SINGLE-CELL OMICS

Integrative single-cell analysis

Tim Stuart[1] *and Rahul Satija*[1,2]*

Abstract | The recent maturation of single-cell RNA sequencing (scRNA-seq) technologies has coincided with transformative new methods to profile genetic, epigenetic, spatial, proteomic and lineage information in individual cells. This provides unique opportunities, alongside computational challenges, for integrative methods that can jointly learn across multiple types of data. Integrated analysis can discover relationships across cellular modalities, learn a holistic representation of the cell state, and enable the pooling of data sets produced across individuals and technologies. In this Review, we discuss the recent advances in the collection and integration of different data types at single-cell resolution with a focus on the integration of gene expression data with other types of single-cell measurement.

Single-cell RNA sequencing
(scRNA-seq). Sequencing of cDNAs derived from RNA molecules (usually polyadenylated mRNAs) from a single cell. It is typically performed for many hundreds to thousands of cells in a single experiment.

Multimodal
Data of multiple types, for example, of RNA and protein.

[1]*New York Genome Center, New York, NY, USA.*

[2]*Center for Genomics and Systems Biology, New York University, New York, NY, USA.*

**e-mail: rsatija@nygenome.org*

https://doi.org/10.1038/s41576-019-0093-7

Recent advances in molecular biology, microfluidics and nanotechnology have given rise to a multitude of single-cell sequencing technologies (FIG. 1). Initial methods have focused on measurements of a single modality (for example, DNA sequence, RNA expression or chromatin accessibility). Although these technologies have yielded transformative insights into cellular diversity and development, this segregation is driven by methodological convenience and limits the ability to derive a deep understanding of the relationships between biomolecules in single cells. Understanding these interactions is key to deriving a deep understanding of the cellular state and remains a challenge for the field of single-cell analysis. Moreover, as the scale and availability of data sets rapidly grow, new computational methods are needed for normalization and joint analysis across samples, even in the presence of significant batch effects or interindividual variation.

Single-cell RNA sequencing (scRNA-seq) is one of the most widely used single-cell sequencing approaches, with a range of technologies for sensitive, highly multiplexed or combinatorially barcoded profiling[1–8]. These advances have accompanied a variety of complementary single-cell genomic, epigenomic and proteomic profiling technologies, including methods for single-cell measurements of genome sequence[9,10], chromatin accessibility[11–15], DNA methylation[11,16–19], cell surface proteins[20,21], small RNAs[22], histone modifications[23,24] and chromosomal conformation[25,26]. Furthermore, recent efforts have pioneered methods to accurately record spatial or lineage information in single-cell studies[27–35] (FIG. 1; TABLE 1).

An idealized experimental workflow would observe all aspects of the cell, including a full history of its molecular states, spatial positions and environmental interactions. Although outside the bounds of current technology, multimodal technologies and integrative computational methods enable us to move closer to this aspirational and exciting goal. In this Review, we describe the currently available methods for single-cell transcriptomics, genomics, epigenomics and proteomics with an emphasis on those methods that provide multimodal data or data that can be integrated into a multimodal analysis. We focus on the analysis of scRNA-seq data in conjunction with other data types, as these are currently the most commonly used and well-established methods. In particular, we discuss those methods capable of integrating data from the same individual cell wherever possible. We discuss these methods and their challenges in depth, as well as their potential applications and future directions.

Multimodal single-cell measurements

Single-cell molecular profiling technologies initially focused on the development of methods capable of accurately detecting a single aspect of the cell state, first with simple semiquantitative readouts[36] and later utilizing high-throughput DNA sequencing[37]. More recently, there has been considerable interest in the simultaneous profiling of multiple types of molecule within a single cell (multimodal profiling) to build a much more comprehensive molecular view of the cell. Often, these methods couple scRNA-seq with the measurement of another cellular characteristic, such as DNA sequence, protein abundance or epigenomic state. Multimodal data can be obtained from single cells using four broad strategies (FIG. 2): first, the use of an initial non-destructive assay before sequencing; second, the separation of different cellular fractions for parallel experimental workflows; third, the experimental conversion of multimodal data into a common molecular format to enable the simultaneous detection of multiple data types via a common methodology, such as DNA sequencing; and fourth, the analysis of different data types encoded in nucleotide sequences, such as RNA abundance and sequence polymorphisms.

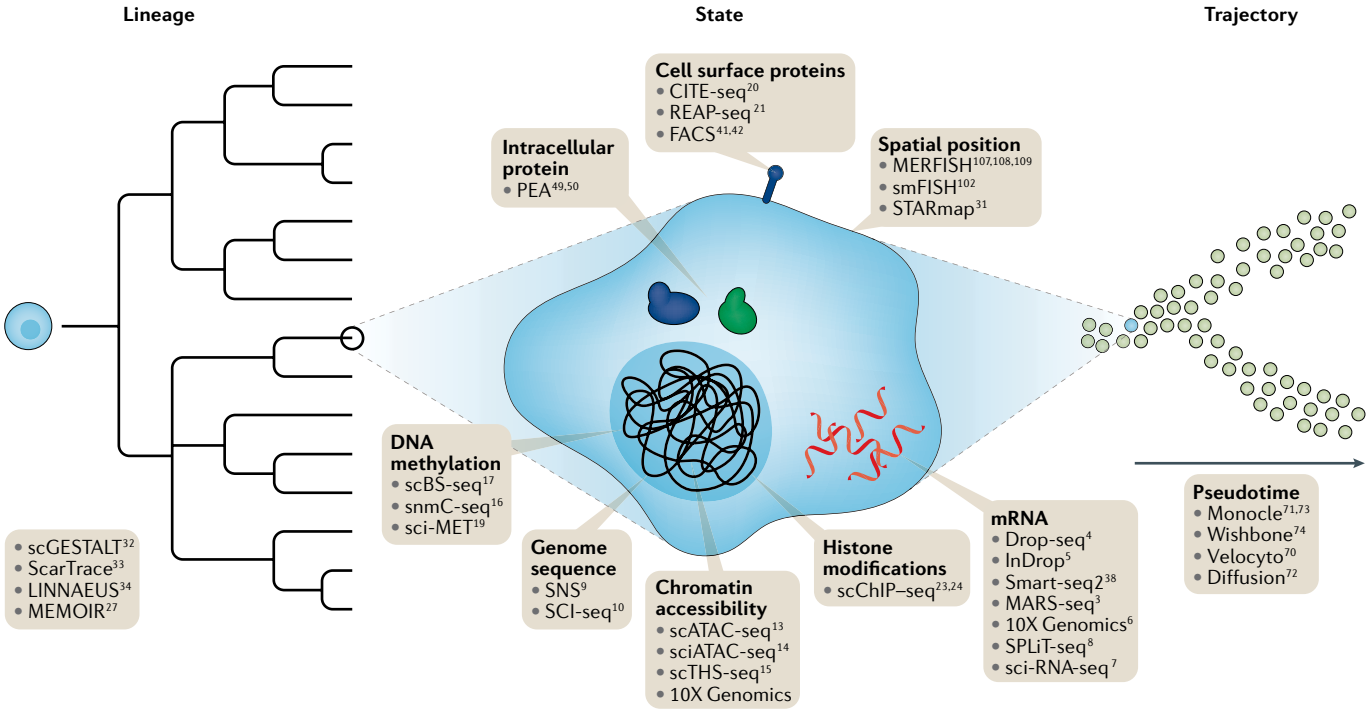


Fig. 1 | **Multimodal and integrative methods for single-cell analyses.** An overview of the current methods for single-cell data integration is shown. A wide variety of single-cell methods have now been developed to measure a broad range of cellular parameters. These methods can be divided into those that determine the current state of the cell, those that determine the cell lineage, and computational methods that order cells along a pseudotemporal trajectory. CITE-seq, cellular indexing of transcriptomes and epitopes by sequencing; FACS, fluorescence-activated cell sorting; LINNAEUS, lineage tracing by nuclease-activated editing of ubiquitous sequences; MARS-seq, massively parallel RNA single-cell sequencing; MEMOIR, memory by engineered mutagenesis with optical in situ readout; MERFISH, multiplexed error-robust fluorescence in situ hybridization; PEA, proximity extension assay; REAP-seq, RNA expression and protein sequencing assay; scATAC-seq, single-cell assay for transposase-accessible chromatin using sequencing; scBS-seq, single-cell bisulfite sequencing; scChIP–seq, single-cell chromatin immunoprecipitation followed by sequencing; scGESTALT, single-cell genome editing of synthetic target arrays for lineage tracing; sci-MET, single-cell combinatorial indexing for methylation analysis; sci-RNA-seq, single-cell combinatorial indexing RNA sequencing; SCI-seq, single-cell combinatorial indexed sequencing; sciATAC-seq, single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing; scTHS-seq, single-cell transposome hypersensitivity site sequencing; smFISH, single-molecule fluorescence in situ hybridization; snmC-seq, single-nucleus methylcytosine sequencing; SNS, single-nucleus sequencing; SPLiT-seq, split-pool ligation-based transcriptome sequencing; STARmap, spatially resolved transcript amplicon readout mapping.

Gathering cytometric information before a destructive assay. An initial and elegant solution for multimodal profiling involves the application of non-destructive cytometric measurements before the application of a destructive single-cell assay. As multiple scRNA-seq workflows utilize fluorescence-activated cell sorting (FACS) to deposit individual cells into microtitre plates[2,3,38], it is a natural extension to combine this single-cell isolation with index sorting to gather additional cytometric data about the cells before sequencing (FIG. 2a). Whereas early studies combined measurements of the cell cycle and semiquantitative measurements of mRNAs from the same cell[39], this approach has been particularly fruitful in immunology and haematology, where well-defined cell surface markers can be used to classify functional cell types and states[40,41] or to enrich for rare cells in heterogeneous populations. Paul et al.[41] and Nestorowa et al.[42] applied this workflow to profile early murine haematopoietic progenitors, revealing the immunophenotypes of transcriptionally defined cell-type clusters. Similarly, Wilson et al.[40] utilized FACS isolation of rare haematopoietic stem cells (HSCs) followed by scRNA-seq and functional assays to identify cell surface markers associated with cells that are able to consistently self-renew[40]. New methods that utilize arrays of picolitre wells have the potential to dramatically increase the scale of these experiments while retaining the ability to gather cytometric data before single-cell assays[43]. However, cytometric methods are fundamentally limited in the number of parameters they can measure for each cell, as they are limited by spectral overlap between the fluorescent reporters.

Separation of cellular components. Alternative approaches are required in order to measure aspects of the cell that cannot easily be read out through cellular fluorescence. This requirement is especially relevant for experiments aiming to simultaneously measure mRNAs alongside genomic DNA or intracellular protein in the same cell. In these cases, the physical separation or selective tagging of different cellular fractions from single cells presents an attractive solution (FIG. 2b). Several groups

Index sorting
Fluorescence-activated sorting of cells into known plate locations.

Table 1 | **Current experimental methods for unimodal and multimodal single-cell measurements**

Data types	Method name	Feature throughput	Cell throughput	Refs
Unimodal				
mRNA	Drop-seq	Whole transcriptome	1,000–10,000	4
	InDrop	Whole transcriptome	1,000–10,000	5
	10X Genomics	Whole transcriptome	1,000–10,000	6
	Smart-seq2	Whole transcriptome	100–300	38
	MARS-seq	Whole transcriptome	100–300	3
	CEL-seq	Whole transcriptome	100–300	1
	SPLiT-seq	Whole transcriptome	≥50,000	8
	sci-RNA-seq	Whole transcriptome	≥50,000	7
Genome sequence	SNS	Whole genome	10–100	9
	SCI-seq	Whole genome	10,000–20,000	10
Chromatin accessibility	scATAC-seq	Whole genome	1,000–2,000	13
	sciATAC-seq	Whole genome	10,000–20,000	14
	scTHS-seq	Whole genome	10,000–20,000	15
DNA methylation	scBS-seq	Whole genome	5–20	17
	snmC-seq	Whole genome	1,000–5,000	16
	sci-MET	Whole genome	1,000–5,000	19
	scRRBS	Reduced representation genome	1–10	18
Histone modifications	scChIP–seq	Whole genome + single modification	1,000–10,000	24
Chromosome conformation	scHi-C-seq	Whole genome	1–10	26
Multimodal				
Histone modifications + spatial	NA	Single locus + single modification	10–100	23
mRNA + lineage	scGESTALT	Whole transcriptome	1,000–10,000	32
	ScarTrace	Whole transcriptome	1,000–10,000	33
	LINNAEUS	Whole transcriptome	1,000–10,000	34
Lineage + spatial	MEMOIR	NA	10–100	27
mRNA + spatial	osmFISH	10–50 RNAs	1,000–5,000	35
	STARmap	20–1,000 RNAs	100–30,000	31
	MERFISH	100–1,000 RNAs	100–40,000	108
	seqFish	125–250 RNAs	100–20,000	29
mRNA + cell surface protein	CITE-seq	Whole transcriptome + proteins	1,000–10,000	20
	REAP-seq	Whole transcriptome + proteins	1,000–10,000	21
mRNA + chromatin accessibility	sci-CAR	Whole transcriptome + whole genome	1,000–20,000	48
mRNA + DNA methylation	scM&T-seq	Whole genome	50–100	46
mRNA + genomic DNA	G&T-seq	Whole genome + whole transcriptome	50–200	44
mRNA + intracellular protein	NA	96 mRNAs + 38 proteins	50–100	50
		82 mRNAs + 75 proteins	50–200	49
DNA methylation + chromatin accessibility	scNOMe-seq	Whole genome	10–20	11

CEL-seq, cell expression by linear amplification and sequencing; CITE-seq, cellular indexing of transcriptomes and epitopes by sequencing; G&T-seq, genome and transcriptome sequencing; LINNAEUS, lineage tracing by nuclease-activated editing of ubiquitous sequences; MARS-seq, massively parallel RNA single-cell sequencing; MEMOIR, memory by engineered mutagenesis with optical in situ readout; MERFISH, multiplexed error-robust fluorescence in situ hybridization; osmFISH, cyclic single-molecule fluorescence in situ hybridization; REAP-seq, RNA expression and protein sequencing assay; scATAC-seq, single-cell assay for transposase-accessible chromatin using sequencing; scBS-seq, single-cell bisulfite sequencing; scChIP–seq, single-cell chromatin immunoprecipitation followed by sequencing; scGESTALT, single-cell genome editing of synthetic target arrays for lineage tracing; scHi-C-seq, a single-cell Hi-C method for chromosome conformation; sciATAC-seq, single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing; sci-CAR, single-cell combinatorial indexing chromatin accessibility and mRNA sequencing; sci-MET, single-cell combinatorial indexing for methylation analysis; sci-RNA-seq, single-cell combinatorial indexing RNA sequencing; SCI-seq, single-cell combinatorial indexed sequencing; scM&T-seq, single-cell methylome and transcriptome sequencing; scNOMe-seq, single-cell nucleosome occupancy and methylome sequencing; scRRBS, single-cell reduced representation bisulfite sequencing; scTHS-seq, single-cell transposome hypersensitivity site sequencing; seqFISH, sequential fluorescence in situ hybridization; snmC-seq, single-nucleus methylcytosine sequencing; SNS, single-nucleus sequencing; SPLiT-seq, split-pool ligation-based transcriptome sequencing; STARmap, spatially resolved transcript amplicon readout mapping.

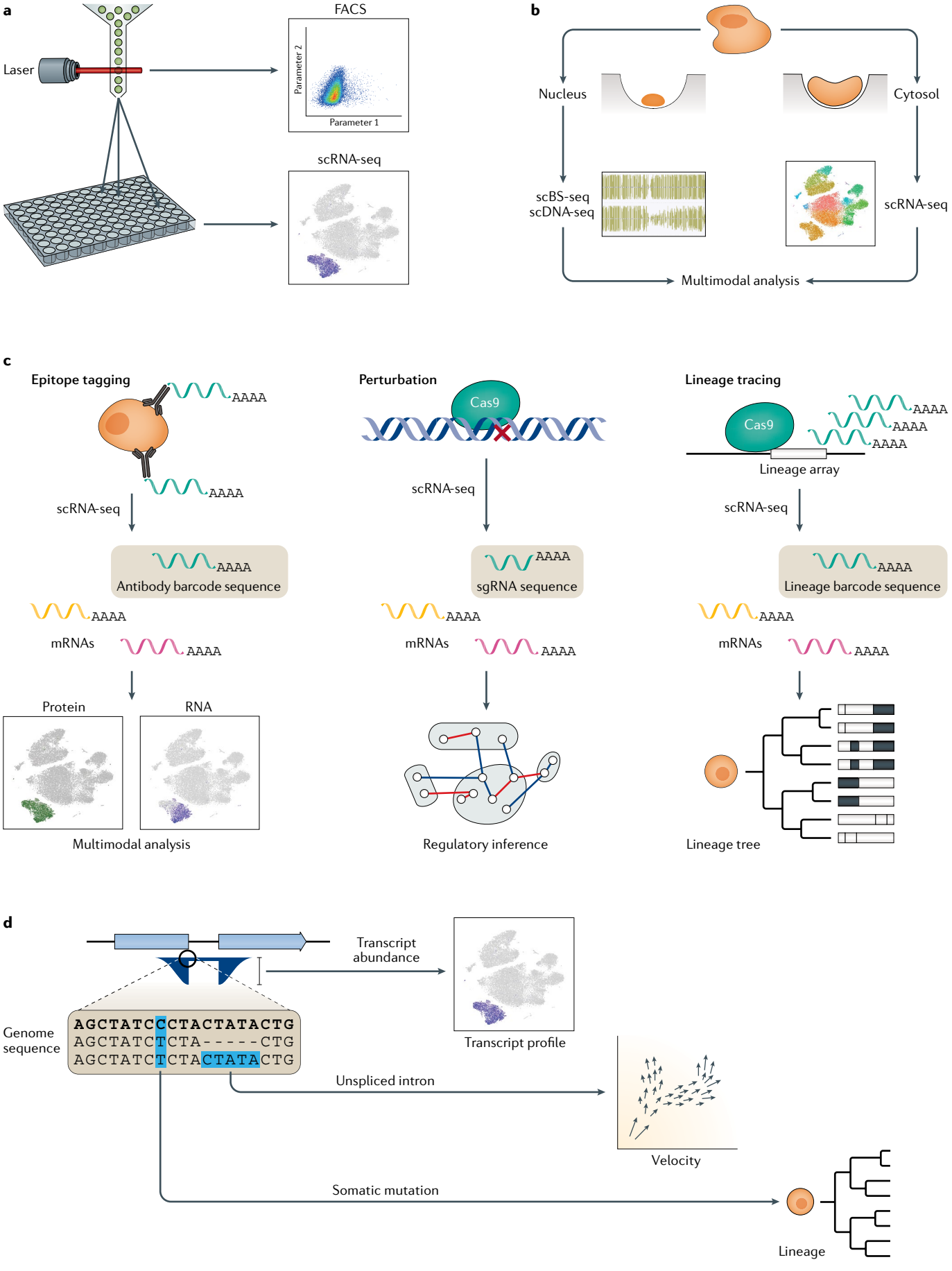

a
Laser
FACS
Parameter 2
Parameter 1
scRNA-seq
b
Nucleus
Cytosol
scBS-seq
scDNA-seq
scRNA-seq
Multimodal analysis
c
Epitope tagging
AAAA
AAAA
scRNA-seq
AAAA
Antibody barcode sequence
AAAA
mRNAs
AAAA
Protein
RNA
Multimodal analysis
Perturbation
Cas9
scRNA-seq
AAAA
sgRNA sequence
AAAA
mRNAs
AAAA
Regulatory inference
Lineage tracing
Cas9
AAAA
AAAA
AAAA
Lineage array
scRNA-seq
AAAA
Lineage barcode sequence
AAAA
mRNAs
AAAA
Lineage tree
d
Transcript abundance
Transcript profile
Genome sequence
AGCTATCCCTACTATACTG
AGCTATCTCTA-----CTG
AGCTATCTCTACTATACTG
Unspliced intron
Velocity
Somatic mutation
Lineage



Fig. 2 | **Experimental methods for performing single-cell multimodal measurements.** **a** | Gathering cytometric single-cell measurements using multiparameter fluorescence-activated cell sorting (FACS) before single-cell RNA sequencing (scRNA-seq) can allow fluorescence-based measurements of protein levels to be later linked to cellular transcriptomes; hence, RNA and protein levels can be analysed jointly in the same cell. **b** | A lyse-and-split strategy can allow parallel workflows to be performed on different cellular fractions. For example, the cytosol can be physically separated from the nucleus to allow measurement of cytosolic mRNAs through scRNA-seq and measurements of the genomic DNA using whole-genome sequencing or bisulfite sequencing to gather complementary data on the cell genotype or methylome, respectively. **c** | Innovative barcoding strategies can enable standard scRNA-seq methods to capture important additional information to enhance the analysis of cell transcriptomes. Cell surface protein abundance can be captured using standard scRNA-seq methods by conjugating polyadenylated antibody barcodes to antibodies targeting cell surface proteins[20,21] (left panel). These antibody barcode sequences can be captured alongside polyadenylated mRNAs and decoded to provide an estimate of protein levels for each cell. Allelic information can be encoded by the single-guide RNA (sgRNA) sequence used to guide Cas9 in pooled genetic screens, allowing gene knockout information to be associated with single-cell transcriptional profiles (middle panel). Cell lineage can also be encoded in a polyadenylated barcode sequence through the cumulative editing of a lineage array sequence by Cas9 (right panel). Over time, Cas9 will cut the lineage array, resulting in mutations at different points in the array. Cells sharing common mutations in the lineage array are likely to have originated from the same progenitor. By placing the lineage array sequence under the control of an RNA polymerase II promoter, these sequences can also be captured alongside endogenous mRNAs. **d** | Additional information can be extracted from scRNA-seq data beyond a typical analysis that provides only estimates of transcript counts in each cell. Somatic mutations can be identified from sequencing reads for each individual cell and can be used to reconstruct lineage relationships between cells. Retained introns can also be detected and can be used to give an estimate of the rate of change in transcript abundance (RNA velocity[70]). scBS-seq, single-cell bisulfite sequencing; scDNA-seq, single-cell DNA sequencing.

have now achieved parallel genome and transcriptome sequencing from the same cell either through the physical separation of mRNA and genomic DNA using biotinylated oligo(dT) primers[44] or the selective incorporation of T7 promoter sequences into cDNAs allowing subsequent selective amplification of cDNAs over genomic DNA through in vitro transcription[45]. This has allowed a direct association between genotype and gene expression to be made, revealing that DNA copy number variations and chromosomal rearrangements may explain some of the variability in mRNA abundance between individual cells. These methods will be of particular interest for tissues with high levels of somatic genetic variation, such as tumours.

Building on methods established by Macaulay et al.[44] and the single-cell bisulfite sequencing methods pioneered by Smallwood et al.[17], a sodium bisulfite treatment step before PCR amplification of the genomic DNA fraction has allowed the capture of single-cell DNA methylation patterns along with gene expression data[46]. This enables a multimodal analysis of DNA methylation patterns and gene expression within the same cells. As DNA methylation patterns are plastic and vary greatly between cell types[47], it is essential to decouple epigenomic variation from cell-type heterogeneity when aiming to decipher the association between DNA methylation marks and gene transcription. By gathering DNA methylation and gene expression data from the same cell, a direct association between epigenomic variation and transcriptional variation can be made. This information allowed the association between gene expression and DNA methylation patterns at regulatory regions to be assessed within the same cell and provided further support for the established negative association between promoter methylation and gene expression, whereas DNA methylation at distal regulatory regions appears to have more variable effects upon gene expression[46]. The simultaneous collection of chromatin accessibility and gene expression data has been performed by selectively tagging genomic DNA and cDNA molecules with specific barcode sequences[48]. Through a combinatorial indexing strategy, the authors were able to co-assay thousands of single cells and identify *cis*-regulatory elements that are likely to influence the expression of nearby genes[48].

The physical separation of the cellular lysate can also enable the simultaneous detection of intracellular proteins and RNAs in single cells[49,50]. One study used a lyse-and-split strategy to detect both RNA and protein, separating the lysed cell into parallel workflows[49]. In one fraction, protein levels were quantified using the proximity extension assay (PEA) followed by quantitative PCR (qPCR), whereas mRNAs were detected using quantitative reverse transcription PCR (qRT-PCR). PEA utilizes two different antibodies targeting the same protein to bring two conjugated DNA sequences into close proximity, providing a measurement that is robust to unbound antibody signals. Another study used a similar approach implemented in a microfluidic chamber where reverse transcription and PEA were performed in the same compartment without separating cell lysis fractions[50]. These studies highlight the complementary nature of protein and RNA data, as cells were able to be more accurately classified when both data types were used than when either was used alone[49]. However, both the quantity and quality of data retrieved by these approaches are limited. These methods were used to profile 82–96 RNAs and 38–75 proteins in a single experiment (TABLE 1), and so, there is a need to greatly increase the number of RNAs and proteins measured[49,50]. Additionally, the requirement for physical separation of cellular compartments or the use of microfluidic circuitry imposes limitations on cellular throughput.

Conversion of cellular information into a common molecular format. The experimental conversion of multiple data types into a single molecular format is one powerful approach for multimodal profiling, enabling multiple data types to be measured in parallel through a single workflow (FIG. 2c). A particularly relevant example involves the simultaneous conversion of cell surface protein information and mRNAs into cDNAs, allowing both to be detected simultaneously through DNA sequencing. Recently, two methods (CITE-seq and REAP-seq) exploited the use of DNA barcodes conjugated to antibodies to enable the measurement of cell surface protein abundance alongside mRNAs with single-cell resolution[20,21]. By attaching poly(A) sequences to the antibody barcodes, the barcode sequences can be hybridized by the cell-specific reverse transcription primers and extended by reverse transcriptase, allowing antibody barcodes to be detected alongside mRNAs (FIG. 2c). These methods circumvent some of the limitations of FACS-based cell surface protein detection through the use of DNA

In vitro transcription
Transcription of a DNA sequence in vitro using the T7 RNA polymerase.

CITE-seq and REAP-seq
Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) and RNA expression and protein sequencing assay (REAP-seq) are methods that are capable of detecting cell surface protein abundance and gene expression within the same single cell. They achieve this through the use of barcoded antibodies captured alongside mRNA transcripts in single-cell RNA sequencing (scRNA-seq) experiments.

CRISPR–Cas9
A protein–RNA complex that allows targeted mutation or binding of DNA sequences as determined by a guide RNA sequence.

Pooled genetic screens
Screening experiments in which each individual cell may receive a different perturbation at random without prior separation of groups of cells and perturbation treatments.

Lineage tracing
The identification of lineage relationships between groups of cells through shared DNA mutations.

Single-molecule fluorescence in situ hybridization
(smFISH). A fluorescence in situ hybridization method capable of detecting the presence of a single molecule (usually RNA) through the recruitment of many fluorophores to the same area. It enables a quantitative readout of the number of molecules present in a cell.

barcodes, rather than fluorescent moieties, to label antibodies[20,21]. DNA barcoding allows an arbitrary number of different antibodies targeting different epitopes to be mixed in a single experiment and later resolved through DNA sequencing, as the number of possible barcodes is 4^N, where N is the length of the barcode. Furthermore, these methods are compatible with high-cell-throughput droplet-based scRNA-seq methods[4–6] and so can potentially be scaled to millions of cells, removing many of the limitations of index sorting methods coupled to scRNA-seq. These studies have allowed a more detailed analysis of single cells than is possible by measuring a single modality, enabling fine discrimination between immune cell types that was not possible with mRNA data alone[20], and may be further applied to study post-transcriptional gene regulation. Extending these methods to detect intracellular proteins alongside mRNAs, rather than only cell surface proteins, remains an important ongoing challenge, especially as the permeabilization of the cell typically results in extensive RNA degradation. Although technically imposing, this problem should not be intractable, and there are likely to be rapid developments in this area, either through the use of PEAs previously used to measure protein levels in single cells[49–51] or the adaptation of newly developed scRNA-seq protocols for fixed tissues[7,8].

These rapid developments demonstrate that as long as cellular information can be converted into a sequenceable barcode, this information can be determined with readouts at single-cell resolution alongside the transcriptome. This strategy extends not only to measurements of the natural cell state but also to cellular perturbations. The development of high-throughput scRNA-seq methods and programmable DNA-editing methods using CRISPR–Cas9 presents an ideal combination of technologies suited for large-scale perturbation experiments for forward genetic studies[52–55]. As guide RNA sequences used to guide Cas9 binding sites are transcribed by RNA polymerase III, they are not polyadenylated; therefore, these molecules are not detected by standard scRNA-seq methods that use oligo(dT) primers for reverse transcription. Just as CITE-seq and REAP-seq use polyadenylated antibody barcodes to convert cell surface protein information into nucleic acid sequences, CRISPR-based pooled genetic screens employ polyadenylated guide RNA barcoding systems to provide a signal that is detectable through scRNA-seq. These guide barcode sequences are captured alongside mRNA expression information in standard scRNA-seq workflows, enabling the inference of Cas9 binding sites and gene expression in the same cell on a massively parallel scale. Several groups have now applied large-scale genetic perturbation experiments to study gene regulatory networks[52], the unfolded protein response[53], immune cell development[55] and T cell receptor activation[54]. This approach holds much promise for the dissection of gene regulatory networks, and although studies have so far focused only on the perturbation of coding sequences, Cas9 targeting could be extended to study the roles of enhancers, insulators and other non-coding sequences in the genome[56]. Furthermore, catalytically inactivated versions of Cas9 could be fused to different effector domains, such as histone modifiers, transcriptional activators or repressors, and DNA methyltransferases, in order to assess the effect of epigenetic modifications on gene expression[57–61] and could target combinations of epigenetic effectors with gene knockouts[62].

Another case in which the conversion of cellular information into a sequenceable readout can provide valuable multimodal single-cell data is in the recent development of lineage tracing methods using CRISPR–Cas9 genome editing[28,32–34,63]. Several single-cell lineage tracing methods capable of simultaneously detecting endogenous mRNAs have now been developed[32–34]. These methods typically induce continuous edits to a lineage barcode sequence using CRISPR–Cas9, and these barcodes can subsequently be transcribed into a polyadenylated mRNA and detected using conventional scRNA-seq. The DNA sequence differences between lineage barcodes from individual cells can then be used to construct a lineage tree for the tissue or organism. Regarding the specific experimental details of these methods, Raj et al.[32], building on previous methods[28], induced edits in a synthetic array of Cas9 target sites in the zebrafish genome. By placing the lineage barcode array under an inducible promoter, transcription of the barcode could be induced before cell collection, allowing the lineage barcode to be reverse transcribed and sequenced along with cellular mRNAs[32]. This enabled lineage trees at single-cell resolution to be assembled for the zebrafish brain and a combined analysis of cell transcriptomes and lineage relationships to be performed[32]. Spanjaard et al.[34] took a similar approach but injected embryos at the single-cell stage with Cas9 and guide RNAs targeting 16–32 red fluorescent protein (*RFP*) transgenes integrated into the zebrafish genome at different loci. Over time, different *RFP* copies were mutated by Cas9, producing a lineage signature that could be read out from transcribed *RFP* sequences using scRNA-seq. Another approach, named ScarTrace, instead, used an array of *H2A–GFP* transgenes and CRISPR–Cas9 editing to perform lineage tracing in whole zebrafish[33]. In this case, editing was induced early in development through the injection of Cas9 into the early embryo. In the mature fish, lineage clones could be detected alongside gene transcripts through a nested PCR strategy amplifying the *GFP* genomic DNA sequence performed in parallel with scRNA-seq[33]. This approach provided unique opportunities to study the plasticity of cell fates. In one experiment, following zebrafish fin regeneration after injury, the authors showed that osteoblast progenitor clones were able to change fate and give rise to mesenchymal cells that populated the regenerated fin. Only by collecting both lineage and transcriptome information from single cells could such a phenomenon be found, highlighting the value of these studies[33].

Frieda et al.[27] took a slightly different approach to lineage tracing by performing single-molecule fluorescence in situ hybridization (smFISH) on lineage barcodes rather than scRNA-seq. This provided a unique ability to detect the lineage relationships between cells along with their spatial context. Applying their method, memory by engineered mutagenesis with optical in situ readout (MEMOIR), to mouse embryonic stem cells in culture,

Expression quantitative trait loci
(eQTLs). Genomic loci that explain variation in the RNA expression levels of genes.

Intron retention
The presence of intronic RNA bases in an RNA transcript. These bases are usually removed by RNA splicing shortly after or during transcription.

Pseudotime
The ordering of cells along a one-dimensional axis describing a continuous differentiation process.

Joint clustering
Grouping cells on the basis of measurements from multiple data modalities.

the authors were able to validate the lineage relationships inferred using lineage barcodes through time-lapse microscopy, providing an important ground truth[27]. If future improvements to MEMOIR can extend this in situ lineage tracing to create a profile for a panel of endogenous mRNAs alongside a lineage barcode, these improvements may provide an incredibly powerful multimodal assay capable of simultaneously detecting cell lineage, spatial position and transcriptional state.

Extracting additional information from scRNA-seq data. Although most scRNA-seq studies focus on transcript abundances, these studies can often provide information not just about transcript levels but also about nucleotide sequence, enabling multiple data modalities to be derived from standard scRNA-seq experiments without changes to experimental methods. In particular, these possibilities include the capture of somatic mutations, genetic variants and RNA splice isoforms (FIG. 2d). Somatic mutations can occur randomly in the genome and will be inherited by all daughter cells. Lineage relationships between cells can then be inferred by detecting somatic mutations in single cells. Somatic mutations in the human brain have been used to reconstruct neuronal lineages for a small number of single cells through whole-genome sequencing[64]. Importantly, somatic mutations in the brain occur in hot spots linked to active transcription and are enriched in coding exons[64], suggesting that neuronal somatic mutations may be able to be detected from scRNA-seq data and used to reconstruct cell lineage relationships along with the transcriptional state. Furthermore, many cancers exhibit an accelerated rate of somatic mutation, often linked to late-replicating regions of the genome. Tirosh et al.[65] inferred copy number variants in melanoma tumours directly from scRNA-seq data by averaging the expression values of 100-gene stretches over the genome, revealing common patterns of aneuploidy in the tumour cells. This allowed single-cell genotype information to be integrated with cell transcriptomes to enable a more informative analysis of cancer biology[65]. Similarly, Fan et al.[66] identified copy number variants and loss-of-heterozygosity events directly from scRNA-seq data and applied their method to study multiple myeloma samples from patients, revealing transcriptional heterogeneity between cancer clones[66].

Single-cell analysis also offers a novel approach to understand how natural variation in DNA sequence influences variation in phenotypes such as gene expression and cell state. Developing a novel set of scRNA-seq data analysis tools to genotype single cells from mixed donors, Kang et al.[67] performed a genome-wide association study with a small cohort of 23 human donors and identified expression quantitative trait loci (eQTLs) associated with cell-type-specific gene expression variation between individuals. Furthermore, they identified a genetic variant associated with altered proportions of immune cell types, highlighting the ability of this approach to model both gene expression and cell-type frequency as quantitative traits that can be associated with genetic variants[67]. Similar approaches were utilized by van der Wijst et al.[68], and future studies utilizing much larger cohorts have the potential to substantially increase the resolution of bulk-level eQTL studies[69] to determine the precise effects of genetic variation on cell state and function.

Notably, key information on transcript isoforms can be obtained from scRNA-seq data, even from methods that capture only the 3′ end of gene transcripts[70], and intron retention data can provide a surprising amount of information to complement transcript abundance measurements. In pioneering work, La Manno et al.[70] demonstrated that the frequency of unspliced introns in scRNA-seq transcripts is related to the relative ratio of mRNA production to degradation, with newly transcribed genes being more likely to contain introns, as they may be captured before being fully processed into mature mRNAs. By measuring the frequency of unspliced introns, the authors were able to derive an estimate of the rate of change in transcript abundance (RNA velocity) and estimate the future transcriptional state for each cell. These estimates of future state additionally allowed cells to be placed on a pseudotemporal trajectory and solved some of the most difficult problems faced by other methods aiming to derive pseudotime measurements from single cells, including 'rooting' the trajectory (identifying the start and end points), branching and dealing with cyclic trajectories[70–75]. These methods have the potential to transform the field of single-cell biology.

Analysis of multimodal data. The rapid development of multimodal profiling strategies has created a subsequent need for innovative analytical approaches for these data types. Although these techniques are largely under development, we anticipate that multimodal data sets are likely to reveal subtle differences in cell state that cannot be captured by a single modality alone (FIG. 3a). This is particularly true for scRNA-seq data, for which incomplete detection ('drop-out') of lowly expressed genes can blur fine-scale distinctions, but complementary data from the same cells can ameliorate this problem. For example, distinct T cell groups (including memory and regulatory subsets) can be challenging to distinguish on the basis of sparse scRNA-seq information but are readily classified according to the expression of cell surface protein markers. This suggests that future methods that perform joint clustering on both immunophenotype and mRNA levels collected from the same cells may achieve dramatically higher resolution in characterizing immune cell states[20] (FIG. 3b). Similarly, co-assays of the transcriptome and chromatin[48] or methylation[46] state may reveal heterogeneity in the regulatory landscape of individual cells, which can bias fate decisions even in advance of transcriptional changes.

Statistical approaches for multimodal single-cell integration are likely to be inspired by bulk approaches (reviewed elsewhere[76]) that are used to perform joint dimensionality reduction on multiple omics data sets to identify conserved or divergent patterns and can be readily extended to single-cell multimodal data. For example, Argelaguet et al.[77] developed a multi-omics factor analysis (MOFA) method capable of identifying a set of factors that explain variance across multiple data modalities and used their method to jointly analyse bulk genomic, DNA methylation and RNA expression data from patients with chronic lymphocytic leukaemia. This integrated analysis

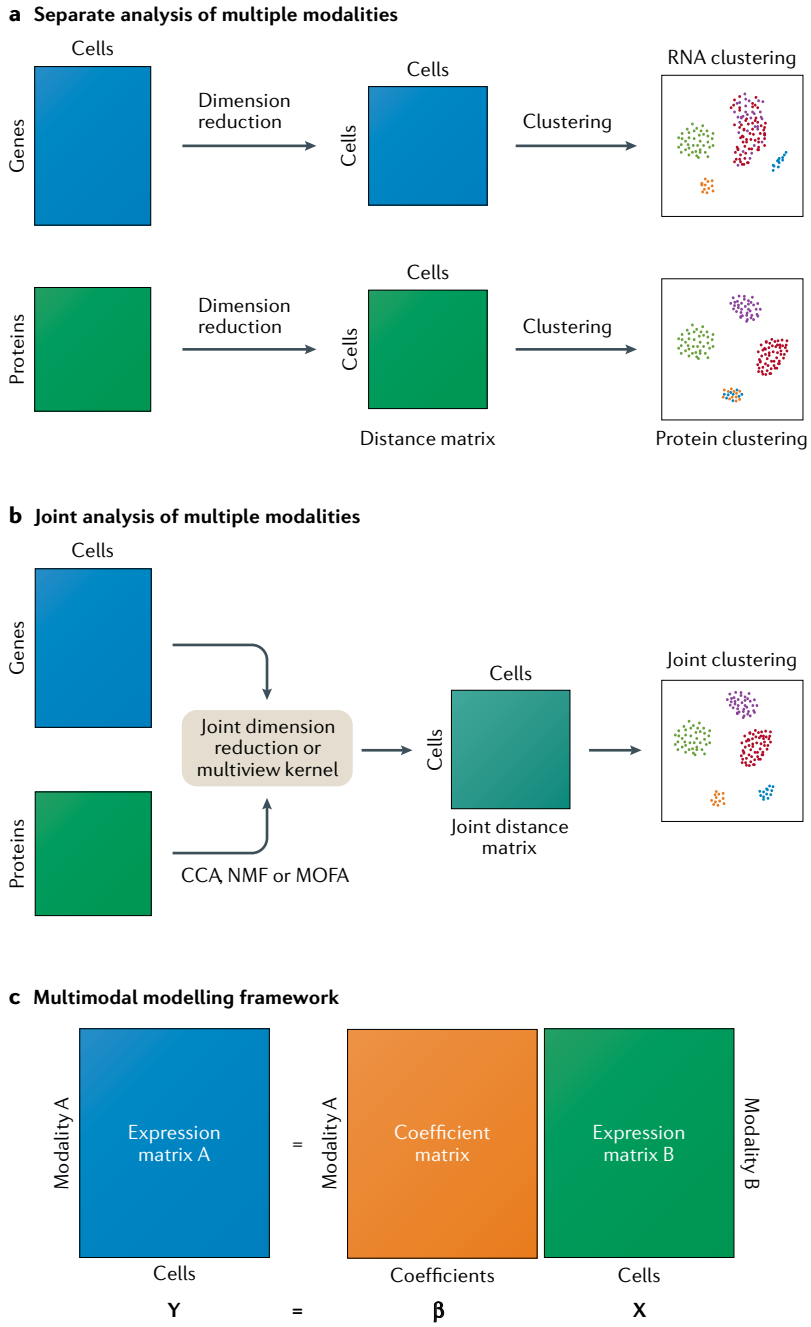


Fig. 3 | **Computational methods for the analysis of multimodal single-cell data.** **a** | Independent analysis of multiple modalities measured from the same group of cells can lead to conflicting identification of clusters of cells. Measurements from one modality may indicate that certain cells are highly similar, while the same cells appear very different when looking at a different modality. In this example, the orange and blue cells form separate clusters when analysed on the basis of their transcriptional profile but cluster together when analysed according to protein measurements. Similarly, the red and purple cells group together based on RNA but separate based on protein. **b** | Joint analysis of multiple modalities measured from the same group of cells can have greater power to identify unique cell states. By taking into account information from multiple modalities, cells can be jointly clustered in a way that outperforms the classifications derived from any one modality alone. In this example, both the red and purple cells and the orange and blue cells form separate groups when protein and RNA information is analysed jointly. **c** | The relationships between data modalities can be studied by building statistical models that aim to explain variance in one modality using a linear combination of components of another modality. CCA, canonical correlation analysis; MOFA, multi-omics factor analysis; NMF, non-negative matrix factorization.

enabled the identification of transcriptional modules whose activation was correlated with the somatic mutation status for each sample[77]. Applying MOFA to 87 single-cell methylation and transcriptome sequencing profiles[46] also identified correspondences between modalities and revealed coordinated DNA methylation and transcriptome changes during the transition from naive to primed pluripotent states in mouse embryonic stem cell differentiation[77]. These findings indicate that these methods are also suitable for the analysis of multimodal single-cell data and may facilitate improved interpretation of such data sets in the future. As single-cell data sets extend far beyond bulk sample sizes, powerful approaches for 'multiview' machine learning (summarized elsewhere[78]) are also likely to give rise to valuable single-cell analysis techniques.

Multimodal single-cell experiments also provide a unique opportunity to study the relationships between different components in a cell. For example, simultaneous mRNA and protein profiling[20,21] can be used to readily identify instances in which the modalities are poorly correlated, indicating active post-transcriptional regulation. Further integration with other measurements, such as RNA velocity[70], can enable the construction of rich dynamic models of protein and RNA production. Statistical models can also be leveraged to assess the ability to predict one modality when given the measurements of another (FIG. 3c). For example, Cao et al.[48] built linear regression models to predict gene expression values from chromatin accessibility data and found that including distal accessibility sites improved the gene expression predictions fourfold compared with models that only included *cis*-regulatory sites. Dixit et al.[52] built regularized linear models to estimate the impact of a given set of guide RNAs in each cell on gene expression levels in order to identify causal drivers of cellular responses and to reconstruct transcriptional networks. These models could be used to determine the variance in gene expression explained by the presence of guide RNAs within the cells, as well as other covariates[52]. We therefore anticipate that multimodal data sets will yield mechanistic insights into complex regulatory processes, including epigenomic, transcriptional and post-transcriptional gene regulation.

Integrating single-cell data across experiments

Whereas, in the previous sections, we focused on the integration of measurements collected in the same single cells, the joint analysis of data sets collected from different single cells poses a key computational challenge for single-cell biology. This challenge echoes similar needs for 'batch-correction' techniques for bulk data sets, which are essential for data sets produced across different laboratories and experimental workflows[79]. However, existing methods tailored for bulk-level measurements cannot be applied to heterogeneous single-cell data, as they cannot distinguish between shifts in the proportional composition of cell types and changes in the molecular programme within a cell type[80,81]. Newly developed approaches that can first identify shared biological states (for example, matched cell types) across data sets can overcome this challenge (FIG. 4a) and have become an area of rapid analytical development.

Canonical correlation analysis
(CCA). A statistical method for investigating relationships between two data sets. CCA aims to identify shared sources of variation in a pair of data sets.

Dynamic time warping
A method for locally stretching or compressing two one-dimensional vectors to correct for lag in one vector relative to another.

Mutual nearest neighbours
(MNNs). Cells that are mutually nearest to one another in normalized gene expression space.

Cell-type classifications
Biologically meaningful labels given to groups of cells on the basis of common molecular profiles and prior knowledge of the cell types.

Computational integration of scRNA-seq data. We recently introduced a method to harmonize single-cell measurements collected across data sets as part of the Seurat v2 R toolkit[80,82]. The method first applies canonical correlation analysis (CCA) to identify shared sources of variation between the data sets (FIG. 4a). The resulting canonical correlation vectors represent the presence of shared cell types across data sets as opposed to batch effects or data set-specific sources of heterogeneity that would typically be captured by a standard principal component analysis (PCA) (FIG. 4b). Next, these canonical correlation vectors are aligned across data sets using dynamic time warping, a nonlinear transformation that corrects for differences in cell population density. These two steps project cells into a low-dimensional space that is shared across multiple scRNA-seq data sets, where cells of the same biological state will be located close together regardless of their experimental origin[80].

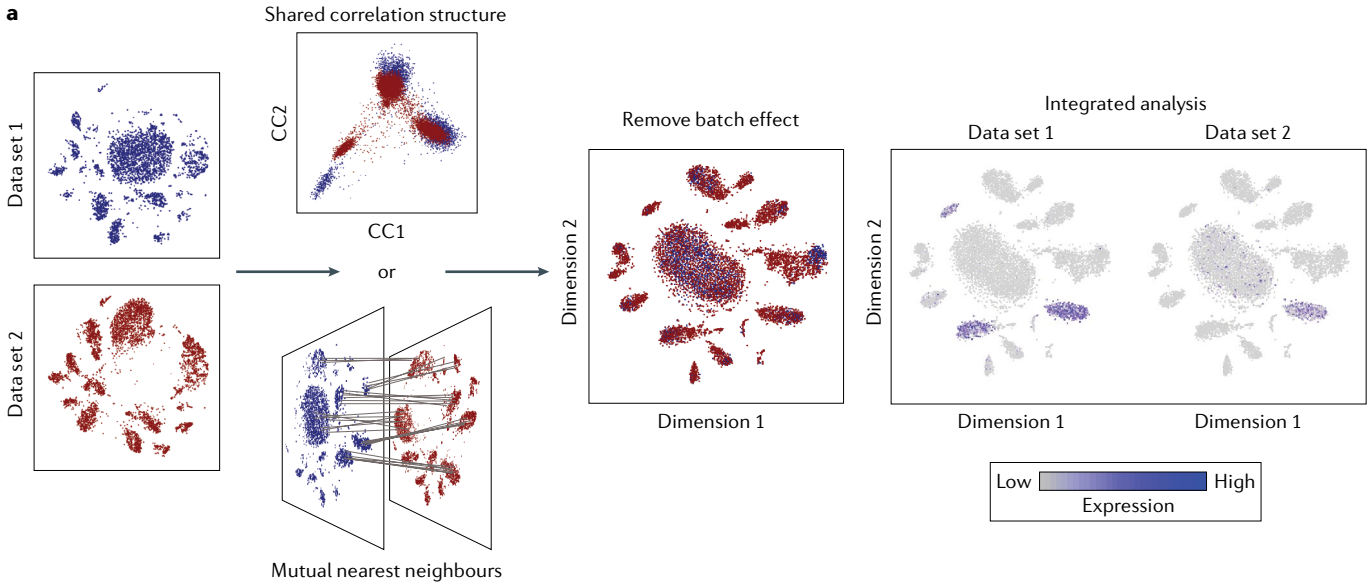
A complementary approach, mnnCorrect, accomplishes similar goals through the innovative application of techniques that have previously been applied to shape and pattern matching across images[83]. This method relies upon the identification of mutual nearest neighbours (MNNs), representing cells that are mutually closest to each other across data sets and therefore are likely to represent a shared biological state[81] (FIG. 4c). The distance between paired MNNs can then be used to compute a batch vector, which can be used to correct the original gene expression matrix. Importantly, it is not necessary for all cells to have an MNN in order for corrected expression values to be calculated for every cell, as the batch vector for each cell is calculated by weighting the batch vectors of nearby matched cells[81].

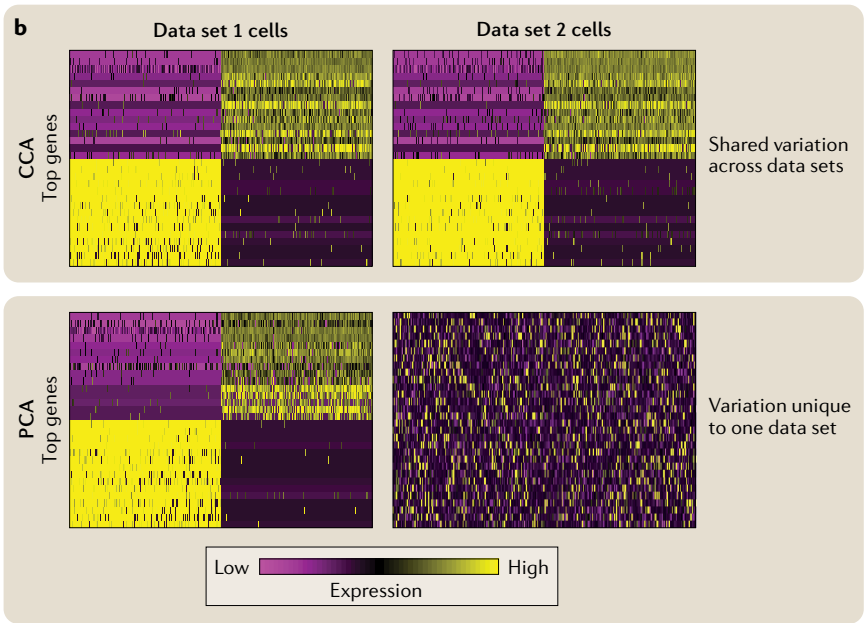
Both CCA and mnnCorrect enable the integration and pooling of scRNA-seq data sets generated by different laboratories and technologies but from the same underlying tissue. By re-analysing data sets from the literature (four independent single-cell analyses of human pancreatic islets), both approaches successfully harmonized the experiments into a single joint data set, enabling a robust meta-analysis with substantially greater sample size compared with that of any individual data set. The resulting increase in statistical power can dramatically boost the ability to discover rare or transcriptionally subtle cell states[80] and the gene expression markers that define them[81]. These methods therefore represent an initial proposal to solve a key challenge for both individual laboratories and large consortia that seek to construct a single reference data set from many individual single-cell experiments. Given the exciting potential for these types of comparison, there has been rapid development of new analytical approaches for scRNA-seq integration, with a particular focus on computational efficiency[84–89]. Of particular note, Korsunsky et al.[87] developed a novel variant of *k*-means clustering that favours clusters containing cells from multiple data sets, enabling scalable integration of 500,000 cells on a personal computer.

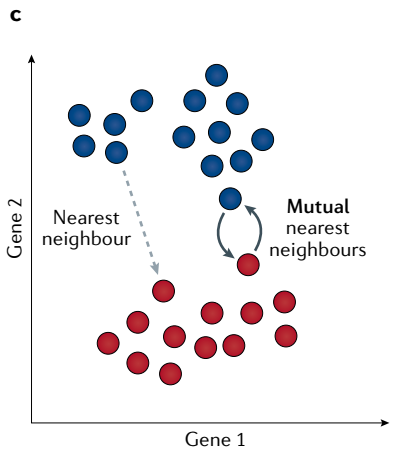
Importantly, effective integration of single-cell data sets can extend far beyond batch correction (FIG. 4d) and enable in-depth comparisons of distinct biological conditions at single-cell resolution. For example, Butler et al.[80] integrated data sets of human peripheral blood mononuclear cells (PBMCs) across both control and interferon-β-stimulated conditions, identifying 13 shared cell types between experiments, and systematically compared their transcriptomes to identify cell-type-specific responses to stimulation. This analysis revealed that plasmacytoid dendritic cells exhibited a striking and unique response to interferon-β, represented by specific gene modules that could be validated by flow cytometry and bulk RNA-seq. We anticipate that similar analyses will help to uncover cell-type-specific responses to environmental and genetic perturbations and even to standardize comparisons between patient samples across disease phenotypes and treatment status.

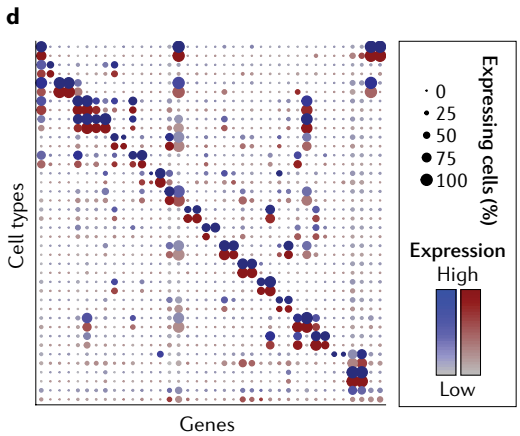
The integration of scRNA-seq data not only applies across technologies and conditions but also can extend to cross-species analyses as well. Indeed, multiple groups have embraced this approach to study the evolution of cell states in diverse systems. Karaiskos et al.[90] compared spatial gene expression patterns between two different *Drosophila* species during the early embryonic stage. By building spatial gene expression maps (discussed further below), the authors were able to systematically compare expression profiles of orthologous genes across species and identify evolutionary changes[90]. Tosches et al.[91] performed scRNA-seq on reptilian brain cells and compared these with mammalian brain cells by computing the correlation between averaged gene expression values within clusters of each species. This enabled the identification of strikingly conserved inhibitory subsets, alongside poorly conserved excitatory groups, between the reptile and the mammalian brain. Similarly, Baron et al.[92] generated scRNA-seq data sets of pancreatic islets in human and mouse tissue, identifying matched subsets and regulatory programmes across species, particularly among four highly conserved endocrine types[92]. In this example, the matching of cell types across species could also be performed using Seurat CCA alignment[80,92] in an unsupervised manner. Finally, Alpert et al.[93] developed cellAlign to compare scRNA-seq data sets of embryonic development between humans and mice by aligning one-dimensional (1D) pseudotime trajectories for the two species (FIG. 4e). They found that human embryos underwent zygotic genome activation later than mice did, whereas many genes that exhibited faster dynamics in mice were involved with protein biogenesis[93]. Although these methods remain at early stages, they reveal a potentially exciting future to apply comparative biology to single-cell resolution and to identify correlates of human cell types in model organisms for further study. However, the accurate identification of gene orthologues across species is essential for the resulting data integration to be reliable and may become one of the most difficult challenges for multispecies data integration spanning many millions of years of evolution.

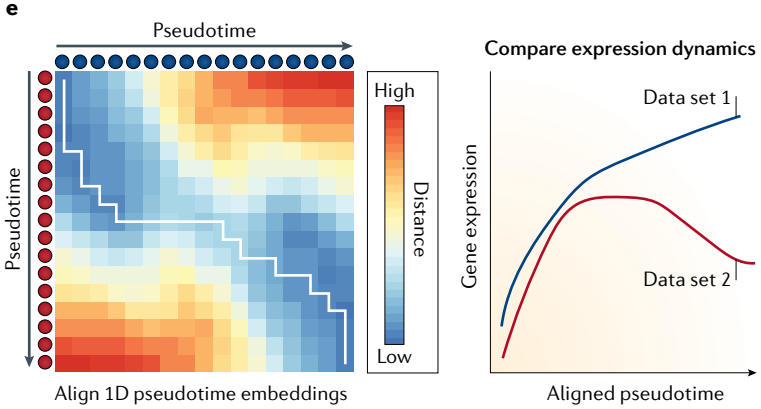
Classification of cells across scRNA-seq data sets. The transfer of information in the form of cell-type classifications from one experiment to another is often highly desirable, as it may outperform de novo clustering of cells, and will become more common as high-quality, annotated cell atlases are developed by the community[94] (FIG. 5a). So far, two related methods


a
Shared correlation structure
Data set 1
Data set 2
CC2
CC1
or
Mutual nearest neighbours
Remove batch effect
Dimension 2
Dimension 1
Integrated analysis
Data set 1
Data set 2
Dimension 2
Dimension 1
Dimension 1
Low
High
Expression


b
Data set 1 cells
Data set 2 cells
CCA
Top genes
Shared variation across data sets
PCA
Top genes
Variation unique to one data set
Low
High
Expression


c
Gene 2
Nearest neighbour
Mutual nearest neighbours
Gene 1


d
Cell types
Genes
Expressing cells (%)
0
25
50
75
100
Expression
High
Low


e
Pseudotime
Pseudotime
High
Distance
Low
Align 1D pseudotime embeddings
Compare expression dynamics
Data set 1
Data set 2
Gene expression
Aligned pseudotime

◀ Fig. 4 | **Computational approaches for integrating multiple single-cell data sets.** **a** | Multiple data sets can be integrated computationally to facilitate downstream comparative analysis. Shared correlation structure can be detected using canonical correlation analysis (CCA) or mutual nearest neighbours (MNNs) identified. The identification of either a shared space or equivalent cells across groups can then be used to eliminate batch-specific variation, enabling direct comparison between the groups. **b** | CCA aims to identify a set of variables that are maximally correlated between two data sets. By contrast, methods such as principal component analysis (PCA) aim to find orthogonal variables that maximize the variance explained in a single data set. **c** | The identification of cells that are mutually nearest to one another in a space, defined by the gene expression profiles of the cells, allows the identification of biologically equivalent cells. Once equivalent cells have been identified across data sets, this information can be used to compute a transformation of the original expression data that would remove data-set-specific expression patterns. **d** | Once equivalent cell states have been identified across data sets, gene expression values within each cell state can be compared across the data sets to identify similarities or differences in gene expression patterns between the data sets. In this example, cells in each data set are grouped by their cell type, and the expression of selected markers for each cell type is shown. The size of each dot corresponds to the percentage of expressing cells in the group, while the colour corresponds to the expression level. This makes it easy to visually compare expression profiles for cells in both data sets for each cell type. Further statistical tests could be used to identify genes that are differentially expressed within a cell type or cluster between data sets. **e** | The alignment of one-dimensional (1D) pseudotime vectors from different data sets can allow temporal differences in cell trajectories to be removed and equivalent points in the trajectory across two data sets to be identified. Gene expression can then be directly compared across the corrected pseudotime trajectories to identify similarities or differences in gene expression across groups. CC, canonical correlation vector.

have been developed that are capable of projecting cells onto an existing data set to facilitate the transfer of cell labels: scmap-cell and scmap-cluster[95]. The scmap-cell method identifies nearest neighbours across data sets, allowing cells to be assigned cell-type labels on the basis of the labels of their neighbouring cells[95]. By contrast, scmap-cluster aims to classify cells in a query data set by finding the nearest cluster centroids in a reference data set defined by correlation-based distance measures or cosine similarity. New methods currently under development employ singular value decomposition[96], linear discriminant analysis[97] or support vector machines[98] to classify cells on the basis of an annotated reference data set. Notably, supervised annotation may have substantially greater power to resolve cell types than unsupervised clustering does, particularly as the size, depth and coverage of reference data sets continue to grow. The identification of cell subpopulations through reference-guided cell annotation may enable the analysis of closely related groups of cells that would otherwise be unable to be resolved using de novo clustering methods (FIG. 5b).

Computational integration of multimodal single-cell data. The integration of scRNA-seq data with different types of single-cell data that do not share common features poses a distinct problem that may require different methods to address. This is particularly true for the comparison of genomic-based measurements (for example, chromatin accessibility or DNA methylome data) with gene-based measurements (gene or protein expression data), as the correspondence between features is unclear. However, by collecting cells from similar populations, common biological states can be identified across data modalities that can assist in the identification of correspondences between modalities. Welch et al.[99] developed a method to align 1D pseudotime trajectories across experiments, named MATCHER, which is tailored to the alignment of different data types. MATCHER assumes a common, underlying, developmental trajectory that similarly impacts both modalities and projects cells from different experiments onto a common 1D pseudotime space. This projection allows the identification of equivalent cells between multiple experiments without requiring prior knowledge of feature correspondence between the modalities. Welch et al.[99] applied MATCHER to integrate scRNA-seq data with single-cell methylome and transcriptome (scM&T-seq) data[46,99] to study transcriptome and DNA methylation dynamics during human induced pluripotent stem cell (iPSC) reprogramming. This revealed that DNA methylation changes often lag behind changes in gene expression[99].

Although cell types can often be identified from scRNA-seq data by the expression of cell-type-specific marker genes, much less is known about the cell-type-specific activity of features of other data modalities measured in single cells, such as accessible chromatin regions. By integrating such data with scRNA-seq, cell-type classifications derived from gene expression data can be used to guide the assignment of cell-type classifications in other modalities. Lake et al.[15] performed single-nucleus RNA sequencing (snRNA-seq) and single-cell transposome hypersensitivity site sequencing (scTHS-seq) on a variety of matched brain tissue sections. The authors leveraged single-cell gene expression data to guide the assignment of cell types in the chromatin accessibility data set by using gradient boosting[15]. By first identifying a subset of corresponding cell types in both the scRNA-seq and scTHS-seq data, the authors were able to train a model relating gene expression patterns to patterns of chromatin accessibility. They then applied this model to classify the remaining scTHS-seq cells, whose type could not be determined from the chromatin accessibility data alone. These classifications, obtained through the integration of multiple data modalities, allowed a more nuanced interpretation of the brain chromatin accessibility data than would be possible from only one data set, including the identification of pathogenic cell types that may underlie common genetic diseases[15]. Similar approaches may also assist in the interpretation of single-cell DNA methylation or single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data sets.

New methods under development can enable the cross-modality classification of cells through the assumption of equivalent features or the identification of features that are assumed to share correlations across the modalities[88,89]. Welch et al.[89] developed an integrative non-negative matrix factorization (iNMF) method, named LIGER, that is capable of integrating data across modalities. They applied LIGER to classify cortical cells profiled by single-cell bisulfite sequencing[16] using a scRNA-seq data set[100] generated from the same tissue[89]. Welch et al.[89] assumed a negative correlation between gene-body methylation and gene expression to integrate the different data modalities, thus allowing the cells to be jointly clustered. We also recently introduced an integration method, implemented in Seurat v3, that is capable of

Gradient boosting
A statistical method that produces a prediction model for classification or regression on the basis of an ensemble of weaker prediction models.

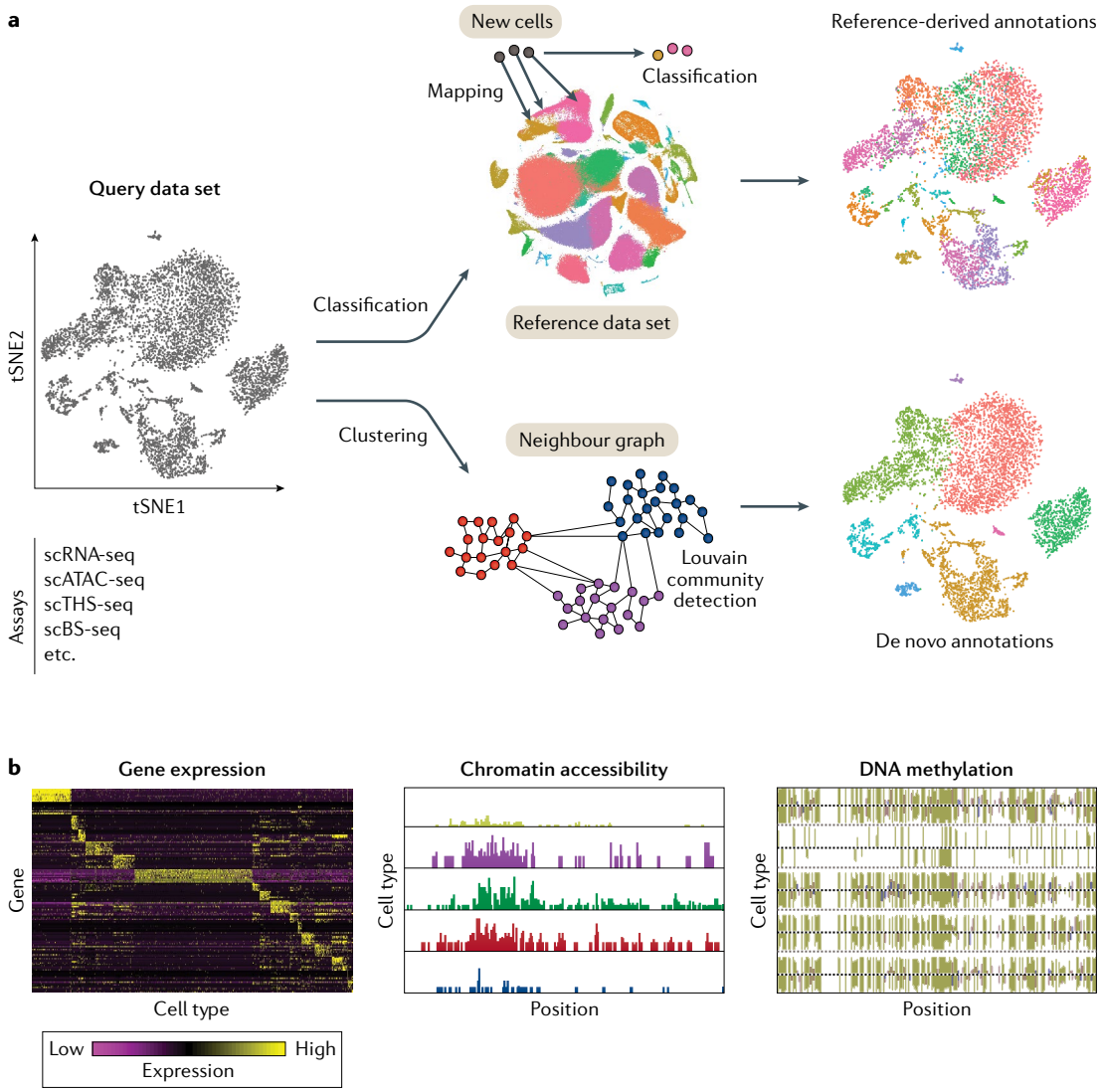


Fig. 5 | **Clustering and classification of cells. a** | The classification of cells in a new data set can be achieved in one of two ways. Cells can be classified on the basis of existing cell-type annotations in a reference data set derived from a similar population of cells, or unsupervised clustering can be performed to identify groups of similar cells. By classifying cells according to reference annotations, more subtle divisions between groups of cells may be identified by leveraging the structure present in a higher quality data set. **b** | By grouping cells on the basis of common properties, through either a reference-based classification or de novo clustering, the differences between groups of cells can be further analysed in different ways depending on the type of data measured in the cells. Gene expression data can be used to identify differentially expressed genes between cell types or clusters, and chromatin accessibility can be used to identify differentially accessible regions and enriched DNA motifs. Similarly, single-cell DNA methylation data can be used to identify differentially methylated regions between cell types or clusters. scATAC-seq, single-cell assay for transposase-accessible chromatin using sequencing; scBS-seq, single-cell bisulfite sequencing; scRNA-seq, single-cell RNA sequencing; scTHS-seq, single-cell transposome hypersensitivity site sequencing; tSNE, *t*-distributed stochastic neighbour embedding.

classifying cells across modalities by assuming equivalent or correlated features in both modalities[88]. This allowed the classification of cells from scATAC-seq data on the basis of annotated cell types from scRNA-seq data from a similar tissue and enabled the identification of subpopulations of cells that could not be separated using the scATAC-seq data alone. We expect that these approaches will enable a more fine-grained analysis of multimodal single-cell data in the future and allow the identification of cell-type-specific patterns of chromatin accessibility and DNA methylation in many different tissues.

Integration of sequencing and spatial data

The spatial organization of cells in tissues often reflects functional distinctions between cells and differences in cell fate and lineage[101]. Spatially restricted patterns of gene expression give rise to the anatomical complexity of multicellular organisms through development, as the expression of different gene sets leads cells along different developmental paths and produces the precise spatial arrangements of cell types that define tissues. Importantly, this spatial information is not fully captured by the RNA expression profile of cells profiled by scRNA-seq, as cells

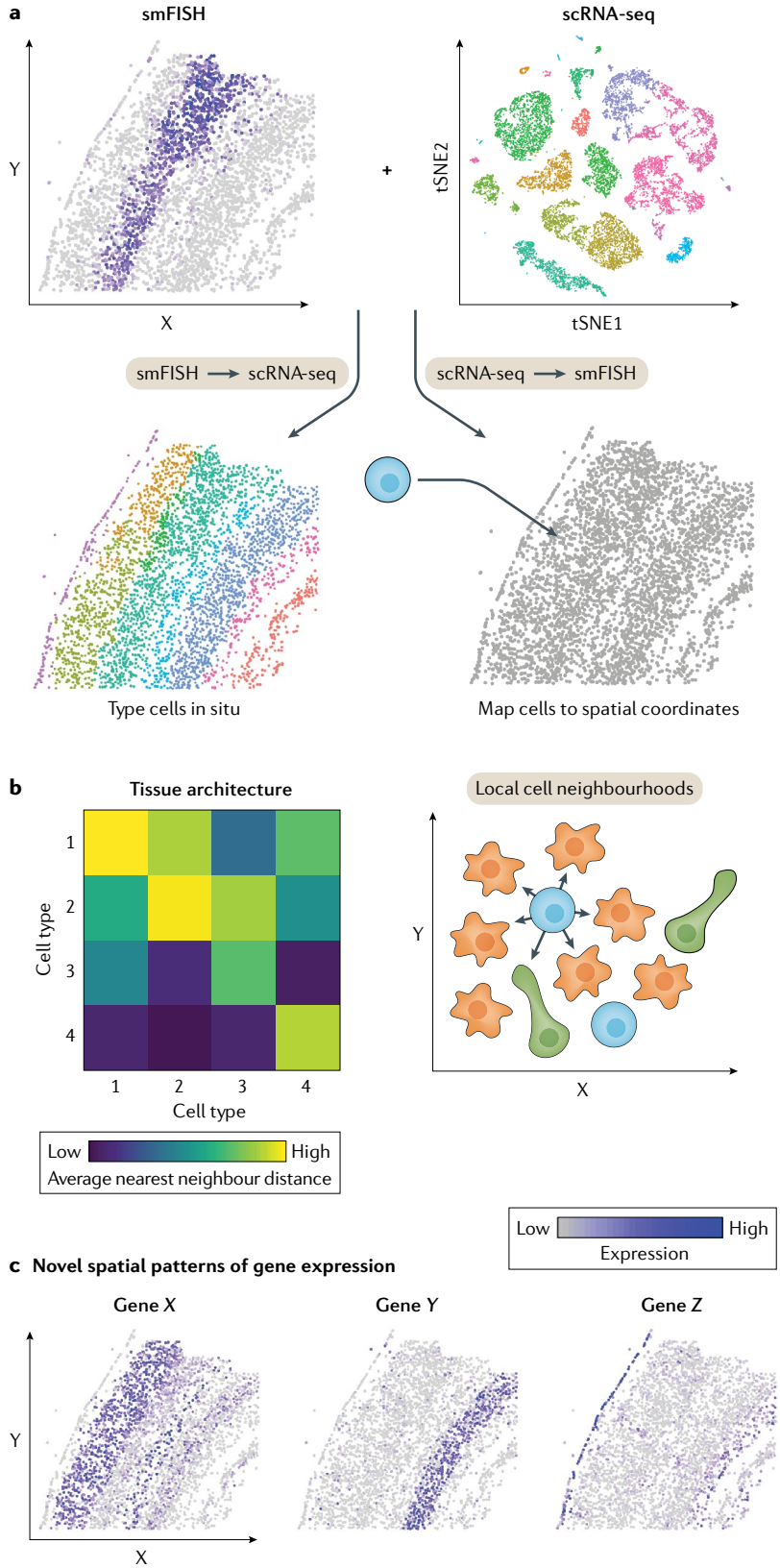


Fig. 6 | **Integration of spatial single-cell data.** **a** | The integration of single-molecule fluorescence in situ hybridization (smFISH) data with single-cell RNA sequencing (scRNA-seq) data can be done in two ways: mapping smFISH-profiled cells onto scRNA-seq clusters or mapping scRNA-seq-profiled cells onto spatially resolved smFISH data. Mapping smFISH cells onto scRNA-seq data allows the transfer of cell-type classifications derived from transcriptome-wide gene expression measurements to be transferred to the spatially resolved cells (left panel), whereas mapping scRNA-seq data onto smFISH-profiled spatial coordinates can allow scRNA-seq data from dissociated cells to be placed back into their spatial context (right panel). **b** | Following spatial integration, tissue architectures can be analysed to determine the cellular composition of tissues and the spatial relationships between cell types. One way of assessing how cell types are spatially organized in the tissue is to look at the local neighbourhood surrounding cells of each type. By measuring the average spatial distance between cell types, it is possible to learn characteristics about the tissue architecture, including which cell types are dispersed throughout the tissue and which cell types often form local neighbourhoods with another cell type, indicating a possible interaction. **c** | By mapping scRNA-seq-profiled cells onto spatially resolved coordinates through the integration with smFISH data, spatial patterns of gene expression can be predicted for any gene measured in the scRNA-seq data set. Through these predictions, novel spatial patterns of gene expression may be identified through the analysis of genes that were not profiled by smFISH. tSNE, *t*-distributed stochastic neighbour embedding.

are dissociated before analyses, typically without retention of information about their original tissue context. Equivalent cell types that may share similar gene expression profiles can occupy distinct spatial domains in situ, and therefore, the loss of spatial information during cellular isolation is a major shortcoming of many single-cell analysis methods. The integration of spatial coordinates with gene expression data from single cells can resolve these experimental shortcomings by combining high-resolution gene expression profiles with spatial expression maps (FIG. 6a). This can be achieved either by using computational methods or by simultaneously collecting spatial coordinates along with gene expression values by quantifying RNAs in single cells in situ.

Several methods are now well established for the measurement of spatially resolved gene expression in situ. The use of fluorescence in situ hybridization (FISH) has become the gold standard for providing in situ gene expression data[102], and new iterations of the technology are approaching 100% detection efficiency[29,35,103,104]. FISH methods are typically used to profile a somewhat small number of cells in a single experiment and are not able to detect the full complement of expressed genes within a single cell[102]. Newer developments employ sequential probe hybridizations with error-correcting codes in order to detect hundreds of genes in a single experiment or use spatial barcoding methods to record spatial information during reverse transcription of mRNAs[31,104–110]. By contrast, scRNA-seq methods typically collect no spatial data but are able to detect many thousands of transcripts within a single cell, for hundreds to millions of cells in parallel[3–8,111]. Through the computational integration of these different data types, the strengths of each method can be leveraged to enable data transfer between data sets, coupling high-throughput scRNA-seq methods with high-resolution spatial information.

The computational integration of spatial gene expression data gathered using FISH and scRNA-seq has now been performed successfully in a number of landmark publications. Initially proposed independently by Satija et al.[82] and Achim et al.[112], and later applied to other tissues[90,113], these computational methods provide integrated spatial expression maps for whole organisms or tissues. These studies typically measure the spatial distribution of key genes that are known to exhibit spatial patterning and use these data to build expression models for each gene. These spatial gene expression models are then used to map single cells captured through scRNA-seq back into their spatial context on the basis of the expression levels of the spatially interrogated landmark genes. In the resulting integrated data sets, the spatial profile of nearly any gene can then be interrogated at high resolution, and the local neighbourhoods inhabited by each cell type can be studied[82,90,112,113] (FIG. 6b). These approaches have enabled the discovery of novel spatially regulated genes, as well as the construction of important resources for the broader research community (FIG. 6c). Further methods have now been developed that enable a systematic analysis of spatial expression trends from spatially integrated scRNA-seq data[114,115]. However, the computational integration of FISH and scRNA-seq data has so far been applied only to organisms or tissues with a well-defined spatial structure, such as the early embryo and the mammalian liver. Extending these methods to more complex spatial structures including mature tissue sections or solid tumours will prove challenging. Some studies have provided coarse integration of spatial gene expression data with scRNA-seq data through the examination of a small set of genes of interest through FISH or immunohistochemical methods following the identification of cell cluster marker genes by scRNA-seq[116,117]. However, such studies have yet to provide the same level of data integration for these tissues as has been demonstrated for other tissues with a more simplistic spatial organization of cells[82,90,112,113].

Recently, two high-resolution spatial gene expression methods were developed that are capable of detecting tens to hundreds of genes in single cells over a large 2D or 3D spatial region[31,35]. These methods greatly reduce tissue background fluorescence to improve the signal-to-noise ratio for gene detection. In one method, cyclic smFISH (osmFISH), the tissue section is covalently bound to the microscope coverslip, and then, tissue clearing is conducted[35]. In another method, spatially resolved transcript amplicon readout mapping (STARmap), modified DNA bases are incorporated during in situ amplification of probes that allow the cDNA to be covalently bound to a polyacrylamide matrix, enabling stringent tissue clearing without the loss of spatial information in three dimensions[31]. These methods can be applied to gather accurate gene expression information in situ for many genes and cells, allowing cells to be molecularly typed in situ and the spatial distribution of cell types to be assessed (FIG. 6a,b). In both studies, the authors applied their method to study the mouse cortex and were able to classify cell types on the basis of the expression of a panel of cell-type marker genes[31,35]. Importantly, the spatial distributions of these cell types could be further analysed, allowing cells to be grouped into anatomical regions[35] and the 3D spatial distribution of cell types to be analysed[31]. Further integration of these spatial data sets with scRNA-seq data or other single-cell data types may provide an opportunity to develop an unprecedented level of understanding of the composition and function of these tissues[88,89].

Perspective

As single-cell technologies continue to grow and mature, both the number of parameters that can be measured per cell and the quantity of cells and molecules detected will inevitably increase. As a result, there is a growing desire in the community to integrate single-cell data across experiments or modalities. Large-scale collaborative efforts are now underway to build a comprehensive Human Cell Atlas that encompasses every cell in the human body and accompanying atlases for key model organisms[94]. Both the construction and use of these atlases will require effective methods for data integration: first, to integrate data from different laboratories, technologies and human donors in a way that is robust to significant technical variation and, later, to facilitate data transfer and comparative analyses between the atlases and new data sets. Just as the initial sequencing and assembly of the human genome allowed subsequent experiments to be performed more quickly and more cheaply than the Human Genome Project by transferring information from the genome to new data sets through read alignment, the development of the Human Cell Atlas will create similar benefits but only in the presence of appropriate computational tools for data transfer between cells, analogous to DNA sequence aligners. Furthermore, we expect that the number of parameters able to be measured in single cells will continue to grow in the coming years. Nanopore sequencing holds much promise for multimodal single-cell applications owing to its ability to directly sequence both RNA and DNA with long reads and to natively detect nucleotide base modifications[118–120]. Further developments could see similar technologies emerge that are capable of detecting other biomolecules, such as proteins. The continued refinement of methods for high-resolution spatial cell profiling will enable cells to be placed into their spatial context, giving important insight into how cell types are arranged in tissues. Ultimately, gathering many different data modalities in single cells across a range of experimental conditions will allow us to move beyond a transcriptome-centric cell view and learn a holistic representation of the cell. By studying the relationships between multimodal data types within single cells, we can begin to uncover the underlying basis for cellular functions and infer causal relationships between modalities. A major outstanding scientific and philosophical question in biology is 'what is a cell type?' If there is an answer to this question, it will be found through a nuanced analysis of single cells, taking into account different modalities and conditions, just as the age-old question of 'what is a gene?' must be answered through the comparative analysis of DNA sequences across species and a multimodal biochemical analysis.

Published online 29 January 2019

1. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
2. Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
3. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
4. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
5. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
References 4 and 5 are two of the first published high-cell-throughput droplet-based methods for scRNA-seq.
6. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
7. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
8. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, eaam8999 (2018).
9. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
10. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
11. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife* **6**, 1127 (2017).
12. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
13. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
14. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
15. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
16. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
17. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
18. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
19. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
20. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **9**, 2579 (2017).
This study presents a method for simultaneously measuring gene expression and proteins in single cells through an innovative barcoding strategy.
21. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **161**, 1202 (2017).
22. Faridani, O. R. et al. Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
23. Gomez, D., Shankman, L. S., Nguyen, A. T. & Owens, G. K. Detection of histone modifications at specific gene loci in single cells in histological sections. *Nat. Methods* **10**, 171–177 (2013).
24. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
25. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 1–6 (2017).
26. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
27. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
28. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
29. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
30. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
31. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
This study greatly increases the number of genes able to be spatially profiled in a single experiment through the development of combinatorial smFISH indexing and tissue clearing methods.
32. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
This is one of the first studies to simultaneously measure the transcriptome and cell lineage relationships.
33. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
34. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
35. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
36. Eberwine, J. et al. Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA* **89**, 3010–3014 (1992).
37. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
38. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
39. Hayashi, T. et al. Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its 'index sorting' function for stem cell research. *Dev. Growth Differ.* **52**, 131–144 (2010).
40. Wilson, N. K. et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **16**, 712–724 (2015).
41. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
This study performs index sorting coupled to scRNA-seq on myeloid progenitor cells and identifies transcriptional heterogeneity within sorted populations.
42. Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
43. Hochgerner, H. et al. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* **7**, 16327 (2017).
44. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
45. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
46. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
This study performs parallel DNA methylome and transcriptome sequencing in the same cell and examines the relationships between DNA methylation and gene expression.
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
48. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
49. Darmanis, S. et al. Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* **14**, 380–389 (2016).
50. Genshaft, A. S. et al. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* **17**, 1–15 (2016).
51. Di Giusto, D. A., Wlassoff, W. A., Gooding, J. J., Messerle, B. A. & King, G. C. Proximity extension of circular DNA aptamers with real-time protein detection. *Nucleic Acids Res.* **33**, e64 (2005).
52. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
53. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1873 (2016).
54. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
55. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR- pooled screens with single-cell RNA-Seq. *Cell* **167**, 1883–1888 (2016).
References 52–55 are the first to perform pooled genetic screens using CRISPR–Cas9 coupled to scRNA-seq to infer causal relationships in gene regulatory networks.
56. Klann, T. S. et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561 (2017).
57. Thakore, P. I., Black, J. B., Hilton, I. B. & Gersbach, C. A. Editing the epigenome: technologies for programmable transcription and epigenetic modulation. *Nat. Methods* **13**, 127–137 (2016).
58. Liu, X. S. et al. Editing DNA methylation in the mammalian genome. *Cell* **167**, 233–247 (2016).
59. Hilton, I. B. et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).
60. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).
61. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
62. Boettcher, M. et al. Dual gene activation and knockout screen reveals directional dependencies in genetic networks. *Nat. Biotechnol.* **36**, 170–178 (2018).
63. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.* **6**, 936–942 (2017).
64. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
65. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
66. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
67. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
68. 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).
69. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
70. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
This study develops a method of deriving the rate of change in gene expression from scRNA-seq data through the measurement of intronic RNA read abundance in each cell.
71. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
72. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
73. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
This study introduces the first method to order individual cells along a pseudotime trajectory.
74. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
75. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476 (2018).
76. Meng, C. et al. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
77. Argelaguet, R. et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
78. Colomé-Tatché, M. & Theis, F. J. Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.* **7**, 54–59 (2018).

79. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
80. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
This study pioneers the use of CCA to jointly reduce dimensionality for a pair of scRNA-seq data sets, allowing common cell states to be identified across data sets.
81. 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).
This study introduces the concept of using MNNs as a method for identifying equivalent cell states across data sets.
82. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
83. Dekel, T., Oron, S., Rubinstein, M., Avidan, S. & Freeman, W. T. in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* 2021–2029 (IEEE, 2015).
84. Hie, B. L., Bryson, B. & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data. Preprint at *bioRxiv* https://doi.org/10.1101/371179 (2018).
85. Barkas, N. et al. Wiring together large single-cell RNA-seq sample collections. Preprint at *bioRxiv* https://doi.org/10.1101/460246 (2018).
86. Park, J.-E., Polanski, K., Meyer, K. & Teichmann, S. A. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape. Preprint at *bioRxiv* https://doi.org/10.1101/397042 (2018).
87. Korsunsky, I. et al. Fast, sensitive, and flexible integration of single cell data with Harmony. Preprint at *bioRxiv* https://doi.org/10.1101/461954 (2018).
88. Stuart, T. et al. Comprehensive integration of single cell data. Preprint at *bioRxiv* https://doi.org/10.1101/460147 (2018).
89. Welch, J. et al. Integrative inference of brain cell similarities and differences from single-cell genomics. Preprint at *bioRxiv* https://doi.org/10.1101/459891 (2018).
90. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–198 (2017).
This study combines scRNA-seq and in situ hybridization data to predict spatial patterns of gene expression in the *Drosophila* embryo.
91. Tosches, M. A. et al. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
92. 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).
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. Regev, A. et al. Science forum: the human cell atlas. *eLife* **6**, e27041 (2017).
95. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
96. Alquicira-Hernandez, J., Nguyen, Q. & Powell, J. E. scPred: single cell prediction using singular value decomposition and machine learning classification. Preprint at *bioRxiv* https://doi.org/10.1101/369538 (2018).
97. Boufea, K., Seth, S. & Batada, N. N. Mapping transcriptionally equivalent populations across single cell RNA-seq datasets. Preprint at *bioRxiv* https://doi.org/10.1101/470203 (2018).
98. Wagner, F. & Yanai, I. Moana: a robust and scalable cell type classification framework for single-cell RNA-Seq data. Preprint at *bioRxiv* https://doi.org/10.1101/456129 (2018).
99. 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).
This study presents a method of aligning pseudotime trajectories developed from different data modalities as a way to compare pseudotemporal changes in each modality.
100. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
101. Scott, M. P. & Carroll, S. B. The segmentation and homeotic gene network in early *Drosophila* development. *Cell* **51**, 689–698 (1987).
102. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
103. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods* **10**, 1127–1133 (2013).
104. 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).
105. Shah, S., Lubeck, E., Zhou, W. & Cai, L. seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron* **94**, 752–758 (2017).
106. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
107. Moffitt, J. R. et al. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl Acad. Sci. USA* **113**, 14456–14461 (2016).
108. Moffitt, J. R. et al. Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
109. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
110. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
111. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
112. Achim, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
113. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
114. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
115. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).
116. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
117. Pandey, S., Shekhar, K., Regev, A. & Schier, A. F. Comprehensive identification and spatial mapping of habenular neuronal types using single-cell RNA-Seq. *Curr. Biol.* **28**, 1052–1065 (2018).
118. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
119. Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).
120. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Preprint at *bioRxiv*. https://doi.org/10.1101/459529 (2018).

Acknowledgements
This work was supported by the US National Institutes of Health through a New Innovator Award (1DP2HG009623-01) and an R01 (5R01MH071679-12) to R.S.

Author contributions
Both authors contributed to all aspects of the manuscript.

Competing interests
The authors declare no competing interests.

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