nature biotechnology

REVIEW ARTICLE

https://doi.org/10.1038/s41587-021-00895-7

Check for updates

Computational principles and challenges in single-cell data integration

Ricard Argelaguet[1,2] ✉, Anna S. E. Cuomo[1,3] ✉, Oliver Stegle[3,4,5] ✉ and John C. Marioni[1,3,6] ✉

The development of single-cell multimodal assays provides a powerful tool for investigating multiple dimensions of cellular heterogeneity, enabling new insights into development, tissue homeostasis and disease. A key challenge in the analysis of single-cell multimodal data is to devise appropriate strategies for tying together data across different modalities. The term 'data integration' has been used to describe this task, encompassing a broad collection of approaches ranging from batch correction of individual omics datasets to association of chromatin accessibility and genetic variation with transcription. Although existing integration strategies exploit similar mathematical ideas, they typically have distinct goals and rely on different principles and assumptions. Consequently, new definitions and concepts are needed to contextualize existing methods and to enable development of new methods.

Technological innovations continue to drive the establishment of single-cell genomics assays, which allow the interrogation of a growing number of molecular dimensions, including the genome, transcriptome and epigenetic modifications, at high resolution and across thousands of cells. Although no single 'omics' technology can fully capture the intricacy of complex molecular mechanisms, collectively, these data have the potential to provide a comprehensive picture of basic biological processes such as early mammalian development[1,2], as well as clinically relevant traits, including cancer[1]. Multimodal measurements, where different molecular features can be probed in the same cell, can already be obtained using a variety of experimental techniques[3–7], which have been reviewed elsewhere[8–10].

One of the most promising features of multimodal sequencing is the opportunity to advance from descriptive 'snapshots' toward a mechanistic understanding of gene regulation. By incorporating prior knowledge about the hierarchical relationship between molecular layers (that is, the central dogma of biology), multimodal assays will have an important role in identifying causal chains of events in gene regulatory networks. Moreover, multimodal assays have already been shown to allow more refined identification of cell types and cell states, for example, in the context of the immune system[11].

None of the biological insights offered by multimodal assays would be possible without concomitant development of computational methods. Each new data modality presents distinct challenges and needs, ranging from low-level processing, quality control and normalization to downstream analysis and interpretation such as quantification of sources of biological variability, which are then used to generate testable biological hypotheses. In particular, a key challenge in the analysis of single-cell multimodal data is to devise efficient computational strategies to integrate different data modalities[8,12–15]. The term data integration has generally been used to describe algorithms and software for this task, encompassing a wide range of distinct computational strategies based on different principles and assumptions. There is a need to define unifying concepts for these data integration tasks, to contextualize existing and future strategies depending on input data structure and the specific integration task at hand.

In this Review, we introduce basic concepts that underpin single-cell data integration techniques and discuss alternative choices of anchors for linking different datasets. We review the established principles, limitations and diagnostics of data integration strategies and highlight parallels between approaches for genetic analysis of single-cell traits and inference of regulatory dependencies between molecular layers. Finally, we discuss future challenges related to the integration of single-cell molecular profiles across physical dimensions, such as space and time, as well as multiscale modeling strategies for tying cellular representations to medically relevant human traits.

Input data and definition of anchors

The first step in any data integration pipeline is the selection of an anchor to link the different data modalities. In practice, this choice is usually driven by the experimental design, but it has fundamental implications for downstream analysis, as different choices for the anchor entail different statistical and biological assumptions and therefore require tailored methodologies. Depending on the choice of anchor, three types of data integration strategies can be distinguished (Fig. 1):

- **Genomic features as the anchor (horizontal integration):** for experimental designs where the same data modality is profiled from independent groups of cells (unmatched assays). An example could be single-cell RNA sequencing (scRNA-seq) experiments profiling cells from the same tissue across different groups of donors or combining data across different scRNA-seq technologies, where the assays are anchored by their common gene set.
- **Cells as the anchor (vertical integration):** for experimental designs where multiple data modalities are simultaneously profiled from the same cells (matched assays). This is exemplified by assays such as single-cell methylome and transcriptome

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [2]Epigenetics Programme, Babraham Institute, Cambridge, UK. [3]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [4]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [5]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [6]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. ✉e-mail: ricard@ebi.ac.uk; acuomo@ebi.ac.uk; o.stegle@dkfz-heidelberg.de; marioni@ebi.ac.uk

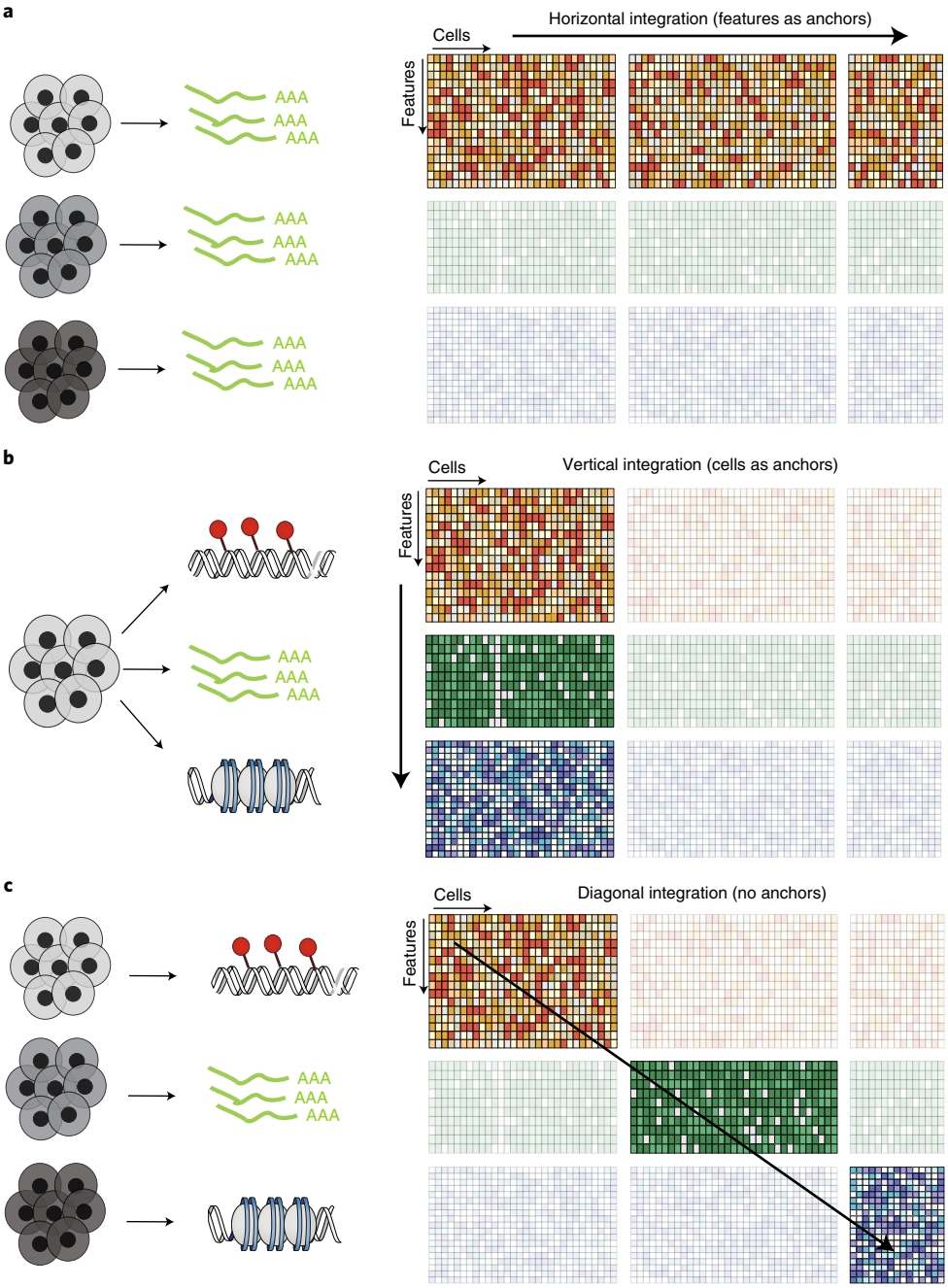

Fig. 1 | Alternative choices of anchors for data integration. a–c, Depending on the anchor choice, three types of data integration strategies can be considered: horizontal integration with features as the anchors (**a**), vertical integration with cells as the anchors (**b**) and diagonal integration with no anchors in high-dimensional space (**c**). The left column shows the data modalities extracted, while the right column illustrates the resulting data matrices to be integrated, depending on the anchor choice.

sequencing (scM&T-seq[6]), single-cell analysis of genotype, expression and methylation (sc-GEM[16]), cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq[5]) and single-nucleus chromatin accessibility and RNA expression sequencing (SNARE-seq[17] and SHARE-seq[4]).

- **No anchor in high-dimensional space (diagonal integration):** for experimental designs where both cells and genomic features are different between experiments. An example is when scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) are applied to separate groups of cells.

Defining the methodology

The choice of anchor determines the data integration strategy, that is, horizontal, vertical or diagonal integration (Fig. 1). For each task, a variety of methods (Table 1) and datasets for benchmarking (Table 2) exist, some of which we discuss below.

Strategies for horizontal integration. Horizontal integration strategies define features as the anchor in unmatched experiments of the same type. This task is commonly faced in large-scale scRNA-seq projects where data are generated across multiple batches and technologies, as technical factors introduce differences that can result in

Table 1 | Overview of common data integration methods classified according to their anchor choice

Integration task	Method	Ref.
Vertical (global)	CCA	112
Vertical (global)	JIVE	70
Vertical (global)	PLS	71
Vertical (global)	MCIA	113
Vertical (global)	MOFA+	65
Vertical (global)	scAI	114
Vertical (global)	iNMF	38
Vertical (global)	Seurat v4	11
Vertical (local)	Spearman's rank correlation coefficient	50
Vertical (local)	LMM	51
Horizontal	MNN	21
Horizontal	Seurat v3	22
Horizontal	LIGER	23
Horizontal	Harmony	24
Horizontal	Scanorama	29
Horizontal	BBKNN	25
Horizontal	scVI	26
Horizontal	scmap	28
Horizontal	conos	27
Diagonal	MATCHER	77
Diagonal	MMD-MMA	78
Diagonal	SCIM	115
Diagonal	UnionCom	116
Diagonal	coupledNMF	117

systematic deviations in the distribution of observed RNA expression counts (or even cell type composition). If left unaccounted for, these sources of technical variation can mask relevant biological variability and thus complicate interpretation of downstream analysis. Frequently, horizontal integration is formulated as a batch correction problem aimed at removing undesired technical variation across experiments while preserving genuine biological variation within and between experiments. With the growing availability of reference atlases at single-cell resolution, epitomized by the Human Cell Atlas project[18], this is arguably one of the most important steps in single-cell analysis workflows.

Naive application of linear batch correction methods that were originally developed for bulk datasets (limma[19] and ComBat[20]) has proven insufficient for single-cell experiments, mainly because these methods implicitly assume identical (or at least known) cell type composition across batches[21]. In practice, however, the abundance of cellular subpopulations can vary even between biological replicates owing to subtle differences in sample collection and library preparation. As a consequence, the majority of horizontal integration methods developed for single-cell data rely on nonlinear (or locally linear) strategies that account for differences in cell type composition. Several integrative methods for batch correction of single-cell data have been developed, including MNN[21], Seurat v3 (ref. [22]), LIGER[23], Harmony[24], BBKNN[25], scVI[26], conos[27], scmap[28], Scanorama[29] and scAlign[30], among others. Despite having common principles, these each use different methodologies. MNN and Seurat v3 match mutual nearest neighbors in a joint low-dimensional space, defined by either principal components or canonical covariates. LIGER performs integrative non-negative matrix factorization (NMF) and disentangles dataset-specific factors from shared factors, followed by construction of a neighborhood graph using only the shared factors. BBKNN performs correction on a neighborhood graph, which results in much faster computation at the expense of losing single-cell resolution. Harmony learns a cell-specific linear correction function through successive rounds of *k*-means clustering on a principal component space. Finally, scVI is a Bayesian variational autoencoder with a probabilistic formulation that accounts for batch-specific variation.

In addition to having similar mathematical principles, most of these methods also have a common set of challenges. First, a classical problem of nonlinear integration methods is overcorrection, which occurs when the batch correction vector is incorrectly estimated and the algorithm forcibly merges nonmatching subpopulations of cells[31]. This can occur, for example, when no shared axes of biological variation are preserved between the datasets (for example, when there are no common cell types). An optimal method should be able to detect this and prevent merging of datasets when no common biological variation exists. Second, most methods perform the integration step with cells embedded in latent space[21–24]. This undoubtedly improves the performance of most batch correction methods by removing noise and decreasing the computational cost. However, the high-dimensional observations (for example, gene expression counts) can be severely distorted as a result of the batch alignment, and other downstream gene-based analyses such as gene marker detection or differential expression analysis can become problematic[32]. Third, when extensive biological variability exists across batches, disentangling batch effects from the underlying biological signal of interest is challenging. For example, when samples are profiled across a developmental time course, it is often difficult to randomize samples at different time points as part of the experimental design.

For a more complete description of the challenges of horizontal integration, as well as benchmarking strategies, we refer the reader to ref. [31], where the researchers compared the performance of 38 methods using increasingly complex datasets and a range of metrics that included scalability, usability and ability to remove batch effects while retaining biological variation.

Strategies for vertical integration. Vertical integration strategies take advantage of the unambiguous assignment between molecular profiles in matched multimodal experiments and thus define cells or groups of cells (for example, cells sampled from the same individuals) as the anchor between data modalities.

Vertical integration methodologies can be further classified into local versus global approaches, a notation inspired by integrative approaches that have been pursued at the bulk level[19,33]. Local analyses refer to associations between specific features across different molecular layers, often with the aim to detect putative interactions between them (for example, associations between genetic variants and gene expression, that is, expression quantitative trait loci (eQTLs[32]); Fig. 2). Global integrations, on the other hand, exploit the full spectrum of measurements to identify broader cellular states, such as cell cycle phase and pluripotency potential[34]. Global analyses typically identify patterns of covariation across genomic features and layers.

Local integration. Prominent examples of local analyses are associations between genetic variants and gene expression (eQTLs) or between the epigenetic status of putative regulatory elements and the expression of nearby genes. Restriction to a defined search space is often necessary, ensuring that the problem remains tractable, from both a computational and statistical perspective, and aiding biological interpretation. For example, *cis*-eQTL mapping, where only genetic variants near a gene's genomic location are interrogated, is

Table 2 | Overview of datasets that can be used for benchmarking horizontal, vertical, diagonal and mosaic integration tasks

Integration task	Biological system	Number of cells	Data modality	Technology	Ref.
Horizontal	Mouse brain	156,049	RNA expression	scRNA-seq (SPLiT-seq)	118
	Mouse brain	509,876	RNA expression	scRNA-seq (10x Genomics Chromium)	119
Horizontal	Human lung	690,000	RNA expression	scRNA-seq (Drop-seq)	120
	Human lung	32,472	RNA expression	scRNA-seq (10x Genomics Chromium)	121
	Human lung	65,662	RNA expression	scRNA-seq (10x Genomics Chromium)	122
	Human lung	9,404	RNA expression	scRNA-seq (Smart-seq2)	122
	Human lung	46,500	RNA expression	scRNA-seq (snRNA-seq)	123
Horizontal	Human pancreas	2,126	RNA expression	scRNA-seq (CEL-Seq2)	124
	Human pancreas	978	RNA expression	scRNA-seq (Fluidigm C1)	125
	Human pancreas	2,209	RNA expression	scRNA-seq (Smart-seq2)	126
	Human pancreas	8,569	RNA expression	scRNA-seq (inDrop)	127
Vertical (local + global)	Human white blood cells	161,764	RNA expression + surface proteins	CITE-seq	11
	Human white blood cells	10,000	RNA expression + chromatin accessibility	Multiome 10x	https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets
Vertical (local + global)	Mouse gastrulation	1,105	RNA expression + chromatin accessibility + DNA methylation	scNMT-seq	64
Vertical (local + global)	Mouse skin	34,774	RNA expression + chromatin accessibility	SHARE-seq	66
Vertical (local + global)	Mouse cerebral cortex	5,081	RNA expression + chromatin accessibility	SNARE-seq	17
Vertical (local + global)	Human brain	4,358	RNA expression + chromatin accessibility + DNA methylation	snmC2T-seq	75
Vertical (local)	Human PBMCs	28,855	RNA expression + genotypes	scRNA-seq (10x Genomics Chromium) + genotype chips	50
Vertical (local)	Human iPSCs	5,447	RNA expression + genotypes	scRNA-seq (Fluidigm C1) + genotype chips	54
Vertical (local)	Human iPSCs (to endoderm)	36,044	RNA expression + genotypes	scRNA-seq (Smart-seq2) + genotype chips	51
Vertical (local)	Human iPSCs (to dopaminergic neurons)	1,027,401	RNA expression + genotypes	scRNA-seq (10x Genomics Chromium) + genotype chips	57
Diagonal	Human fetal tissue	4,000,000	RNA expression	scRNA-seq (sci-RNA-seq)	128
	Human fetal tissue	800,000	Chromatin accessibility	scATAC-seq (sci-ATAC-seq)	129
Diagonal	Bone marrow and white blood cells	35,582	RNA expression + surface protein	CITE-seq	74
	Bone marrow and white blood cells	35,038	Chromatin accessibility	scATAC-seq	74
Diagonal	*Drosophila* eye disc	3,531	RNA expression	scRNA-seq (10x Genomics Chromium)	130
	Drosophila eye disc	15,766	Chromatin accessibility	scATAC-seq	130
Diagonal	Mouse gastrulation (E8.25)	15,935	RNA expression	scRNA-seq (10x Genomics Chromium)	88

Continued

Table 2 | Overview of datasets that can be used for benchmarking horizontal, vertical, diagonal and mosaic integration tasks (continued)

Integration task	Biological system	Number of cells	Data modality	Technology	Ref.
	Mouse gastrulation (E8.25)	19,453	Chromatin accessibility	scATAC-seq	131
Diagonal	Mouse brain	509,876	RNA expression	scRNA-seq (10x Genomics Chromium)	119
	Mouse brain	5,000	Chromatin accessibility	scATAC-seq	https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k
	Mouse brain	15,000	Chromatin accessibility	scATAC-seq	132
Diagonal	Human lung	46,500	RNA expression	scRNA-seq (snRNA-seq)	123
	Human lung	90,980	Chromatin accessibility	scATAC-seq	123
Mosaic	Mouse gastrulation	116,312	RNA expression	scRNA-seq	88
	Mouse gastrulation	19,453	Chromatin accessibility	scATAC-seq	131
	Mouse gastrulation	1,105	RNA + chromatin accessibility + DNA methylation	scNMT-seq	64
Mosaic	Human white blood cells	10,000	RNA expression	scRNA-seq	https://support.10xgenomics.com/single-cell-gene-expression/datasets
	Human white blood cells	10,000	Chromatin accessibility	scATAC-seq	https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_pbmc_10k_v1
	Human white blood cells	161,764	RNA expression + surface proteins	CITE-seq	11
	Human white blood cells	10,000	RNA expression + chromatin accessibility	Multiome 10x	https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets
Mosaic	Human brain	2,784	DNA methylation	snmC-seq2	133
	Human brain	4,358	RNA expression + chromatin accessibility + DNA methylation	snmC2T-seq	75
	Human brain	23,005	RNA expression	scRNA-seq	75
	Human brain	12,557	Chromatin accessibility	scATAC-seq	75
	Human brain	4,200	DNA methylation + chromatin conformation	sn-m3C-seq	134
Mosaic	Mouse brain	5,000	RNA expression + chromatin accessibility	Multiome 10x	https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/e18_mouse_brain_fresh_5k
	Mouse brain	3,377	DNA methylation	scBS-seq	133
	Mouse brain	15,000	Chromatin accessibility	scATAC-seq	132
	Mouse brain	509,876	RNA expression	scRNA-seq (10x Genomics Chromium)	119

PBMCs, peripheral blood mononuclear cells.

most prevalent because the multiple-testing burden is reduced and because the underlying molecular mechanisms can be more directly interpreted[35,36]. Similarly, the first parallel RNA expression, DNA methylation and/or chromatin accessibility assays have been used to probe dependencies between proximal regulatory elements and gene expression levels[3,6,37–39].

Local analyses are typically supervised tasks, where specific combinations of features are tested for association on the basis of some prior knowledge of mechanism. From a methodological perspective, most local analyses use regression models[40]. In local analyses, the challenge is to distinguish true interactions between features from spurious associations that can result from confounding

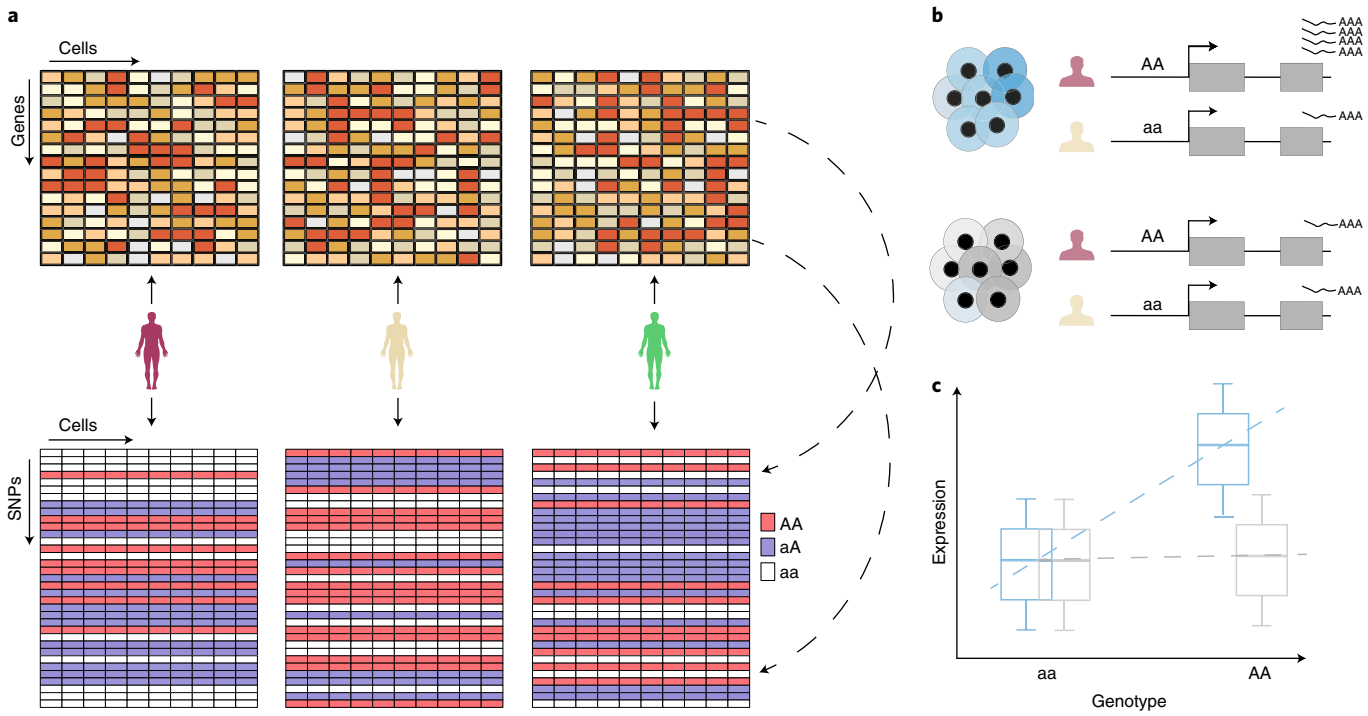

Fig. 2 | Cell-type-specific eQTL mapping as an example of local vertical integration. a, Single-cell eQTL mapping is an example of vertical integration, where the two modalities considered are RNA expression (top) and genotype at genetic loci (SNPs; bottom). Within an individual, and at a given SNP, all cells have the same genotype value. *Cis*-eQTL mapping is a local vertical integration task, where specific gene–SNP pairs are tested for association, as indicated by the arrows. **b**,**c**, Example of a cell-type-specific eQTL. In the illustration (**b**), genotype A results in increased expression of a gene of interest in one cell type (blue) but not in the other (gray). Alternative representation using boxplots (**c**). The average expression of the gene of interest (*y* axis) is increased as a function of the genotype at the SNP of interest (*x* axis) in one cell type (blue) but not in the other (gray).

sources of variation or upstream factors that drive coordinated changes across many elements. For example, in eQTL analyses, sample substructure, due to relatedness and population stratification as well as the repeated nature of the observations (that is, cells) for the same donor, is an important confounder to take into account (Fig. 2). Linear mixed models (LMMs) are a popular analytical framework to address this challenge[41–46]. In LMMs, a random effect component is added to the linear regression framework to adjust for such confounding dependencies. LMMs are widely used in genetics[41–46] but have also been applied in molecular association analyses (for example, see ref. [47]) (Fig. 2).

Moreover, latent variable models such as principal-component analysis (PCA) and probabilistic estimation of expression residuals (PEER)[48] can be used to identify, in an unsupervised fashion, factors affecting gene expression in a global manner, thus efficiently capturing known and hidden covariates affecting expression across all genes. These factors can then be added to the models as covariates (alongside other factors such as sex, the age of individuals or the experimental batch) to control for additional phenotypic variation, often leading to improved statistical power.

Mapping eQTLs using single-cell expression profiles has led to the identification of cell-type-specific eQTLs (which are prevalent, as previously demonstrated[49]) in rare cell populations, which would have been masked using bulk assays[43]. Additionally, van der Wijst et al.[50] and Cuomo et al.[51] combined differentiation of induced pluripotent stem cells (iPSCs) across multiple donors and single-cell expression profiles as well as allele-specific expression to show how eQTLs influence expression dynamically along the developmental axis (extending work from ref. [52]) and cellular context. Single-cell eQTL mapping is growing as a field and promises to provide an additional layer to understanding of genetic regulation at the molecular level[53–57]. As methods to assay various molecular traits at single-cell resolution become more established, other flavors of single-cell QTL mapping, where genomic variants are associated with changes in DNA methylation, histone modification or protein level at single-cell resolution, will likely become routine. Similar methods can be applied to test for regulatory effects of genetic perturbations (rather than natural variation) on different molecular readouts, through the use of assays such as Perturb-seq[58,59],CROP-seq[60] and TAP-seq[61]. As an example, candidate enhancers were perturbed using CRISPR technology followed by testing for associations with changes in gene expression, identifying hundreds of high-confidence *cis* enhancer–gene pairs in the K562 chronic myelogenous leukemia cell line[62]. As another example, Mimitou et al. introduced ECCITE-seq, a multimodal CRISPR-based screen that allows for the simultaneous detection of changes in transcriptome and protein abundance in the context of different combinations of induced mutations[63].

Besides eQTL mapping, local integration can be performed for any pairs of molecular layers. For example, gene expression has been correlated with matching promoter DNA methylation and chromatin accessibility during mouse gastrulation, revealing that the expression of pluripotency genes is decreased by the induction of a repressive epigenetic landscape[64,65]. In the past several years, highly scalable approaches to simultaneously measure chromatin accessibility and gene expression in thousands of cells[66] have increased the power to identify *cis*-regulatory interactions. These kinds of association analyses can be performed using the same LMM frameworks commonly used for eQTL mapping. The challenge, once again, is to account for hidden common drivers of variability across modalities that can result in spurious associations. A common confounding factor, especially in epigenetic analyses, is sequence context. For example, ATAC-seq peaks with high G+C content are associated with higher levels of chromatin accessibility, and DNA methylation regions with high C+G density are associated with low DNA

Box 1 | Statistical challenges associated with single-cell multimodal analysis

There are several statistical challenges for data integration. Below, we highlight key aspects.

- **Heterogeneous data modalities.** Molecular readouts collected using different assays generally have distinct statistical properties and require bespoke methods with different statistical assumptions. For example, scM&T-seq[6] yields mRNA expression counts that can be modeled using a negative binomial distribution. However, DNA methylation readouts are binary (each CpG site is either methylated or unmethylated). Combining different likelihood models in a single inference framework is not a trivial statistical task.

- **Overfitting.** As the number of molecular layers increases (and the number of features), modeling strategies face the risk of overfitting if not appropriately regularized. Following from the example above, scM&T-seq captures the methylation status for potentially millions of CpG sites, but experimental designs are typically restricted to only a few hundred cells. This is a classic case of a large p (number of features) and small n (number of observations) problem in high-dimensional statistics[135].

- **Missing data.** A major problem associated with some single-cell methodologies is the large amount of missing information. Importantly, assays differ in terms of how missing data is defined. For example, for bisulfite sequencing methods (scM&T-seq[6], scNMT-seq[3] and scCOOL-seq[136]), missing values are distinguishable from observed values. However, for scRNA-seq and scATAC-seq, an absence of sequence reads does not distinguish between the event that the genomic feature was not measured and the event that the readout was indeed zero[137]. Handling of missing information is an important aspect of multiomics data integration, as some of the conventional implementations of popular statistical methods such as linear regression and PCA do not handle missing information.

- **Delineating biological versus technical noise.** Multiomics datasets from complex experimental designs typically contain multiple sources of heterogeneity, both technical and biological. If not accounted for, technical variability can mask biological signals of interest[138]. Understanding and correcting technical variation is a critical step to ensure successful computational analysis[72].

- **Scalability.** As sequencing cost decreases and technologies improve, we anticipate that multimodal datasets will follow a trend similar to that seen with scRNA-seq, where in the span of less than 10 years the size of experiments increased from the order of tens to potentially millions of cells[139]. Querying exceptionally large datasets requires fast computational methods that typically rely on stochastic inference schemes[26,65].

- **Assay noise.** Because of the small amounts of starting material, single-cell technologies are inherently noisy and result in large amounts of technical noise[140]. To overcome this challenge, computational frameworks use information on the similarities between cells and/or genes to delineate signal from noise. Prominent examples are normalization methodologies based on Bayesian approaches that are able to borrow information across cells and/or genes and propagate uncertainty when performing inference and predictions[141,142].

- **Principled validation and assessment of model outputs.** Assessment of data integration outputs is one of the most challenging steps. Accurate ground truth information is seldom available, and hence this assessment relies on a combination of statistical quality metrics, as well as qualitative assessment of the impact of alternative integration strategies on downstream analysis tasks (that is, differential expression, dimensionality reduction, clustering, etc.).

methylation levels. To account for this, methods such as chromVAR[67] perform association testing by designing, for each feature, a null distribution using randomly selected features with a matching sequence context. Additionally, there can be confounding factors that affect only one modality, but across all features—for example, differences in global methylation levels caused by differences in cell type. Similarly to eQTL analyses, where tools such as PCA and PEER are used to guard against this, such cell-centric attributes can be added as covariates in a linear regression framework.

Global integration. Although local analyses are useful for identifying putative regulatory elements and their effect on gene expression, they have limited capacity to discover complex molecular maps resulting from interactions between multiple genomic features. An alternative strategy for data integration is to exploit the full spectrum of measurements to identify broader cellular states. For example, cell cycle phase, pluripotency potential and differentiation state are properties that are determined by gene regulatory networks and cannot be studied with local analyses that test one feature at a time. Thus, global integration is typically performed using unsupervised dimensionality reduction approaches that exploit covariation patterns across genomic features.

PCA is the paradigmatic method for global analyses of data when focused on a single modality. PCA infers an orthogonal projection of the data onto a low-dimensional space that maximizes the variance explained by the projection. The key to the popularity of PCA is its linearity assumption, which ensures that the resulting principal components are simple and interpretable. Generalizations of PCA for the integration of multiomics data have been devised by adapting multiview learning methods from the statistics literature[68]. Although most of these methods were originally developed for bulk data or applications outside genomics, the majority can be adapted to single-cell multimodal data. These approaches include unsupervised dimensionality reduction methods that represent different flavors of matrix factorization, such as canonical correlation analysis (CCA, implemented in Seurat[18,22]), MOFA[69], JIVE[70], PLS[71], MCIA[72] and iNMF[62], among others. Each of these methods builds on the matrix factorization framework, which has been notably successful owing to its simplicity, interpretability, computational speed and reduced risk of overfitting[27].

A key challenge of global analysis is to quantify the coupling between data modalities. In the case of CCA, the aim is to find a latent representation that maximizes the covariation between two data modalities, thus neglecting variation that is not shared by the data modalities. MOFA can be regarded as a generalization of CCA that builds on the Bayesian group factor analysis framework[73] to handle an arbitrary number of data modalities. By using structured sparsity priors, MOFA is capable of detecting sources of covariation between different sets of data modalities as well as sources of variation that are only present in a single data modality.

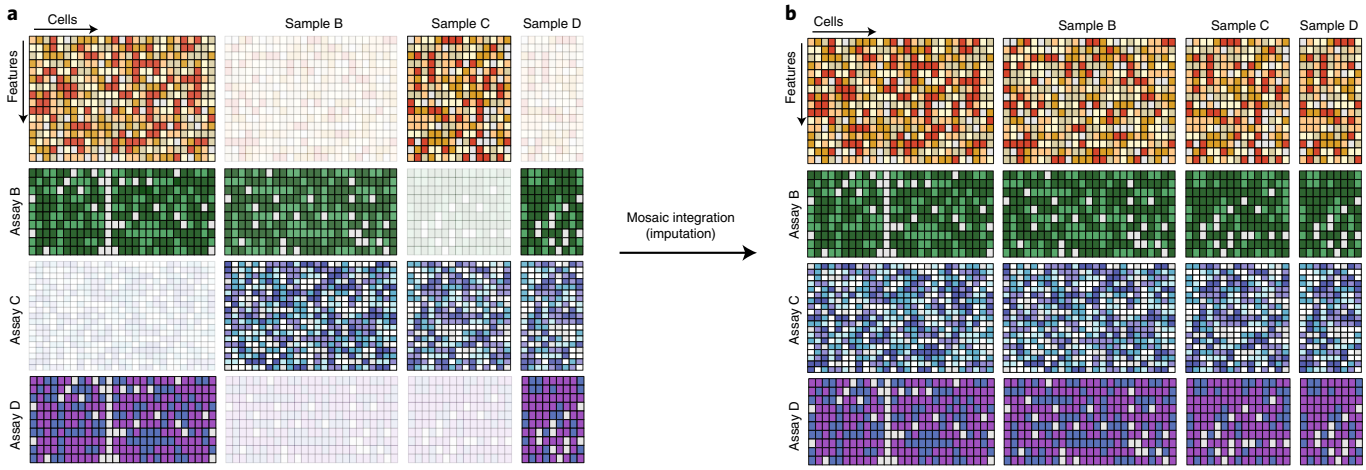

Fig. 3 | Mosaic integration. a, Overview of an experimental design where different data modalities (each block in the rows) are profiled in different subsets of cells (each block in the columns). Transparent matrices denote missing information. **b**, Resulting data matrices after applying a mosaic integration approach aimed at imputing missing data modalities.

Some of the matrix factorization methods have been applied to single-cell multimodal experiments and have revealed important biological insights. In a dataset of mouse preimplantation embryos where scCAT-seq was used to simultaneously profile chromatin accessibility and RNA expression, iNMF enabled the extraction of interpretable molecular signatures associated with distinct cell states during the blastocyst stage[18]. Similarly, in a dataset of mouse postimplantation embryos where scNMT-seq was used to simultaneously profile RNA expression, DNA methylation and chromatin accessibility, an integrative analysis using MOFA revealed the existence of lineage-specific enhancers that are associated with germ layer commitment[64].

Most of these methods have a common set of challenges and diagnostics (Box 1). First, molecular readouts collected using different techniques generally have heterogeneous statistical properties and thus have to be modeled under different assumptions. For example, the RNA component of scM&T-seq yields mRNA expression counts, whereas the DNA methylation component yields binary readouts (each CpG site is either methylated or unmethylated). Combining different likelihood models in the same statistical framework is not a trivial task. Second, different data modalities can have vastly different numbers of features. For example, a typical CITE-seq experiment consists of dozens of antibodies but thousands of gene expression measurements. This feature imbalance can have a strong influence on dimensionality reduction models, such that bigger data modalities disproportionately contribute to the latent space. A related challenge is differences in the signal-to-noise ratio across assays, which can be taken into account during the integration step to appropriately weight different assays[11]. Third, all matrix factorization methods report a solution, but its quality can be hard to assess. The amount of variance explained by the latent representation can be a useful diagnostic, but determining the model fit more broadly can be difficult. Apart from assessing statistical measures of the goodness of fit, it is recommended to explore how robust the matrix factorization solution is to perturbations in the input data. This generally involves assessing the consistency of the latent factors when bootstrapping or downsampling the dataset. Finally, linearity is arguably the biggest advantage of most matrix factorization methods, but it comes at the cost of a substantial loss of explanatory power. Nonlinear alternatives, such as deep generative models in the form of variational autoencoders, have proven to be powerful generalizations of factor analysis and have been successfully applied to a variety of single-cell genomics technologies[36,64], albeit at the cost of reduced interpretability.

Although we have focused our discussion on approaches based on matrix factorization, other strategies for vertical data integration have also been conceived. Hao et al. extended nearest neighbor graphs to a multimodal setting[11]. This method, called weighted nearest neighbor (WNN) analysis (implemented in Seurat v4), yields a latent representation that enables joint definition of cellular states across data modalities.

Strategies for diagonal integration. The third type of data integration problem arises when no anchor exists in the high-dimensional space. This task occurs in unmatched experiments where different molecular layers are profiled in different subsets of cells, for example, when performing scRNA-seq and scATAC-seq in separate groups of cells. In comparison to the horizontal and vertical integration tasks, diagonal integration is much more challenging and the biological insights that can be obtained from diagonal integration are often more difficult to interpret and validate.

Diagonal integration methods generally aim to reconstruct a low-dimensional manifold that captures covariation across two (or more) data modalities. Thus, a critical assumption of this strategy is the existence of a latent manifold that is, to some extent, preserved between the data modalities. For example, this could represent cells sampled from a common differentiation trajectory or cells sampled from a common set of subpopulations.

Several studies[74–76] have addressed the diagonal integration task by simplifying it to a vertical or horizontal integration task. In the case of vertical integration, the data have to be summarized such that the samples unambiguously match between assays. For example, this can be achieved by aggregating cells into cell types that are confidently identified in both assays. However, this strategy makes strong assumptions on the definition of cell types, and single-cell resolution is in part lost. In the case of horizontal integration, the data have to be summarized such that the features are linked between assays by one-to-one mapping (for example, gene body accessibility and gene expression). Using this strategy, horizontal methods such as LIGER[33] and Seurat v3 (ref. [22]) have successfully integrated unmatched epigenetic and transcriptomic experiments from the same tissue and even across different species. However, this strategy relies on fragile biological assumptions and can fail in scenarios where such linkages are incomplete and when the

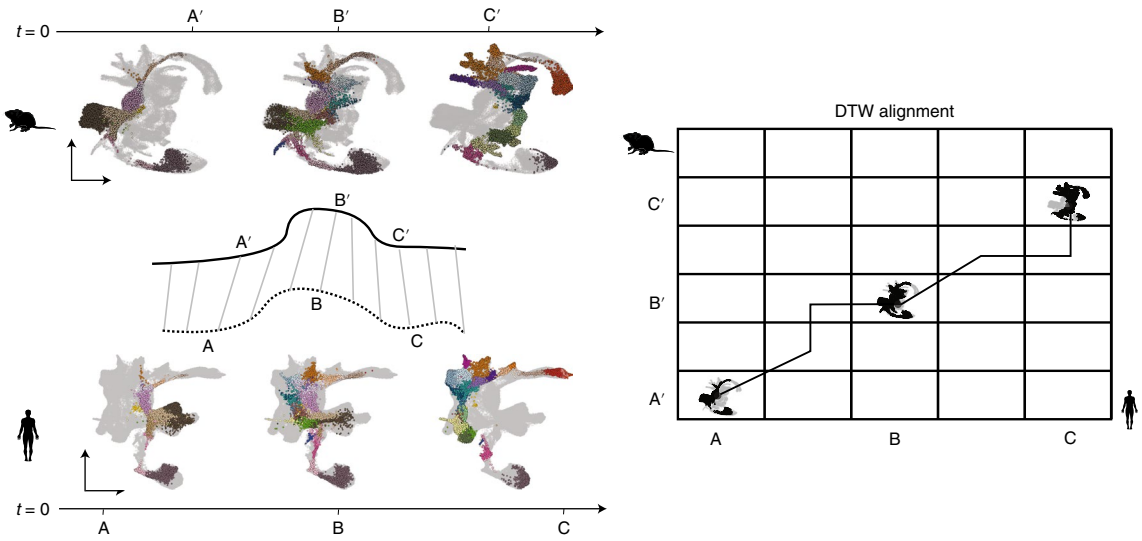

Fig. 4 | Mapping time-resolved single-cell genomics experiments across species. Illustration of a time course single-cell genomics experiment in mouse and human embryonic development. The left panel shows dimensionality reduction resulting from single-cell measurements at three developmental stages for each species (A, B and C). Note that the developmental stages are not necessarily equivalent but can be computationally matched by exploiting the single-cell expression profiles. The right panel illustrates alignment of human and mouse developmental trajectories using a DTW alignment algorithm[91].

relationship between the molecular layers is complex. A good example is early embryonic development, where gene body DNA methylation and/or chromatin accessibility are not good predictors of gene expression[52]. Thus, the question that arises is how to perform data integration when epigenetic and transcriptomic measurements cannot be associated with a common genomic locus.

Some methods have attempted to solve the diagonal integration problem by reconstructing technology-invariant integrated latent spaces. The first of these methods was MATCHER[77], a Gaussian process latent variable model that was successful at reconstructing a differentiation trajectory in embryonic stem cells by exploiting covariation between transcriptomic and epigenetic measurements. However, this method relies on the strong assumption that biological variation is defined by a unidimensional axis of variation. More recent methods, including MMD-MA[78], SCIM[67] and UnionCom[68], have generalized MATCHER to account for complex multivariate trajectories.

Diagonal integration is arguably the hardest of the aforementioned data integration tasks and faces important challenges on how to define the data input and validate and interpret the model output. One of the reasons is that understanding of the properties of latent biological manifolds is incomplete. Even if RNA expression and chromatin accessibility are (partially) correlated, there is no guarantee that their latent manifolds can in fact be aligned. Matched multimodal assays can be used as gold standard datasets to address this question and benchmark integration strategies[76]. We envisage that diagonal integration will receive increasing attention in the near future and that some of these questions will be explored and perhaps answered.

Mosaic integration. Despite the maturation of single-cell multimodal technologies, simultaneously capturing multiple molecular layers from the same cell in an efficient and scalable manner is still a challenging task. A more feasible and common experimental design is to profile individual data modalities on different populations of cells from the same biological sample (Fig. 3). This leads to an incomplete dataset where entire data matrices are missing, a scenario that is not handled by most data integration strategies. Nevertheless, this missing value structure will become ubiquitous as current transcriptomic atlases are extended with other molecular layers. Examples of biological systems where such data already exist include human white blood cells, mouse gastrulation and brain development (Table 2).

An instance of mosaic integration could consist of the imputation of missing molecular layers with the goal of building a self-consistent multimodal dataset. Nonetheless, in this setting, there is no simple choice of anchor, as some pairs of matrices are anchored by the cells whereas other pairs of matrices are anchored by the features and some pairs of matrices do not share anchors at all. An intuitive and simple approach could be to perform successive but independent rounds of horizontal, vertical or diagonal data integration by selectively exploiting cells and features as anchors (Fig. 3), but the impact of feature selection and the order of integration complicate this approach. A more comprehensive computational solution would be to use multitask learning models that are capable of combining all three types of integration simultaneously while propagating uncertainty in the imputation estimates.

Transfer learning. Large single-cell atlases are becoming increasingly available for a diverse collection of tissues, organs and species, thus providing valuable references that can aid the analysis of succeeding datasets. For example, instead of performing unsupervised clustering and cell type annotation de novo, one can leverage a matching reference dataset to generate a joint embedding and transfer the cell type labels from the reference atlas to the query. As described above, this strategy is frequently addressed as a horizontal integration task by multitask learning methods that exploit a common feature space to learn a joint representation. However, multitask learning methods load the full reference dataset together with the query dataset, which can require large computational resources as existing single-cell atlases grow in size. An alternative strategy is to extract a compressed representation of the reference dataset and use this to inform analysis of the query dataset. This strategy, known as transfer learning or knowledge transfer, has revolutionized fields such as computer vision[79] and natural language processing[80], where neural network models are trained using extensive amounts of data and then repurposed as a starting point to train new related models[79,80]. Note that, in this case, the anchor between experiments is the

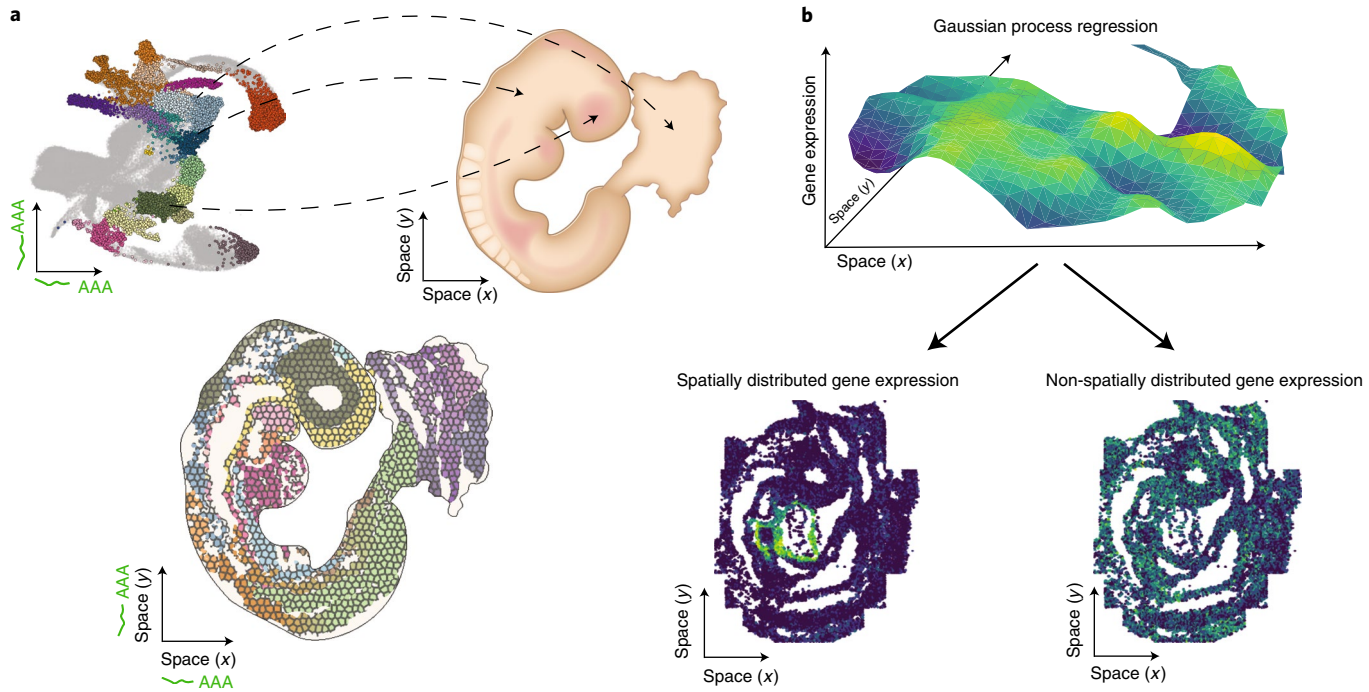

Fig. 5 | Data integration of spatially resolved transcriptomics. a, Example of horizontal data integration: mapping cell types from a dissociated sample (latent representation shown on the left) to a spatial distribution (embryo diagram shown to the right). The integrated dataset shown at the bottom has cell type assignments overlayed onto spatial coordinates. **b**, Example of vertical integration: detection of genes that display (nonlinear) spatial expression variation across a tissue.

feature space and the task can be classified as horizontal integration. However, the fundamental difference is that, instead of treating each experiment as an independent observation, transfer learning implies a hierarchical relationship where a reference dataset is used to inform a second query dataset. Transfer learning strategies have been devised for single-cell genomics for tasks such as denoising[81], cell type classification[82] and shared embedding[83]. With the increasing availability of single-cell atlases, we envision that transfer learning will massively facilitate data reuse and will be at the core of many future single-cell data analysis pipelines.

Integration of molecular measurements with physical dimensions

Until recently, most integrative methods for single-cell genomics did not explicitly consider physical dimensions such as time and space. However, several experimental protocols have been developed that allow molecular measurements to be made, at the single-cell level, while maintaining some information about a cell's location in the tissue of interest[84–87]. In this section, we consider principles and challenges of data integration in the context of time-resolved and spatially resolved multimodal data.

Integration of time-resolved data. When integrating samples from time course experiments, most horizontal integration methods treat each sample independently, ignoring the time component. This is a reasonable strategy if the populations of cells are sampled from a similar time point or from a stationary state. However, accounting for temporal variation is essential in some scenarios, particularly for the study of dynamic biological processes such as embryonic development. As an example, Pijuan-Sala et al. constructed a mouse gastrulation atlas by profiling a total of 116,312 single-cell transcriptomes from embryonic day (E) 6.5 to E8.5 (ref. [88]). Instead of integrating all embryos simultaneously, the researchers performed horizontal integration with MNN[21] in a bottom–up fashion, starting with the closely related E6.5 samples followed by incorporation of later time points in a sequential manner.

In this example, time-aware data integration was possible because the time points for the different embryos were known and comparable. However, there are cases where the time correspondence between samples is not known, such as in evolutionary single-cell genomics, an emerging field aimed at mapping single-cell variation across different species[89,90].

From a computational perspective, a challenge in this task is how to align differentiation trajectories when the molecular clock ticks at varying rates across different species. To address this, computational strategies such as dynamic time warping (DTW; originally developed for speech recognition[91]) have been successfully adapted and applied to align time series (Fig. 4). In the context of single-cell genomics, DTW permits the mapping of differentiation trajectories[92–95]. For example, Kanton et al. generated a gene expression atlas of chimpanzee and human cerebral organoid development from iPSCs and used DTW to align the differentiation trajectories[96]. Notably, they found that their differentiation trajectories did not match in late differentiation stages, thus revealing human-specific gene expression programs, some of which persist into adulthood. Evolutionary comparisons come with fundamental challenges of how to define the anchors. For instance, when performing horizontal integration, which features should be used as anchors? In the case of RNA expression measurements, the simplest choice is to restrict the feature space to genes with high sequence similarity (that is, homologs). Yet, there is no guarantee that the function and transcriptional profiles of homolog genes are preserved when mapping across large evolutionary distances[97]. An alternative approach could be to summarize gene expression profiles over gene sets or pathways, which might be better preserved by natural selection and thus provide a more robust anchor choice[98].

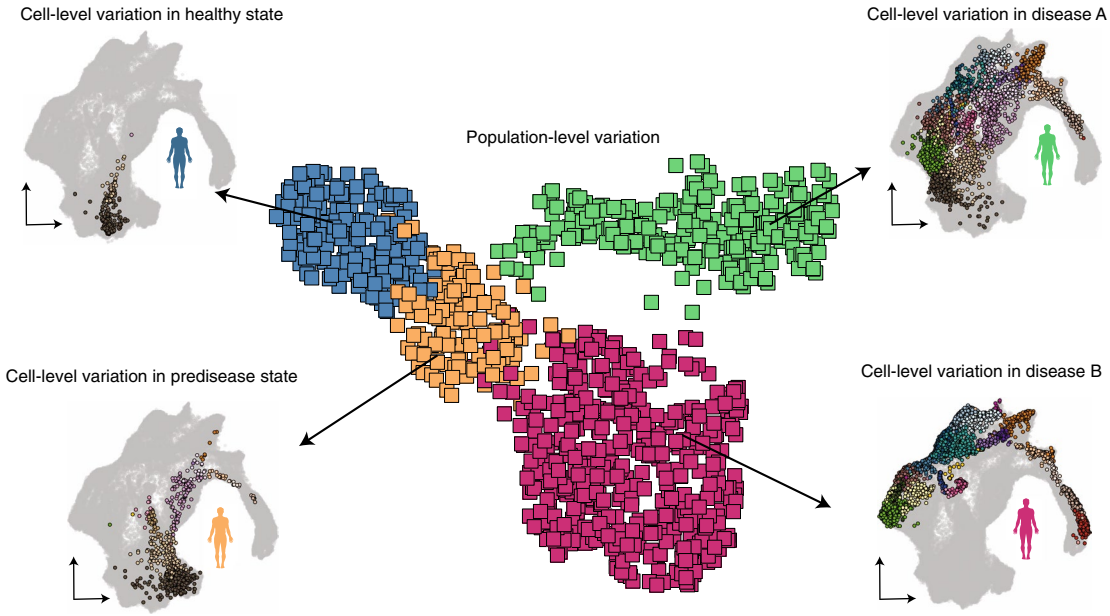

Fig. 6 | Exploiting molecular variation at single-cell resolution to construct population-level maps of human phenotypic variation. The central plot is a schematic of a latent representation of human population variation (each square corresponds to a human individual, and individuals are the anchor between the two representations), inferred from properties of single-cell profiles. Individuals are colored by phenotypic state (blue, healthy; orange, predisease; green, disease A; magenta, disease B). The position of each individual in the latent representation captures the cell type distribution as quantified using single-cell genomics in disease-relevant tissue. For each phenotypic state, we illustrate the single-cell profile of an individual by highlighting the location of their cells on a latent manifold that contains all cells from all donors (in gray). Note that each phenotypic state is associated with a different distribution of cell types.

Integration of spatially resolved single-cell data. The use of single-cell omics has been instrumental in improving understanding of cellular biology. Yet, dissociation of cells from their native spatial context results in fundamental limitations in our ability to understand the interplay between intrinsic and extrinsic factors that underlie cellular communication and organ function. Bridging the gap between single-cell molecular readouts and tissue-level variation from histopathological and microscopy assays has been challenging because of the lack of a simple anchor choice between these two layers. Recent technological advances in multiplexed imaging and sequencing[84–87] permit the quantification of a large number of genes in individual cells in situ and hold promise for making this type of multiscale modeling a forthcoming reality. However, from a computational perspective, the integration of high-resolution molecular information and spatial information raises challenges, and analytical tools are just beginning to emerge.

Most methods introduced for the analysis of spatially resolved data can be framed using the anchor framework. An example of a horizontal integration task is the definition of resident cell type identities in situ by using information from existing dissociated datasets (Fig. 5a). Although cell type assignment can be performed de novo, this strategy is limited by the resolution of the experiment, owing to the fact that fixed pixel locations can overlap multiple cells. Because of this, gene expression measurements at a given pixel can be the result of a mixture of cell types. Thus, a natural approach to associate pixels with cell type identities is to use multitask or transfer learning to exploit the cell type information contained in (dissociated) reference atlases[99–103]. As an example, SpiceMix combines NMF with a hidden Markov random field to jointly model spatial location and latent factors of cell identity to transfer cell type information from a dissociated reference dataset to a spatially resolved dataset[99].

An example of a vertical integration task is the detection of genes that display spatial expression variation across a tissue (Fig. 5b), a task that is only possible when RNA expression and spatial location are available for the same cell (or pixel). Although a simple Pearson correlation coefficient could be applied for this task, recognizing the complex gene expression patterns that some tissues display in space requires the use of nonlinear methodologies. This is the aim of SpatialDE[104] and spatial variance component analysis (SVCA)[105], both of which build upon Gaussian process regression, a class of models commonly used in geostatistics. SVCA decomposes gene expression variation into intrinsic effects (that is, cell cycle), environmental effects and, most importantly, an explicit cell–cell interaction component.

Finally, there are some largely unresolved data integration tasks where the definitions of horizontal, vertical and diagonal integration tasks will have to be revisited. One example is the generation of spatially resolved atlases where tissue samples are derived from multiple donors. The optimal strategy for designing an anchor that accounts for anatomical differences between individuals is unclear; such an anchor should incorporate biological parameters such as age, sex and ancestry, while remaining sensitive to clinically relevant variation[106]. Thus, there is a need for computational methods that are able to map samples onto a reference while at the same time acknowledging the variation between samples[83].

Multiscale integration for personalized medicine

The enormous efforts to build atlases of single-cell variation in the context of human health will prove essential to understanding disease heterogeneity at the cellular level. However, querying these datasets will require a new set of methodologies for data integration that connect variation at the single-cell level with biomedical traits in human populations[107].

Intuitively, the goal of such multiscale methodologies is to extract information from the cellular representation that explains phenotypic variability at the individual human level. If clinical data are available as a predictor, this task can be formulated as a supervised learning problem, where the covariates correspond to features

extracted from the cellular representation. Such features can be extracted using model-based approaches or manually defined by expert knowledge. The latter strategy was adopted for generation of a tumor immune atlas by integrating scRNA-seq datasets from 217 patients and 13 different cancer types[108]. By calculating cell type proportions from the single-cell representation, the investigators were able to provide a cellular basis for patient stratification by immune cell composition. More generally, we envisage that the extraction of interpretable and predictive features from single-cell representations using statistical models will be an area of active research. If a predictor is only available as a multidimensional variable, the task becomes substantially more challenging, as two multidimensional datasets have to be tied together. In this setting, the anchor between the two representations is the human individuals in the study, thus defining a new type of integration problem where cells and genes cannot be used as anchors (Fig. 6).

This class of multiscale modeling strategies will have a key role in enabling the application of single-cell omics in personalized medicine, as these approaches will enable linkage of attributes specifically associated with a donor's cell ecosystem to medically relevant traits[108,109]. The ultimate goal of such an integration task is to provide tools that will facilitate an understanding of the etiology and progression of diseases at single-cell resolution, predict pathological phenotypes before their onset and allow intervention in the most personalized fashion possible[110,111].

Concluding remarks

In this Review, we introduce a set of concepts to contextualize single-cell data integration techniques and discuss alternative choices of anchors for linking different datasets. We review established principles, limitations and diagnostics of data integration strategies and highlight parallels between approaches for genetic analysis of single-cell traits and inference of regulatory dependencies between molecular layers. Finally, we extend the basic data integration concepts to more challenging future applications, including the integration of single-cell omics data with physical dimensions and the construction of reference atlases of human variation for personalized medicine.

Received: 6 November 2020; Accepted: 16 March 2021;
Published online: 03 May 2021

References

1. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25**, 1499–1507 (2015).
2. Peng, G., Cui, G., Ke, J. & Jing, N. Using single-cell and spatial transcriptomes to understand stem cell lineage specification during early embryo development. *Annu. Rev. Genomics Hum. Genet.* **21**, 163–181 (2020).
3. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
4. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* https://doi.org/10.1016/j.cell.2020.09.056 (2020).
5. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
6. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
7. Swanson, E. et al. TEA-seq: a trimodal assay for integrated single cell measurement of transcription, epitopes, and chromatin accessibility. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.04.283887 (2020).
8. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
9. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).
10. Chappell, L., Russell, A. J. C. & Voet, T. Single-cell (multi) omics technologies. Annu. Rev. *Genomics Hum. Genet.* **19**, 15–41 (2018).
11. Hao, Y., Hao, S., Andersen-Nissen, E. & Mauck, W. M. Integrated analysis of multimodal single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.12.335331 (2020).
12. Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **22**, 20–29 (2021).
13. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.* **38**, 1007–1022 (2020).
14. Colomé-Tatché, M. & Theis, F. J. Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.* **7**, 54–59 (2018).
15. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
16. Cheow, L. F. et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* **13**, 833–836 (2016).
17. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0290-0 (2019).
18. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
19. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
20. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
21. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
22. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
23. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
24. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
25. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
26. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
27. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
28. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
29. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
30. Johansen, N. & Quon, G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* **20**, 166 (2019).
31. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.22.111161 (2020).
32. Schadt, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
33. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
34. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
35. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).
36. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014).
37. Hu, Y. et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 88 (2016).
38. Liu, L. et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 470 (2019).
39. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
40. Packer, J. & Trapnell, C. Single-cell multi-omics: an engine for new quantitative models of gene regulation. *Trends Genet.* **34**, 653–665 (2018).
41. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
42. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
43. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
44. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
45. Henderson, C. R. *Applications of Linear Models in Animal Breeding* Univ. Guelph (1984).
46. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
47. Furlotte, N. A., Kang, H. M., Ye, C. & Eskin, E. Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics* **27**, i288–i294 (2011).
48. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

49. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
50. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific *cis*-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
51. Cuomo, A. S. E. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
52. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
53. Wills, Q. F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).
54. Sarkar, A. K. et al. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* **15**, e1008045 (2019).
55. van der Wijst, M. et al. The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
56. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
57. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
58. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
59. Rubin, A. J. et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* **176**, 361–376 (2019).
60. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
61. Schraivogel, D. et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
62. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 1516 (2019).
63. Mimitou, E. P. et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
64. Argelaguet, R. et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
65. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
66. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
67. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
68. Xu, C., Tao, D. & Xu, C. A survey on multi-view learning. Preprint at https://arxiv.org/abs/1304.5634 (2013).
69. Argelaguet, R. et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
70. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7**, 523–542 (2013).
71. Singh, A. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
72. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162 (2014).
73. Klami, A., Virtanen, S., Leppäaho, E. & Kaski, S. Group factor analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **26**, 2136–2147 (2015).
74. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
75. Luo, C. et al. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. Preprint at *bioRxiv* https://doi.org/10.1101/2019.12.11.873398 (2019).
76. Wang, C. et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 198 (2020).
77. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138 (2017).
78. Liu, J., Huang, Y., Singh, R., Vert, J.-P. & Noble, W. S. Jointly embedding multiple single-cell omics measurements. Preprint at *bioRxiv* https://doi.org/10.1101/644310 (2019).
79. Zheng, H. et al. Cross-domain fault diagnosis using knowledge transfer strategy: a review. *IEEE Access* **7**, 129260–129290 (2019).
80. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* 15–18 https://doi.org/10.18653/v1/n19-5004 (2019).
81. Wang, J. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
82. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe—classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE* **13**, e0205499 (2018).
83. Lotfollahi, M., Naghipourfar, M., Luecken, M. D. & Khajavi, M. Query to reference single-cell integration with transfer learning. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.16.205997 (2020).
84. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
85. Eng, C.-H. L., Shah, S., Thomassie, J. & Cai, L. Profiling the transcriptome with RNA SPOTs. *Nat. Methods* **14**, 1153–1155 (2017).
86. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
87. Giacomello, S. et al. Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* **3**, 17061 (2017).
88. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
89. Marioni, J. C. & Arendt, D. How single-cell genomics is changing evolutionary and developmental biology. *Annu. Rev. Cell Dev. Biol.* **33**, 537–553 (2017).
90. Shafer, M. E. R. Cross-species analysis of single-cell transcriptomic data. *Front. Cell Dev. Biol.* **7**, 175 (2019).
91. Vintsyuk, T. K. Speech discrimination by dynamic programming. *Cybernetics* **4**, 52–57 (1972).
92. Cacchiarelli, D. et al. Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Syst.* **7**, 258–268 (2018).
93. Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15**, 267–270 (2018).
94. Do, V. H. et al. Dynamic pseudo-time warping of complex single-cell trajectories. Preprint at *bioRxiv* https://doi.org/10.1101/522672 (2019).
95. Velten, B., Braunger, J. M., Arnol, D., Argelaguet, R. & Stegle, O. Identifying temporal and spatial patterns of variation from multi-modal data using MEFISTO. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.03.366674 (2020).
96. Kanton, S. et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
97. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).
98. Arendt, D. et al. The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
99. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* gkab043 (2021).
100. Chidester, B., Zhou, T. & Ma, J. SpiceMix: integrative single-cell spatial modeling for inferring cell identity. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.29.383067 (2021).
101. Kleshchevnikov, V. et al. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.15.378125 (2020).
102. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
103. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00830-w (2021).
104. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
105. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling cell–cell interactions from spatial molecular data with spatial variance component analysis. *Cell Rep.* **29**, 202–211 (2019).
106. Rood, J. E. et al. Toward a common coordinate framework for the human body. *Cell* **179**, 1455–1467 (2019).
107. Camp, J. G., Platt, R. & Treutlein, B. Mapping human cell phenotypes to genotypes with single-cell genomics. *Science* **365**, 1401–1405 (2019).
108. Nieto, P., Elosua-Bayes, M. M., Trincado, J. L. & Marchese, D. A single-cell tumor immune atlas for precision oncology. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.26.354829 (2020).
109. Keener, A. B. Single-cell sequencing edges into clinical trials. *Nat. Med.* **25**, 1322–1326 (2019).
110. Rajewsky, N. et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* https://doi.org/10.1038/s41586-020-2715-9 (2020).

111. Shalek, A. K. & Benson, M. Single-cell analyses to tailor treatments. *Sci. Transl. Med.* **9**, eaan4730 (2017).
112. Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
113. Meng, C. et al. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
114. Jin, S., Zhang, L. & Nie, Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 25 (2020).
115. Stark, S. G. et al. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).
116. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).
117. Duren, Z. et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl Acad. Sci. USA* **115**, 7723–7728 (2018).
118. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
119. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
120. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
121. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
122. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
123. Wang, A. et al. Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. *eLife* **9**, e62522 (2020).
124. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394 (2016).
125. Lawlor, M. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).
126. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
127. 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 (2016).
128. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
129. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
130. Bravo González-Blas, C. et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol. Syst. Biol.* **16**, e9438 (2020).
131. Pijuan-Sala, B. et al. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat. Cell Biol.* **22**, 487–497 (2020).
132. Preisel, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
133. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
134. Lee, D.-S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16**, 999–1006 (2019).
135. Johnstone, I. M. & Titterington, D. M. Statistical challenges of high-dimensional datal. *Philos. Trans. A Math. Phys. Eng. Sci.* **367**, 4237–4253 (2009).
136. Guo, F. et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).
137. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).
138. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
139. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
140. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
141. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
142. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

Acknowledgements

R.A. and A.S.E.C. are supported by a PhD fellowship from the EMBL International PhD Programme. O.S. is supported by core funding from EMBL and the DKFZ, as well as the BMBF, the Volkswagen Foundation and the European Union (810296). J.C.M. acknowledges core funding from EMBL and core support from Cancer Research UK (C9545/A29580).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to R.A., A.S.E.C., O.S. or J.C.M.

Peer review information *Nature Biotechnology* thanks Carl Herrmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature America, Inc. 2021