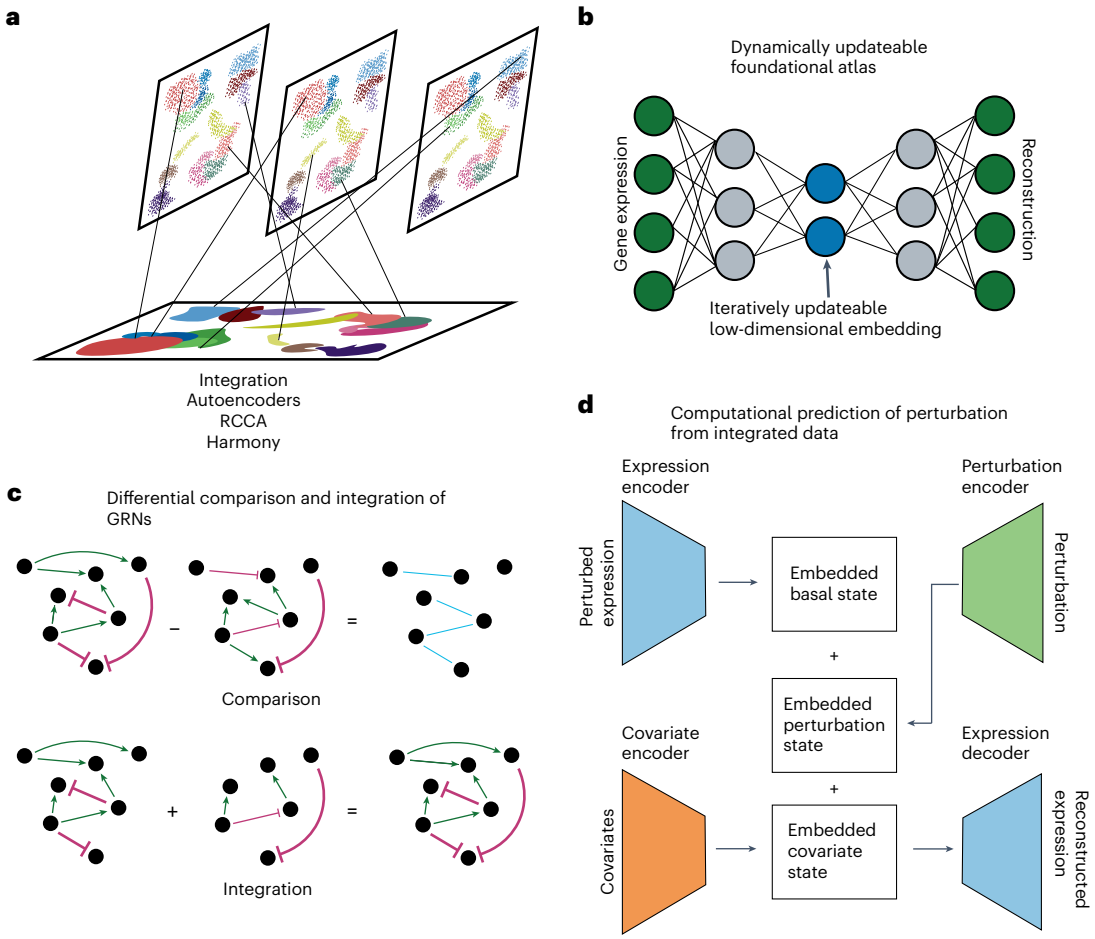

Fig. 2 | Applications of single-cell integrative foundational models. **a**–**d**, Single-cell integration (**a**), machine learning models for creating updateable single-cell atlases (**b**), post-integration GRN comparison and integration (**c**) and a machine learning model for computation of gene perturbation effects from integrated scRNA datasets (**d**), inspired by the computational perturbation autoencoder.

and DNA accessibility matrices to facilitate integrative computational analysis of RNA and epigenetic perturbation impact. This database, and others of its kind, provides a foundation for the computational analysis and prediction of gene perturbation impact via machine learning models such as the compositional perturbation autoencoder[50] or GEARS[33].

A second use case involves analysis of differential GRN alteration in the context of disease[51–53]. While integrated atlases cannot (currently) take the place of paired control animals for GRN studies, they still provide useful comparisons in two ways. First, comparison of inferred GRNs from control and atlas data provides a measure of expected statistical variation between samples, thereby enabling an additional significance measure for condition- and perturbation-dependent GRNs[54,55]. Second, identified GRNs can themselves be integrated within databases and frameworks, which will allow researchers to compare gene regulatory alterations in their disease, to those within similar or disparate conditions, thereby enabling contextualization of this information within the wider scheme of pathology.

Neuroscience in particular is primed for application of such methods, due to the large number of disease-associated mouse models[56,57], and the enormous influx of omics data from both specific brain regions, and whole brains[16,58,59]. A possible specific application of this approach would analyze transcriptomic alterations in various mouse models of Alzheimer's disease. Currently, there are dozens of Alzheimer's mouse models, exhibiting varying features (amyloid plaque and tau tangle deposition distribution and time frame, among others)[60,61], created using different genetic strategies. However, comparison of cell transcriptome alterations between models is frequently limited to comparison tables of differential gene or gene ensemble expression[62]. An integrated Alzheimer's atlas would enable precise comparison of disease progression and its impact on neural transcriptomics, including perturbation and gene regulatory analysis.

Finally, collaborative atlas creation would further enable cross-species comparisons across disease states and drug treatments[63]. This could improve the translation ability of medical approaches from basic to clinical science, by providing a common framework for determining whether treatment mechanisms of action are similar across species. As failure rates for translation of treatment approaches from animal models to humans remain over 90%[64], an integrative cross-species disease atlas could play a critical role in developing medical treatments. Overall, there is enormous potential for integrated transcriptomic atlases across molecular science and beyond.

Concluding remarks and future perspectives

Ultimately, a combination of small data methods will likely be used by scientists to collaboratively train models and build references atlases. Model training and sharing along with reference atlas updating allows users to both create their own custom models and references atlases, and to contribute to public models and atlases. This can pave the way for automated and standardized analyses of single-cell studies of brain tissue. By using transfer learning methods, users will share the most complete and recent models and references atlases, which can be trained and updated either locally or centrally. Thus, the entire field of biology will collaborate to generate a joint embedding, without the need to share full datasets, by mapping their own small-scale

datasets into the public reference atlas. Generalization to multimodal datasets will allow for reference atlas representations of nucleomics, epigenomics and proteomics, in addition to transcriptomics[65–67]. This effort will be enormously beneficial to individual labs as identification of subtle state-specific biological changes present in their one-off small-scale data will be discoverable when contextualized within the reference dataset.

A centralized effort to store, maintain, integrate and improve access to already existing databases is critical for enabling researchers to maximize the value of their individual assays, and would enable rapid comparison and analysis of animal and human disease and treatment models. This approach has the potential to improve our ability to identify molecular mechanisms of action across animal models, which may translate into an improved ability to translate discoveries in basic science into therapeutic approaches to human disease.

References

1. Ngai, J. BRAIN 2.0: transforming neuroscience. *Cell* **185**, 4–8 (2022).
2. BRAIN Initiative Cell Census Network. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
3. Regev, A. et al. The human cell atlas. *Elife* **6**, e27041 (2017). **Perhaps the largest single-cell atlas in the world.**
4. Landhuis, E. Neuroscience: big brain, big data. *Nature* **541**, 559–561 (2017).
5. Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013).
6. Todman, L. C., Bush, A. & Hood, A. S. 'Small data' for big insights in ecology. *Trends Ecol. Evol.* **38**, 615–622 (2023).
7. Ferguson, A. R. et al. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* **17**, 1442–1447 (2014).
8. Hekler, E. B. et al. Why we need a small data paradigm. *BMC Med.* **17**, 133 (2019).
9. Cai, C. et al. Transfer learning for drug discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).
10. Qi, G. -J. & Luo, J. Small data challenges in big data era: a survey of recent progress on unsupervised and semi-supervised methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 2168–2187 (2020).
11. Yang, L., Hanneke, S. & Carbonell, J. A theory of transfer learning with applications to active learning. *Mach. Learn.* **90**, 161–189 (2013).
12. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).
13. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
14. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
15. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* https://doi.org/10.1093/database/baaa073 (2020).
16. Yao, Z. et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023). **An incredible resource for analysis of transcriptomic diversity in the brain.**
17. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
18. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
19. Angerer, P. et al. Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).
20. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022). **An important resource for updateable atlas creation.**
21. Lopez, R. et al. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
22. Zhang, Z. et al. scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data. *Nat. Commun.* **15**, 912 (2024).
23. Zhou, Y. et al. Accurate integration of multiple heterogeneous single-cell RNA-seq data sets by learning contrastive biological variation. *Genome Res.* **33**, 750–762 (2023).
24. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* https://doi.org/10.1093/database/baz046 (2019).
25. Papatheodorou, I. et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2019).
26. Tarhan, L. et al. Single Cell Portal: an interactive home for single-cell genomics data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.13.548886 (2023).
27. CZI Single-Cell Biology Program et al. CZ CELL×GENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.30.563174 (2023).
28. Camps, J. et al. Meta-analysis of human cancer single-cell RNA-seq datasets using the IMMUcan database. *Cancer Res.* **83**, 363–373 (2023).
29. Li, X. -W. et al. SCAD-Brain: a public database of single cell RNA-seq data in human and mouse brains with Alzheimer's disease. *Front. Aging Neurosci.* **15**, 1157792 (2023).
30. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* https://doi.org/10.1038/s41592-024-02201-0 (2024).
31. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
32. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
33. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935 (2024).
34. Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by $\log(x+1)^{\dagger}$ or $\log(1+x)^{\dagger}$. *Bioinformatics* **37**, 2223–2224 (2021).
35. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
36. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
37. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).
38. Xi, N. M. & Li, J. J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.* **12**, 176–194 (2021).
39. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
40. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
41. Zhang, F., Wu, Y. & Tian, W. A novel approach to remove the batch effect of single-cell data. *Cell Discov.* **5**, 46 (2019).

42. Chacon, S. & Straub B. *Pro Git*. (Apress, 2014).
43. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. Preprint at https://arxiv.org/abs/1811.12808 (2018).
44. Verbraeken, J. et al. A survey on distributed machine learning. *ACM Comput. Surv.* **53**, 1–33 (2020).
45. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
46. Stuart, T. et al. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
47. Zhang, K. et al. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat. Methods* **21**, 217–227 (2024).
48. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
49. Peidli, S. et al. scPerturb: harmonized single-cell perturbation data. *Nat. Methods* **21**, 531–540 (2024).
50. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
51. Bocci, F., Zhou, P. & Nie, Q. spliceJAC: transition genes and state-specific gene regulation from single-cell transcriptome data. *Mol. Syst. Biol.* **18**, e11176 (2022).
52. Wang, J., Chen, Y. & Zou, Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. *PLoS Genet.* **19**, e1010942 (2023).
53. Badia-i-Mompel, P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
54. Duren, Z. et al. Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat. Commun.* **12**, 4763 (2021).
55. Kim, Y. et al. DiffGRN: differential gene regulatory network analysis. *Int. J. Data Min. Bioinform.* **20**, 362–379 (2018).
56. Götz, J., Bodea, L. -G. & Goedert, M. Rodent models for Alzheimer disease. *Nat. Rev. Neurosci.* **19**, 583–598 (2018).
57. Moulin, T. C. et al. Rodent and fly models in behavioral neuroscience: an evaluation of methodological advances, comparative research, and future perspectives. *Neurosci. Biobehav. Rev.* **120**, 1–12 (2021).
58. Zhang, M. et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).
59. Zu, S. et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature* **624**, 378–389 (2023).
60. Hall, A. M. & Roberson, E. D. Mouse models of Alzheimer's disease. *Brain Res. Bull.* **88**, 3–12 (2012).
61. McKean, N. E., Handley, R. R. & Snell, R. G. A review of the current mammalian models of Alzheimer's disease and challenges that need to be overcome. *Int. J. Mol. Sci.* **22**, 13168 (2021).
62. Li, Q. S. & De Muynck, L. Differentially expressed genes in Alzheimer's disease highlighting the roles of microglia genes including *OLR1* and astrocyte gene *CDK2AP1*. *Brain Behav. Immun. Health* **13**, 100227 (2021).
63. Bakken, T. E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
64. Marshall, L. J. et al. Poor translatability of biomedical research using animals—a narrative review. *Altern. Lab. Anim.* **51**, 102–135 (2023).
65. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
66. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future. *Science* **358**, 69–75 (2017).
67. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

Acknowledgements

This work was supported by National Institutes of Health (NIH) grants UM1MH130994, U01AG076791, U01DA052769, R01AG067153, R01AG082127 and RF1AG065675 to X.X. and the Knights Templar Eye Foundation grant KTEF-5646361 to S.F.G. F.J.T. acknowledges support from the German Federal Ministry of Education and Research (BMBF; 031L0210A) and from the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI (ZT-I-PF-5-01). Q.N. acknowledges support from National Science Foundation grants DMS1763272, MCB202842 and CBET2134916, and NIH grants R01AR079150, R01ED030565 and U01AR073159. K.G.J. acknowledges support from NIH grant T32 DC010775-14.

Author contributions

K.G.J. and S.F.G. wrote the paper and created the figures. Q.N. and F.J.T. and oversaw the writing. X.X. oversaw and supported the work.

Competing interests

F.J.T. consults for Immunai, Singularity Bio B.V., CytoReason and Omniscope, and has ownership interest in Dermagnostix GmbH and Cellarity.

Additional information

Correspondence and requests for materials should be addressed to Qing Nie, Fabian J. Theis or Xiangmin Xu.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team.