Trends in Biotechnology

CellPress

Review

Concepts and new developments in droplet-based single cell multi-omics

Arthur Chow [1,@], and Caleb A. Lareau [1,*,@]

Single cell sequencing technologies have become a fixture in the molecular profiling of cells due to their ease, flexibility, and commercial availability. In particular, partitioning individual cells inside oil droplets via microfluidic reactions enables transcriptomic or multi-omic measurements for thousands of cells in parallel. Complementing the multitude of biological discoveries from genomics analyses, the past decade has brought new capabilities from assay baselines to enable a deeper understanding of the complex data from single cell multi-omics. Here, we highlight four innovations that have improved the reliability and understanding of droplet microfluidic assays. We emphasize new developments that further orient principles of technology development and guidelines for the design, benchmarking, and implementation of new droplet-based methodologies.

Highlights

Single cell methods have proliferated in recent years, made feasible by droplet microfluidic technologies, including commercialized technologies.

Although the concept of single cell barcoding using these methodologies is straightforward, these assays require several assumptions about the underlying molecular biology for data interpretation.

Following the initial droplet-based single cell demonstrations, new experimental and bioinformatics methods have evolved the baseline expectations of these assays, producing more complex and higher quality data, including single cell multi-omics.

By generalizing themes from prior bioinformatics and experimental innovation, new users and experienced technology developers can identify opportunities to maximize data generation, improve analytical interpretation, and avoid failure modes in establishing new assays.

Lessons learned from droplet-based single cell genomics

Following the first demonstration of single cell sequencing in 2009 [1], single cell genomics assays have become widely used staples in the molecular biology toolkit, ushering in a new appreciation for molecular heterogeneity underlying cells in complex tissues [2]. A variety of approaches enable the partitioning and barcoding of individual cells, including microwells [3], split-pool combinatorial indexing [4], and droplet emulsions [5]. Although each of these methods in contemporary workflows can yield massive-scale, high-quality data, the most widely used single cell methods leverage **droplet microfluidics** (see Glossary) [6], whereby microfluidic instrumentation is used to create oil emulsions for the barcoding of cellular nucleic acids over thousands of reactions in parallel. Thus, we focus the scope of this review on concepts and developments in droplet-based technologies. In addition to ~100× greater throughput compared with prior plate-based methods, droplet-based single cell genomics provided a technical platform for multi-omics, including transcriptome [7–9], accessible chromatin [10,11], protein abundance [12,13], perturbation [14,15], and many combinations thereof [16,17]. For example, **cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)** quantifies surface protein expression and gene expression [12,17,18] via the covalent attachment of antibodies to oligonucleotides. Furthermore, **DOGMA-seq** [16], TEA-seq [19], and NEAT-seq [20] span all measurements of the central dogma (DNA, RNA, and protein).

The first methods to describe droplet-based single cell RNA sequencing (scRNA-seq), inDrop [8] and Drop-seq [9], established this key technological capacity, which has transformed single cell genomics. Both methods introduced massively parallel barcoding of single cells using oil emulsion droplets containing oligonucleotides comprising cell barcodes (14–16 bases) and **unique molecular identifiers (UMIs)** (8–12 bases), lengths that were selected to ensure molecular diversity without creating an onerous sequencing burden. Although the overall concept and design were similar, these assays differed in details that shaped the adoption and extensions of these single cell assays. For example, Drop-seq used hard resin beads that were limited by **Poisson loading** into droplets, whereas inDrop utilized soft hydrogel beads to achieve **sub-Poisson loading**, concentrating the distribution of beads per droplet closer to one per droplet, to enable higher

[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA

*Correspondence: lareauc@mskcc.org (C.A. Lareau).
@X: @arthurwchow (A. Chow) and @caleblareau (C.A. Lareau).

Box 1. Summary of assumptions and recent innovations in droplet-based single cell genomics

In an ideal droplet-based single cell technology, every cell would be individually partitioned into a droplet containing reagents needed to de-identify its nucleic acid content. Each cell would be associated with exactly one barcode, and nucleic acids from any other cell would not be present in the droplet emulsion. We articulate this idealistic framework because most downstream analyses rely on these assumptions and violations may confound the interpretation of existing and new single cell technologies. Fortunately, an appreciation for these technical details has catalyzed new methodologies.

First, experimental frameworks used to verify and enhance single cell measurements are outlined (see Figure 1 in the main text). The species-mixing experiment is the most common design, which provides a high-sensitivity measure for instances where two or more cells are contained within an individual droplet. In most experimental settings designed to profile new biology, only one species is prevalent in the experiment, requiring exogenous reagents or computational methodologies to identify and remove cell doublets. Further emerging approaches that purposefully overload cells into droplets and rationally deconvolve cell doublets can increase cell throughput by an order of magnitude in these workflows.

Second, we extend this concept of validating single cell measurements to detecting nucleotides that did not originate from the cell contained in the droplet (see Figure 2 in the main text). The degree of the ambient signal can be assessed from analyses of the knee plot and computationally mitigated. Furthermore, explicitly modeling the abundance of ambient molecules from empty droplets can create a high-quality transformation of raw count data using more sophisticated statistical methods.

Third, the other key ingredient within the droplet chemistry is analyzed: the bead conjugated with oligonucleotides. Both possibilities of multiple barcodes associating with the same cell through two or more beads encapsulated in the droplet or sequence heterogeneity on an individual bead, collectively termed 'barcode multiplets' (see Figure 3 in the main text). Barcode multiplets can be leveraged not only for new assays, such as bead overloading in microfluidic reactions, but also for benchmarking the robustness and reliability of single cell measurements in new assays.

Finally, we illustrate considerations for new single cell multi-omic assays (see Figure 4 in the main text). Specifically, to ensure proper interpretation of new analytes, there is an imperative to measure as close to the source as possible, a guiding principle underlying multi-omic method development. As motivating examples, we discuss the variable efficacy of strategies for measuring genetic perturbations and quantifying proteogenomic features alongside single cell genomics measurements from the cells.

Through these four vignettes, this review summarizes how idealistic assumptions can be relaxed, and how newer methods have enhanced these assays through a synthesis of computational and experimental methods development.

throughput of cell profiling. Notably, using soft hydrogel beads has similarly been commercialized and utilized by 10x Genomics [7], the most widely adopted commercial kits for single cell sequencing. The choice of beads is one of many concepts that has emerged and evolved over recent years of single cell genomics, including strategies for detecting multiple cells or beads in droplets.

Motivated by the evolving understanding of the nuances of droplet-based single cell sequencing, here we review four experimental and computational innovations that have improved the accuracy or expanded the scope of single cell genomics technologies. In particular, we reflect on vignettes where iterative method development has improved the state-of-the-art assays (summarized in Box 1 and part (A) of all figures in this article). We review how new methods have expanded beyond assay baselines, which have catalyzed new computational and experimental method development. These innovations have produced single cell assays with vastly improved signal-to-noise ratios, accurate cell yields, and complex multi-omic measurements. Furthermore, although we focus on cell-based measurements, we note that many analogous advances have been similarly described for profiling nuclei. In the years ahead, we envision continued innovation to stem from a similar appreciation of the technical principles underlying the evolution of droplet-based single cell technologies.

Validating and expanding single cell measurements

Baseline: at most, one cell must be loaded into a droplet for accurate single cell profiling.

New developments: computational and experimental strategies can identify droplets with more than one cell and remove doublets, increasing cell throughput per experiment.

Glossary

Ambient molecules: cell-free nucleic acids (typically RNA) in the solution where cells are suspended. These nucleotides can become encapsulated in droplets with cells and barcoded alongside nucleic acids from those cells. The abundance of ambient molecules can vary widely between tissue sources and dissociation protocols [39].
Antibody-derived tags (ADTs): oligonucleotides covalently attached to antibodies for proteogenomic characterization, including surface marker expression via CITE-seq [12] or cell hashing by targeting B2M [24].
Apt-seq: assay that allows for the simultaneous profiling of transcriptomes of cells and surface proteins using aptamers, which are nucleic acid probes that can bind to specific target epitopes [63]. Critically, the same molecule (nucleic acid) that does the binding is what is barcoding during the droplet-reverse transcription (RT) step.
ATAC with selected antigen profiling by sequence (ASAP-seq): single cell multi-omic assay that profiles surface or intracellular protein abundance alongside accessible chromatin and mitochondrial DNA via a droplet-based ATAC-seq reaction [16]. A critical innovation in developing this workflow was a bridge oligo, which allowed for repurposing existing CITE-seq reagents for compatibility with a different capture sequence for the bead oligos on the ATAC gel beads.
Barcode multiplets: instances where a cell is barcoded by two or more oligonucleotides due to multiple barcodes on a single gel bead or from multiple beads in the same droplet.
Bead-based ATAC processing (bap): computational approach for identification of barcode multiplets using the base-pair resolution abundance of Tn5 insertion sites shared between pairs of single cell barcodes [10,49].
Bridge oligo: oligonucleotide used to facilitate the barcoding of a feature in a droplet-based single cell reaction. Used in the ASAP-seq workflow [16], the bridge oligos allow the capture sequence in the ATAC beads to barcode reagents developed for CITE-seq.
Cell Hashing: approach for increasing the cell throughput of a single cell run while controlling for the doublet rate [24]. Antibodies recognizing conserved proteins, such as B2M, are conjugated

Species mixing experiments

When establishing a new assay, an essential validation is that the intended analyte is accurately measured. Intrinsic to their name, single cell technologies profile the molecular contents of individual cells, which requires an experimental system to readily distinguish whether one, two, or more cells are profiled within the individual droplet reaction (Figure 1A). As throughput increases, more cells are processed in parallel, loading a higher density of cells into the system (e.g., in droplet-based platforms, such as the 10x Genomics Chromium). Consequently, higher cell densities increase the probability that two or more cells will be encapsulated in a single droplet.

In this sense, **species mixing** experiments, often using a combination of human and mouse cell lines, are a gold-standard technique for benchmarking and quantifying cell doublets [21], that is, artifacts where two or more cells are mistakenly encapsulated together (Figure 1B). Since the cells from different species have distinct genetic sequences, any doublet-containing cells from both species (a **heterotypic doublet**) can be identified by their mixed-species expression profile (Figure 1A). Originally introduced in Drop-seq for droplet-based single cell sequencing [9], the authors reported that 0.4–11% of cell barcodes were cell doublets through analyses in a 'barnyard plot' (Figure 1C), and the rate varied as a function of cell loading under standard statistical assumptions. While human–mouse doublets are readily identified analytically, **homotypic doublets** cannot be discriminated, requiring a *post hoc* correction from the heterotypic doublet rate to report the true frequency of doublets [9]. Mixing human and mouse cell lines at 50:50 ratios to quantify the doublet rate results in 50% heterotypic and 50% homotypic doublets and has been used for benchmarking most droplet-based assays, including single cell assay for transposase accessible chromatin by sequencing (scATAC-seq), where genomic DNA from human and mouse cells can similarly be used to detect doublets [10,11].

Increasing throughput while controlling doublet rate

Although species-mixing experiments are conducted when initially validating and benchmarking new assays, the design is rarely compatible with experiments aiming to study the molecular heterogeneity of complex tissues typically derived from a single organism. Furthermore, because cells are Poisson loaded into droplets, meaning the number of cells per droplet follows a Poisson distribution, ~80–90%+ of droplets lack cells following the standard guidelines for single cell sequencing [22]. Given that many studies have demonstrated the utility of uncovering cellular heterogeneity by profiling more cells, new approaches have been developed to increase the number of cells per experiment (i.e., cell throughput). Thus, **droplet overloading** is a common feature whereby many more cells are loaded into the droplet microfluidic device than recommended, increasing the number of cells barcoded while controlling the doublet rate, which limits the erroneous interpretation of faulty single cell states. A common method for cell doublet detection utilizes exogenous diverse barcodes that are co-detected alongside cellular features in single cell assays [23]. Common formats involve either an oligo-conjugated antibody via **Cell Hashing** [24] or an oligo-lipid conjugate in **MULTI-seq** [25] (Figure 1D). Following single cell encapsulation, amplification, and sequencing, **antibody-derived tags (ADTs)** can be enumerated per cell, resulting in a barcode-by-cell matrix, and subsequent analyses allow for each cell to be assigned to an original sample. Conceptually, these methods can control the rate of doublets when increasing the cell throughput by identifying droplets with two or more distinct hash barcodes (Figure 1E,F). The droplets identified as doublets are filtered from the analysis, resulting in a reduced doublet rate at the expense of discarded data. Under previously described workflows and conditions [24], the combined experimental and computational framework of doublet removal can increase the throughput of *bona fide* singlets by nearly an order of magnitude for an equivalent doublet rate (Figure 1E).

with a diverse set of oligonucleotide barcodes and can be used for sample multiplexing in addition to increasing the number of cells profiled per experiment.

Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq): high-throughput, single cell multiomic technology for quantifying mRNAs and surface proteins via oligo-conjugated antibodies [12]. Successful applications of CITE-seq measure transcriptome-wide gene expression alongside up to 220 or more surface proteins [18].

Close-packed ordering: method used to surpass the Poisson distribution of loading beads into droplets, primarily for increasing the abundance of droplets with exactly one bead to 90% or more [7,83]. This is applicable for deformable hydrogel beads (used originally in InDrop [8], among other assays) but not hard resin beads (used in Drop-seq [9]).

CRISPR droplet-sequencing (CROP-seq): introduced a new vector design that allowed for the direct detection of the gRNA protospacer [15], rectifying a limitation of the original Perturb-seq workflow [14].

Denoised and scaled by background (dsb): method that transforms integer count data from proteogenomic assays, such as CITE-seq, and produces a more sensitive measure by using the occurrence of ambient molecules barcoded by droplets that are not annotated by cells [48].

DOGMA-seq: assay that profiles all elements of the central dogma, including accessible DNA, RNA abundance, and protein abundance, via droplet-based single cell multiomics [16].

Droplet microfluidics: key technology in single cell genomics profiling; a microfluidics device can be engineered to create oil–water emulsions in which individual cells are partitioned for barcoding nucleic acids for single cell genomics.

Droplet overloading: attempts to profile cells beyond the manufacturer's recommendations, often leading to two or more cells per droplet. This workflow is accompanied by exogenous barcoding or combinatorial indexing of nucleic acids upstream of loading the microfluidic reaction.

Droplet-based combinatorial indexing: or combinatorial pre-indexing; inclusion of a diverse barcode upstream of the microfluidic instrument, allowing for the overloading of cells into

Computational detection of cell doublets

In addition to the physical techniques for identifying cell doublets, methods such as DoubletFinder [26] and scrublet [27] have been developed to identify cell doublets under specific data assumptions. Each requires a supervised set of parameters, including the expected doublet rate, typically estimated from prior species mixture experiments underlying the technology. These tools then generate synthetic doublets by randomly merging gene expression profiles from pairs of cells, creating a set of true-positive doublets (Figure 1G). The real and synthetic data are jointly analyzed in a neighbor graph, typically computed in a reduced dimensionality space. Based on the fraction of synthetic doublets that are the nearest neighbors, each cell is assigned a score proportional to the likelihood of being a heterotypic doublet (Figure 1H). Related methods have similarly been developed for scATAC-seq analyses, including ArchR [28] and Amulet [29]. Ultimately, the inference of doublets can be useful in some settings. An important limitation of computational doublet detection is that these approaches cannot identify homotypic doublets (Figure 1A) or cells of closely related cell types. Although these homotypic doublets remain in downstream analyses, straightforward statistical extensions allow for estimating the number of remaining doublets [27].

Donor-mixing experiments

Another increasingly common approach for increased throughput without modifying the single cell reaction is using distinct donors and **genetic demultiplexing** [30] to identify and filter cell doublets (Figure 1I and Box 2). Although demuxlet [31] achieves doublet identification and donor assignment with remarkably high accuracy, this tool requires donor genotypes to be established prospectively. Other approaches, including SouporCell [32], freemuxlet (available as a preprint [33]), and scSplit [34], can identify distinct clusters of genetic variants from single cell-sequencing data directly, mitigating the requirement for genotypes to be known and broadening the use of donor demultiplexing and doublet detection. These donor-mixing studies have become increasingly popular in single cell expression quantitative trait loci studies to resolve the impact of genetic variation on gene expression within a subset of cells [35].

Inline barcoding

The doublet-detection methods discussed thus far identify droplets with evidence of heterotypic doublets, discarding those data and limiting efficiency at higher levels of droplet overloading. This deficiency motivated the development of new methods that utilize inline barcoding, termed **droplet-based combinatorial indexing**, where a cell barcode is defined by the concatenation of diverse sequences from a split-pool step in addition to the droplet-derived oligos (Figure 1J).

The combination of a split-pool and droplet barcode allows for the unambiguous assignment of nucleic acids to single cells, even when multiple cells are encapsulated within the same droplet. Thus, instead of discarding droplets identified with two or more cells, the nucleic acids in these droplets are deconvolved into their constituent cells (up to some lower doublet rate). These workflows have been demonstrated independently for each of scATAC-seq (accessible chromatin) [10], scRNA-seq (transcriptome) [36], and oligo-based antibody detection (proteo-genomics) [37] (Figure 1J and Box 2). Given that each approach challenges the conventional approach of one cell per droplet, species-mixing experiments were critical in establishing the single cell nature of these assays. We expect new methods will benchmark their single cell performance with a similar mixing framework.

Lessons learned: applications of single cell technologies must consider the frequency and proportion of cell doublets for valid inference. Typically, new assays are benchmarked with species-mixing experiments to quantify the doublet rate and assume that these measurements generalize

droplets because individual cells can be de-identified from a combination of the bead barcode and the pre-index barcode. Successful applications of this approach have been demonstrated for accessible chromatin [10], RNA [36], and proteins [37].

Empty droplets: droplets that do not receive a cell during the single cell assay. All nucleic acids barcoded in these droplets are ambient, which can be used for statistical modeling to subtract the ambient RNA signal from cells, including RNA molecules (as in CellBender and SoupX [45,46]) as well as multiomic settings including protein signals [48] and DNA [44].

Genetic demultiplexing: approach to increase cell yields in single cell workflows by leveraging genetic differences between individuals; cells barcoded by bead oligos can be assigned as a heterotypic doublet by specifying individual donor genotypes [31] or without prior knowledge [32,34].

Heterotypic doublet: instances where two sufficiently distinct cells can be readily distinguished by a doublet detection algorithm (e.g., different species, a unique donor, or different cell types).

Homotypic doublet: instances where two cells of the same cannot be readily distinguished by any doublet detection algorithms (e.g., same species, donor, or cell type).

MULTI-seq: scRNA-seq and snRNA-seq sample multiplexing using lipid-tagged indices [25] has the advantage of utilizing lipid-based incorporation into cells or nuclei for multiplexing, whereas other multiplexing approaches, such as Cell Hashing [24], require protein-based detection via a barcoded antibody.

Perturb-seq: emerging single cell multi-omics assay that couples gRNA detection with single cell transcriptomics. In the original design of Perturb-seq [14], a separate barcode was linked to the gRNA protospacer, but was shown to be prone to mismatches during lentiviral packaging. The vector used in CROP-seq [15] enables direct barcoding of the protospacer, eliminating the need for the proxy barcode and improving data quality.

Phage-ATAC (PAC)-tag: tag containing the Illumina Read 1 sequence (Rd1) required for adding the single cell bead oligo barcode to the nanobody CDR3 sequence in the droplet amplification reaction [62].

to other contexts. Emerging methods of overloading droplets utilize diverse methods to increase cell throughput while controlling doublet rates.

Leveraging ambient (non-cell) measurements

Baseline: all nucleic acids barcoded within a droplet are aggregated and assigned to a specific cell.

New developments: statistical models can leverage droplets without cells to estimate **ambient molecules**, which can be used to regress out potential contamination.

Ambient nucleic acids

Upstream of the droplet microfluidics steps of single cell profiling, cells or nuclei are typically isolated from tissues, blood, or a similar heterogeneous environment. Even in high-viability settings, some debris, comprising nucleic acids from dead cells among other biological analytes, will be present in these mixtures. From free-floating nucleic acids, ambient molecules (often RNA) will nonspecifically associate with the contents of viable cells in the microfluidic reaction (Figure 2A) [38]. Microfluidics workflows then barcode the nonspecific nucleic acids (Figure 2A, gray) and those from the captured cell (Figure 2A, blue), introducing noise and potential artifacts in the data. At a basic level, abundant ambient RNA will contaminate the true gene expression profile, leading to noise in dimensionality reduction, clustering, and annotating cell types [39]. More critically, high ambient RNA can create artificial differences between conditions from differential expression analyses due to variations in ambient RNA levels rather than underlying biology (Box 3) [40]. Ambient nucleic acids can be particularly problematic in single nuclei RNA-seq (snRNA-seq) settings, including tissues [41] and tumors [42].

Ambient molecules revealed through the knee plot

A noteworthy feature of droplet microfluidic single cell profiling is that the quantification of ambient molecules can be achieved with no special experimental design or modification to the technology. Specifically, estimates from the first generation of the 10x Genomics Chromium chemistry indicated that ~500 000 droplets will contain a bead for amplifying nucleic acid, but only ~10 000 of these droplets contain a cell under recommended loading conditions [7]. Consequently, hundreds of thousands of **empty droplets** in a single cell genomics reaction will contain only ambient molecules when molecules are sequenced for these barcodes. Although improved in subsequent technology updates, the principle of an experiment yielding at least an order of magnitude more empty droplets than contained within a cell remains [43,44]. Hence, the quantity and identity of ambient features can be identified through analyses of the knee plot, a rank-ordered arrangement of the number of molecules barcoded by each bead barcode (Figure 2B). We provide examples of libraries with high and low contamination from knee plots where the abundance of unique reads that separates the two plateaus represents the separation in nucleic acid content between droplets containing cells or ambient molecules (Figure 2B).

Leveraging ambient molecules to improve data quality

Knee plot analyses are typically the first step in identifying which barcodes contain cells. When an experiment focuses on well-established major cell types (e.g., cell line-mixing experiments), a generic cell-calling algorithm that considers the marginal distribution of UMIs per barcode can sufficiently discriminate cell-containing droplets from empty droplets. In experiments to profile rare cell subtypes, advanced methods that analyze the full barcode-by-feature matrix to identify real cells with low RNA content and further correct for background contamination can recover rare cell types that would otherwise be discarded [43]. Motivated by this concept, EmptyDrops improves the discrimination of droplets containing cells from those lacking cells by using parametric

Poisson loading: distribution of the abundance of cells or beads per droplet in a single cell experiment. The count data follows a Poisson distribution, where the mean number of features per droplet equals the variance. Under an optimal loading of one bead per droplet, $\lambda = 1$, resulting in 36.8% of droplets with zero or one bead each and 26.4% of droplets with two or more beads.

Species mixing: experimental setup designed to benchmark the occurrence of two or more cells profiled in an assay. Human and mouse cells are often pooled together before running the single cell protocol.

Sub-Poisson loading: distribution of beads per droplet in a single cell experiment when combined with close-packed ordering, where the variance is much lower than the mean. The empirical distribution is skewed to where most droplets have exactly one bead loaded per barcode. For example, one droplet-based technology reported ~15% of droplets with zero beads, ~80% with one bead, and 5% with two or more beads per droplet [7].

Unique molecular identifier (UMI): in the context of scRNA-seq, UMIs provide complex barcode diversity before PCR amplification steps, enabling the identification of PCR duplicates, and thereby yielding accurate gene expression count values for downstream bioinformatics analyses. In most scRNA-seq applications, UMIs are 8–12 random nucleotides.

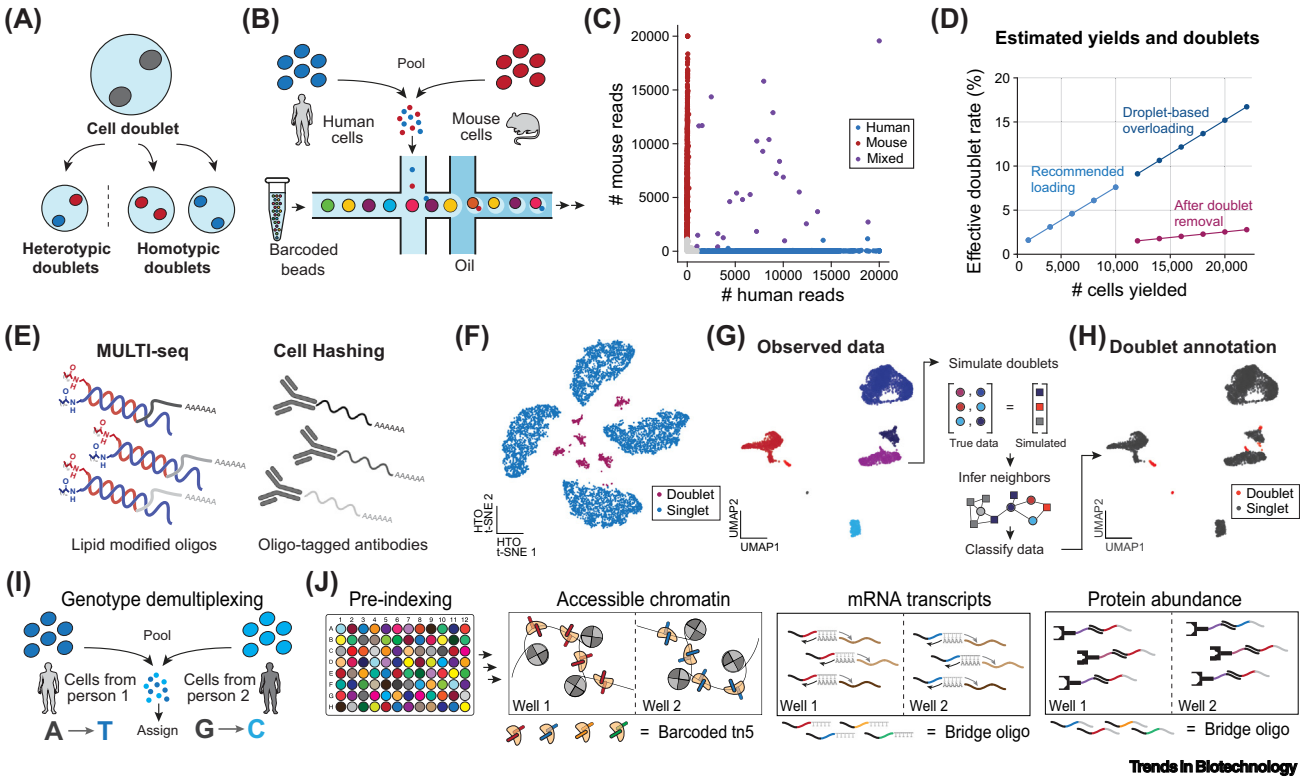

Figure 1. Approaches for validating and optimizing single cell measurements. (A) Representations of potential cell doublets from mixing experiments in droplet-based assays, including homotypic and heterotypic doublets. (B) Schematic of species-mixing experiments to validate single cell measurements. (C) Example of typic mixing experiment highlighting high-confidence human, mouse, and mixed doublet cells. (D) Estimates of cell yields and doublet rates under recommended, overloading, and after doublet detection and removal (purple). Rates are estimated from previous species-mixing estimates [7] and Cell Hashing inferences of a six-plex hashing experiment [24]. (E) Schematic of oligo-based methods for doublet-detection, including MULTI-seq [25] and Cell Hashing [24]. (F) Low-dimension embedding of data from oligo-based hashing methods to identify cell doublets from a real hashing experiment [62]. (G) Example of low-dimensional embedding of single cell data from cell state measurements [e.g., single cell (sc)RNA or single cell assay for transposase accessible chromatin (scATAC-seq)] used to infer potential heterotypic doublets via simulated doublets and neighbor graph analyses. (H) The result of the computational procedure described in (G) where predicted heterotypic doublets are classified in the high-dimensional space. Doublet calls (via scrublet) and clusters were generated from real scRNA-seq data [84]. (I) A general approach for overloading cells into droplets and identifying doublets via donor genotypes. (J) Schematic of approaches for combinatorial pre-indexing of accessible chromatin [10], RNA [36], and protein [37].

assumptions about ambient RNA or DNA molecules [43,44], and cell calling is determined based on a likelihood function rather than on the simple aggregate of all UMIs. Given that cell types may contain variable abundances of RNA molecules, a more sophisticated model can discriminate cells from droplets containing a surplus of ambient UMIs [43].

In addition to better discriminating true cells from background molecules, recent methods have been developed to regress the ambient count signal from the cell profiles, mitigating issues related to errant molecular barcoding. For example, CellBender [45], SoupX [46], and DecontX [38] estimate and remove the contribution of ambient RNA from cell profiles, thereby refining the gene expression interpretation for individual cells (Figure 2B and Box 3). Analogous tools, such as DIEM [47], utilize semi-supervised machine learning to remove ambient RNA contamination from snRNA-seq data. Although the benchmarking of each of these methods demonstrated clear utility in real-world data sets, there is a trade-off between removing true noise (sensitivity) and retaining true signal (specificity). To our knowledge, no systematic framework exists for determining ideal methods or hyperparameters, and proper ambient signal removal likely involves manual curation and validation via known marker genes in a data set-specific manner.

Box 2. Methods for identifying cell doublets during microfluidic overloading

Infecting cells into droplets for single cell reactions follows a Poisson distribution. As the multiplicity of infection increases (i.e., the number of cells loaded), so too does the number of droplets with two or more cells, which produces an invalid single cell measurement by definition. Thus, under manufacturers' guidelines, cells are loaded at low concentrations, but lowering doublet rates also increases the number of empty droplets that do not participate in the barcoding reaction. Alternatively, more cells can be loaded into droplets and heuristics then used to identify doublets and remove them from downstream analyses.

Orthogonal barcodes for doublet detection

Orthogonal sequences are designed to facilitate retention of the exogenous sequence throughout the microfluidic reaction. For example, MULTI-seq [25] utilizes lipid-tagged barcodes that embed into the cell membrane, enabling the labeling and pooling of a range of cell types, including those without well-characterized surface proteins for antibody binding. This method is particularly advantageous for cells not amenable to Cell Hashing [24], which labels cells via an antibody that recognizes a highly expressed surface protein (e.g., B2M) with a diverse set of oligonucleotides, termed hashes. A limitation of these methods has been the frequency of 'negative' cell barcodes, whereby no hash barcode is detected at insufficient levels to assign the barcode to a corresponding hash identifier.

Genetic demultiplexing

The same input material can be pooled together from different donors before loading into the droplet microfluidic device. After sequencing, cell doublets where each cell is derived from a different donor (in this case, heterotypic doublets) can be readily identified and filtered using rigorous computational approaches. For example, demuxlet [31] utilizes a small set of SNPs that are well detected by scRNA-seq techniques and defined for each donor *a priori*. With these data as input, demuxlet uses a mixture model and maximum likelihood estimation to determine the most likely donor for each droplet in a given sample, including whether the droplet contained cells from two different donors [31]. Given that the probability of two cells coming from the same individual decreases linearly with the number of donors pooled together, common strategies include eight or more individual donors per batch [35]. In situations where genotype references are unavailable or incomplete, Vireo uses a Bayesian approach to demultiplex donors in pooled samples instead of a genotype reference but achieves high-confidence classification [71]. For unrelated donors, these methods report the incorrect assignment only ~1–2% of the time [31]. However, multiplexing multiple samples from the same donor or closely related donors would invalidate this approach and lead to a higher false positive rate.

Combinatorial pre-indexing

Each workflow involves introducing one or more rounds of unique barcoding to label nucleic acids in cells from eight to 96 wells, where each well will contain a different pre-barcoding sequence. Subsequently, cells are pooled together before loading into the microfluidic instrument, ultimately allowing multiple cells per droplet to be de-identified via a combination of pre-index and droplet barcodes, termed '**droplet-based combinatorial indexing**'. In addition to these three approaches, which introduce a split-pool barcode upstream of droplet microfluidic capture, preliminary data have emerged from a new framework called 'Overloading And unpacKing (OAK)', which appends a barcode after droplet barcoding [72] to increase cell throughput by orders of magnitude.

Similar to the interference caused by ambient RNA, noise from multimodal single cell-profiling techniques, including oligo-conjugated antibodies to measure protein expression [e.g., cellular indexing of transcriptomes and epitopes (CITE)-seq], can obscure true biological variation within a data set. Although normalization approaches have been developed for proteo-genomic ADT data, **denoised and scaled by background (dsb)** [48] provides a rational normalization to mitigate ambient signal in proteo-genomic assays (Box 3). The utility of this model-based transformation is revealed through the quantitative output of the ADT data, including the delineation of proteins that express intermediate values of surface antigens, such as CD4 expression in monocytes (Figure 2D). In other words, the raw count data from proteogenomic methods, such as CITE-seq, is influenced by nonspecific binding, and models that account for ambient molecules in empty droplets can mitigate noise and improve quantitative signals in these multimodal assays.

Lessons learned: not all nucleic acid barcodes arise from a cell in a droplet, and this rate of ambient nucleic acid can vary between experiments. Analyses of the empty droplets can identify ambient molecules, which can be regressed out of the cell-by-feature matrix or leveraged in more

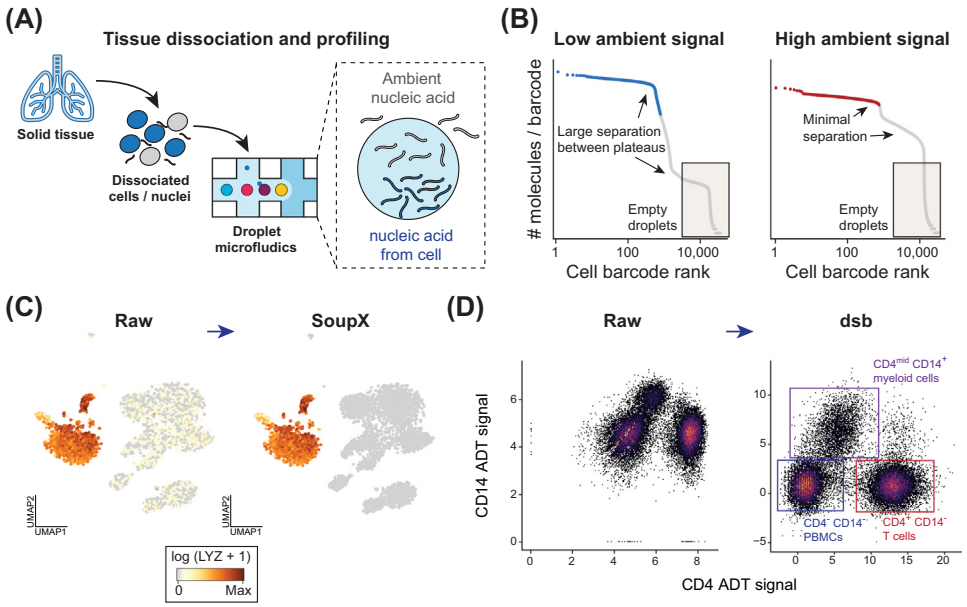

Figure 2. Quantifying and mitigating noise from single cell assays. (A) Schematic of RNA molecules captured in a droplet microfluidic reaction, including ambient RNA molecules, which can arise from cell/nuclei dissociation of tissues before droplet microfluidics and single cell genomics. (B) Knee plots of exemplar data comparing single cell samples with low or high ambient signal. Droplets with no cells are highlighted in the gray box. Knee plots were adapted from preliminary data from [85]. (C) Summary of *LYZ* expression from raw single cell data (left) compared with expression accounting for ambient RNA via SoupX [46]. (D) Proteogenomic signal before (left) and after (right) modeling of ambient signal with denoised and scaled by background (dsb) [48], highlighting $CD4^{mid}$ myeloid cells (purple). (C,D) reproduced from source code from [46,48] via a Creative Commons License. Abbreviation: ADT, antibody-derived tag.

sophisticated models. New assays and computational methods that consider measurements beyond those in the barcodes identified as cells can improve the sensitivity and specificity of single cell measurements.

Barcoding the contents of a cell only once (unless you do not want to)

Baseline: at most one oligo-conjugated bead should be present per droplet for single cell barcoding.

New developments: computational approaches to detect **barcode multiplets** can improve the data quality of existing assays, and relaxing the one barcode for one cell assumption can enable assay development.

Sources of barcode multiplets

A common assumption in single cell assays is that a single oligonucleotide barcode represents the nucleic acid of an individual cell. Given that the previous two sections discussed the benchmarking and quantification of contamination of cellular nucleic acids (either from cell doublets or ambient molecules), we now consider the complementary scenario where an individual single cell has its contents barcoded by multiple distinct oligonucleotides, termed barcode multiplets. These events can arise from two potential scenarios: (i) a single droplet containing multiple beads; or (ii) multiple oligonucleotides present on the same individual bead (Figure 3A). Every single cell assay has barcode multiplets to varying degrees. Although this artifact is unlikely to have invalidated any biological conclusions, the presence of barcode multiplets can create challenges and opportunities in establishing and benchmarking assays.

Box 3. Removing and utilizing ambient molecules for normalization

If ambient nucleic acids stem from a highly abundant cell type, the annotation of biologically relevant features can be obscured for lowly abundant cell types, such as the percent of cells positive for a feature [73]. Although the abundance of ambient molecules will vary between different tissues and upstream preservation/processing workflows [39], the complete elimination of ambient nucleic acid in library preparation is unavoidable in most contexts, motivating computational approaches to mitigate their impact on downstream analyses.

Regressing out ambient RNA counts

Multiple methods now account for ambient RNAs captured randomly and use statistical analyses to mitigate their impact on downstream interpretation. For example, SoupX first assembles a profile of ambient RNA molecules from the **empty droplets** in an experiment (highlighted in Figure 2B in the main text) and estimates the fraction of ambient molecules attributed to each gene. Whereas SoupX directly uses a relatively simple heuristic to subtract the impact of ambient expression, newer approaches, such as CellBender, allow for more complicated models of ambient expression, albeit at a high computational cost [45]. Using a cluster-aware model, SoupX adjusts the raw cell count data by regressing out the ambient RNA profile from the empty droplets, vastly improving the specificity of marker genes. For example, *LYZ*, a marker gene of myeloid cells in peripheral blood mononuclear cells, was highly expressed in empty droplets and resulted in errant *LYZ*$^{+}$ B and T cells (see Figure 2C in the main text) [46]. After correcting this experiment with SoupX, the expression of *LYZ* was restricted to the myeloid compartment, as expected, validating the method. Methods such as DropletQC [74] and SiftCell [75] implement quality control checks and statistical analyses to evaluate the integrity of droplets to ensure that the final data set used for downstream analyses is of high quality and representative of single cells.

Modeling protein tags in empty droplets

The dsb [48] model addresses two sources of noise in protein expression data: (i) protein-specific noise, which stems from ambient unbound ADTs captured during the creation of droplets; and (ii) cell-specific technical variance, evidenced by shared variances linked to isotype antibody controls and background protein counts within each cell. Using a rationally motivated model incorporating both noise terms, dsb similarly leverages the abundance of empty droplets to estimate these parameters, resulting in a transformed ADT signal that allows for quantitative interpretation when comparing cells in an experiment. One limitation of dsb is the reliance on isotype control antibodies, which may be absent from some panels, resulting in the limited performance of data normalization.

Detection of barcode multiplets in scATAC-seq data

The first software to detect barcode multiplets in droplet-based single cell data was **bead-based ATAC processing (bap)**. In brief, this workflow quantifies the degree of overlap of transposition events between pairs of bead barcodes to identify instances where two barcodes share more transposition events than expected, leading to a highly sensitive and specific classification of barcode multiplets [10,49]. When applying bap to scATAC-seq data sets released from 10x Genomics, instances of nine or more bead barcodes were annotated by bap of belonging to the same droplet (all barcodes annotated as 'cells' having a sequence in the whitelist and read abundances above the knee threshold) [49]. These barcodes were enriched in the same biological cluster, supportive of the idea that they may be 'replicates' of the same cell (Figure 3B, top). Examining the underlying barcode sequence, a common seven-nucleotide variable region followed by a nine-base constant in these barcodes (Figure 3B, bottom) were discovered, consistent with the idea that multiple 'whitelisted' barcode sequences were installed on an individual bead during synthesis.

Although bap was the first software solution to identify barcode multiplets, a new implementation from 10x Genomics was incorporated into CellRanger-ATAC v1.2 that similarly analyzes shared transposition events between pairs of barcode sequences to identify multiplets (https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/release-notes). Whereas bap merges different bead barcode sequences into a single cell barcode, the implementation in CellRanger-ATAC simply discards multiplet barcodes after retaining the one bead barcode with the highest number of fragments. Compared with v1.1, which did not consider barcode multiplets, there was a substantial difference in the number of cells called for the same data set. For example, a publicly available species-mixing data set analyzed by 10x Genomics resulted in a loss of 2347 of 11 721 cells called between subsequent versions of the software (Figure 3C). In other words, the

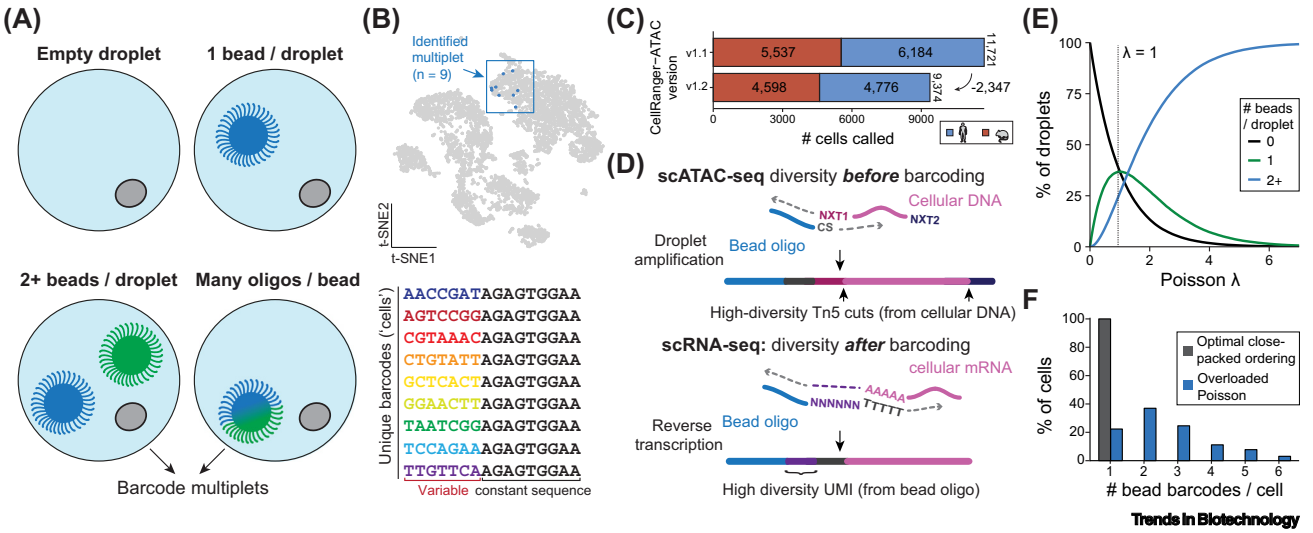

Figure 3. Detection and impact of barcode multiplets in single cell assays. (A) Schematic of potential bead/barcode outcomes assuming that a single cell or nucleus is loaded into a droplet, highlighting two instances where barcode multiplets can arise. (B) Example of a barcode multiplet where nine valid single cell barcodes (all passing the knee cutoff) marked the same cell (top). Based on analysis of the underlying barcode sequence (bottom), this multiplet originated from a bead with many distinct oligonucleotide barcodes associating with an individual cell [49]. (C) Impact of barcode multiplets on cell count from a selected single cell assay for transposase accessible chromatin sequencing (scATAC-seq) experiment (same data set). Version v1.1 of the CellRanger-ATAC software did not consider barcode multiplets, whereas v1.2 retained the barcode with the highest read count, discarding other barcodes. (D) Schematic of key nucleic acid barcoding steps for scATAC-seq (top) and single cell (sc)RNA-sequencing (bottom), noting that the diversity of molecules is variable between the two methods. (E) Distributions of beads per droplet under Poisson assumptions of the mean barcodes per droplet (defined by λ). The theoretical optimum for % droplets with exactly one bead is highlighted at λ = 1. (F) Summary of bead abundance per cell following computational bead oligo merging, including cells with two or more bead oligos. Data for (B,C) adapted from [49] and for (F) adapted from [10]. Abbreviation: UMI, unique molecular identifier.

loss of nearly 20% of barcodes called as cells was based solely on the same data, reflecting the magnitude of the artifact. Barcode multiplet detection is still utilized in contemporary versions of CellRanger-ATAC and CellRanger-ARC (for processing multiome data), limiting the impact of multiplets on data derived from these commercial kits when processed through the official software. However, the marked change in cell yields due to this artifact (Figure 3C) warrants consideration in new assay development.

Although multiple algorithms can now detect barcode multiplets for scATAC-seq, no such software exists for scRNA-seq due primarily to an intrinsic difference in the assay barcoding. Namely, the base-pair transposition events from scATAC-seq allow for a high-diversity identifier for each fragment (Figure 3D). As a result, when two bead oligos are present in the same droplet, they will amplify the same molecule at a low, but detectable rate. These shared amplifications can be identified across the full set of molecules in an experiment due to the diversity of transposition events (for more discussion and algorithm performance, see [10,49]). Conversely, for scRNA-seq, diversity per molecule is in the form of UMIs, which are added in the reverse transcription (RT) step. As a consequence, two different bead oligos would not share UMIs even if they amplify the same original mRNA molecule (Figure 3D). This intrinsic difference in the assays allows for the sensitive and specific detection of barcode multiplets in scATAC-seq but not in scRNA-seq.

Leveraging barcode multiplets for method development

Although barcode multiplets confound the one bead–one droplet assumption, their occurrence can be leveraged when benchmarking and developing new assays. For example, in establishing the droplet single-cell (dsc)ATAC-seq assay [10], barcoded beads were incompatible with **close-packed ordering**. Indeed, the implementation of dscATAC-seq and the development of bap

produced a workflow with near-perfect detection and merging of barcode multiplets [10], resulting in most cells being barcoded by two or more beads in this assay (Figure 3E,F). Annotated barcode multiplets can provide a built-in form of replication that can be leveraged to assess assay or computational performance, including when developing and internally benchmarking three other assays (Box 4). Using barcode multiplets as internal validations is conceptually similar to preliminary bioinformatics approaches, such as molecular cross-validation [50] or self-supervision [51], which have been used to benchmark technical or algorithmic measurements. To our knowledge, barcode multiplets have not been analyzed in the context of ambient molecules (Figure 2), although these artifacts likely coexist in nearly every single cell experiment, which may enable more sophisticated models of ambient nucleic acid estimation and subtraction.

Lessons learned: most single cell analyses assume that each cell has all its nucleic acids tagged by one unique barcode. This assumption can be violated either by accident (impacting perceived cell number) or on purpose for assay development (e.g., overloading microfluidic devices with many beads). Utilizing this technical nuance of single cell genomics data can enable new dimensions to assay development or interpret existing data.

Measuring the many facets of a cell simultaneously

Baseline: one 'omic measurement (e.g., RNA or chromatin) is measured per assay.

New developments: strategies for co-opting in-droplet chemistry can measure additional modalities, but careful consideration is required when measuring a barcoded surrogate of the intended analyte.

Multi-omic detection

Although the most common applications of single cell genomics remain single modality measures of either transcriptome or accessible chromatin, new methods have emerged that facilitate the

Box 4. Leveraging barcode multiplets for assay development

Under standard assumptions for single cell sequencing, barcode multiplets can confound interpretation because the same cell will be profiled by multiple barcodes. However, barcode multiplets have also been utilized in developing and validating new droplet-based assays. We note some of these examples here.

Bead overloading for increasing cell capture

Typical Poisson loading of beads into droplets requires the balance of empty droplets (where zero beads would be loaded into droplets) versus barcode multiplets (where two or more beads would be loaded per droplet). Under Poisson statistics, the maximization of exactly one bead per droplet occurs at a loading of $\lambda = 1$, resulting in 36.8% of droplets receiving zero beads, 36.8% of droplets receiving one bead, and 26.4% of droplets having two or more beads (see Figure 3E in the main text). However, assuming that barcode multiplets could be detected and subsequently merged, beads could be loaded at a higher concentration to increase λ to 3, resulting in <5% of droplets receiving zero beads under Poisson loading conditions.

Use in benchmarking and validating in assay development

In HyPR-seq [76], the authors used hard resin beads from the Drop-seq protocol to establish an assay for detecting selected genes via RNA-based probe detection. The authors utilized the fact that UMIs were encoded in the probes for gene detection, thus enabling two different bead oligos to amplify the same (diverse) gene probe during in-droplet amplification. This barcode multiplet correction allowed for more accurate cell counting and greater sensitivity [76]. In another example, the Sperm-seq assay [77] was developed to quantify the prevalence of aneuploidy and recombination in tens of thousands of individual gametes. Here, the authors identified 1201 barcode multiplets (termed 'bead doublets' in their work [77]) among 31 228 genomes profiled. Patterns in the multiplets were used to validate concordant measurements of crossovers when the same cell was profiled multiple times. Finally, a recent approach called Slide-Tags [78] collapses the presence of bead oligos originating from multiple beads patterned in a spatial array to establish a *bona fide* spatial single cell assay. Together with dscATAC-seq, these four assays demonstrate the strategies for accounting for barcode multiplets in establishing new genomic, epigenomic, multi-omic, and spatial assays.

detection of additional 'omic features, such as perturbations, protein content, and cellular clonality. The defining feature of these single cell multi-omics assays is that all measurements occur for the same cell (Figure 4A), enabling integrative analyses of gene regulatory networks [52–54], the impact of perturbations [55,56], or lineage biases [57]. A key concept in these innovations is that these assays leverage existing detection methods without fundamentally changing the key chemistry in droplets. Thus, many of these methods append new measurements to existing high-quality profiles of the transcriptome or accessible chromatin. These additions are enabled by molecular mimicry, where analytes are made to 'look like' their intended targets in the single cell assay (Box 5). Moreover, these concepts have also been extended to single nuclei profiling, including proteogenomic measurements, such as transcription factors [58].

Perturbation barcoding and detection

The development of RNA-guided gene-editing tools via CRISPR nucleases readily enables the co-detection of genetic perturbations alongside measurements of cell state by the parallel barcoding of the guide RNA (gRNA). This capability was first realized in **Perturb-seq** [14]. In the first version of Perturb-seq, a designed gRNA library was individually barcoded, and long read sequencing was used to connect barcodes to protospacers because the protospacer was distal from the 3′ end of the RNA that would be detected with the standard Drop-seq chemistry. The flaw with this design, as revealed by subsequent studies, showed high levels of protospacer/gRNA swapping, arising from reverse transcriptase recombining between templates of RNA genomes, within lentiviral particles during DNA synthesis (Figure 4B) [59,60]. This template switching, especially prevalent when various gRNA-barcode viruses are co-packaged and most frequently at homologous regions, disrupts the accurate identification of gRNAs. Ultimately, this critical design failure was quickly corrected via a new vector used in **CRISPR**

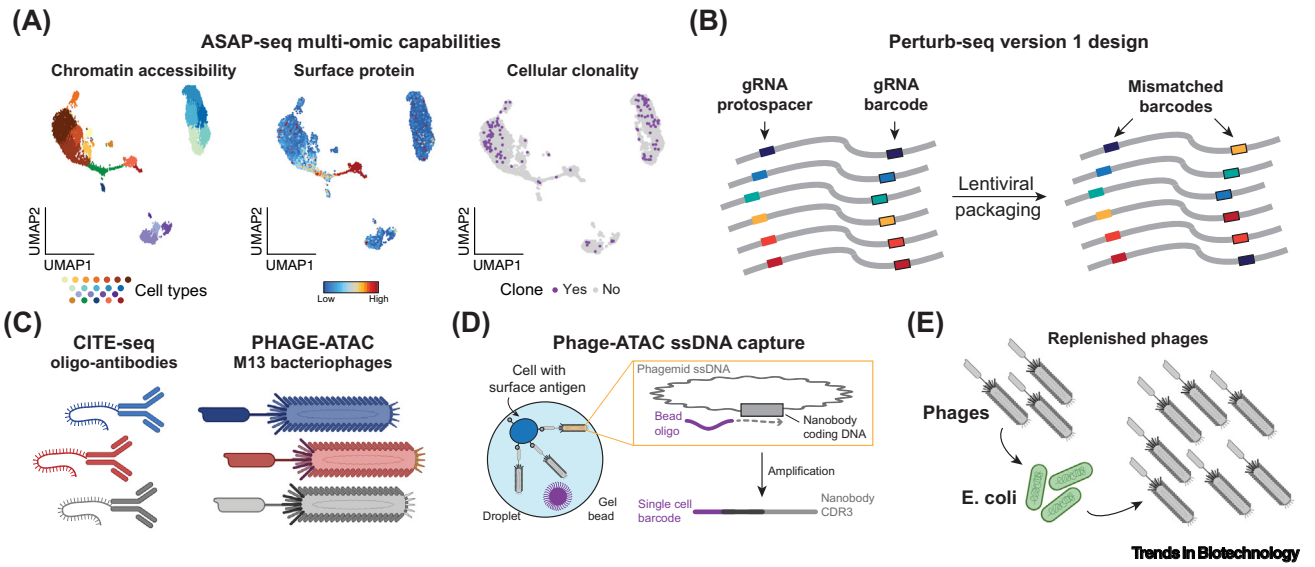

Figure 4. Concepts and opportunities in single cell multi-omics assays. (A) Summary of single cell multi-omic measurements. The same cells are represented in a reduced-dimension space. Bone marrow mononuclear cells profiled with assay for transposase accessible chromatin sequencing (ATAC) with select antigen profiling by sequencing (ASAP-seq) [16] are embedded with complementary measurements of cell types (from chromatin accessibility), cell surface protein, and clonality from mitochondrial (mt)DNA mutations. (B) Summary of barcode swapping occurring during lentiviral packaging, a known confounder in the original Perturb-seq design. (C) Schematic of reagents for proteogenomic workflows, including oligo-conjugated antibodies for assays such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) [12] and M13 bacteriophages displaying nanobodies in PHAGE-ATAC [62]. (D) Summary of PHAGE-ATAC method [62], which amplifies single-stranded (ss)DNA encoding protein binders directly. (E) Schematic indicating that PHAGE-ATAC reagents are renewable and are a key and novel feature in a single cell assay. Abbreviations: *E. coli, Escherichia coli; gRNA, guide RNA.*

Box 5. Examples of multi-omic detection methods

Most existing single cell kits have been optimized to detect a specific form of cellular molecule (i.e., poly-A transcripts in scRNA-seq or Nextera-flanked DNA fragments in scATAC-seq). A common strategy to enable multi-omic detection is to engineer molecules for detection in these assays to 'look like' what the assay intended to measure, thereby extending one high-quality measurement into multi-omic detection. Examples of this molecular mimicry are discussed here.

M13 bacteriophages with genetic engineering

A genetic system was specifically designed for the M13 phagemid to comprise: (i) a nanobody that binds to a specific epitope; (ii) a PAC-tag, which includes the Nextra Read 1 sequence to hybridize to the bead oligos; and (iii) the p3 phage coat protein for displaying these elements on the phage surface (see Figure 4D in the main text). In this sense, the single-stranded DNA in the phagemid looks like an ATAC fragment, allowing for efficient amplification using the existing ATAC-seq kit.

Aptamers containing capture sequences

Aptamers are RNAs that fold in a 3D confirmation and bind to specific targets of interest. Given that aptamers are RNA molecules, adding a polyadenylation sequence during synthesis readily allows for efficient capture and sequencing. Each aptamer is directly identifiable by its nucleic sequence, obviating the requirement for barcoding the binding agent. Although parsimonious, Apt-seq has low adoption, likely due to limited aptamer reproducibility and stability compared with antibodies.

gRNA detection with scATAC-seq

Additional approaches have used this mimicry approach for perturbations in other multi-omic assays. Spear-ATAC [61] flanks the lentiviral gRNA with Nextera Read1 and Read2 sequences to facilitate the capture of gRNA fragments that will be amplified during droplet scATAC-seq chemistry. This design capitalizes on a similarity of DNA molecules cut by Tn5 transposase, which are similarly flanked by Nextera sequences in the ATAC-seq chemistry, directly detecting the protospacer to ensure no barcode swapping. A caveat of this approach is that most existing guide RNA libraries do not contain Nextera sequences, requiring custom gRNA synthesis for compatibility with Spear-ATAC.

Barcoding heterogenous RNA molecules

In addition to measuring different analytes in multi-omic assays, emerging strategies have suggested paths for quantifying diverse RNA species in single cell reactions. For example, preliminary evidence in other split-pool contexts indicates that random hexamers [79] or rationally designed probes [80] can be used instead of poly-T, RT-based detections of transcripts. Other approaches, such as Vasa-seq [81], append poly-A sequences to arbitrary RNA molecules, resembling polyadenylated transcripts that can readily be recovered using standard commercial scRNA-seq kits. These modifications may allow for more faithful sampling of the total RNA content of the cell [82].

droplet-sequencing (CROP-seq), which directly barcodes the protospacer for the gRNA via a rational vector engineering design and has remained reliable ever since [15]. Here, CROP-seq allows the perturbation to occur closer to the critical functional unit (the protospacer guide RNA), obviating potential barcode swapping. Analogous perturbation detection has been engineered for concomitant readouts with scATAC-seq [61].

Proteogenomic barcoding and detection

In a similar vein of multi-omic innovation, proteogenomic tools use the same 'mimicry' approach for facile detection inside droplets. For instance, **ATAC with select antigen profiling by sequencing (ASAP-seq)** enables a multimodal readout that simultaneously profiles accessible chromatin, protein levels, and optionally mitochondrial DNA for cellular clonality by repurposing the existing antibody–oligonucleotide conjugates (Figure 4A) [16]. Given that the bead oligo capture sequences on the scATAC-seq kit are complementary to the Nextra Read 1 sequence attached to DNA fragments after Tn5 transposition, the poly-A CITE-seq reagents are not immediately compatible. Thus, the critical innovation in ASAP-seq was a **bridge oligo** complementary to the poly-A antibody-oligo sequence on one end and the Read 1 sequence on the other. This development allows existing CITE-seq reagents to be used directly in ASAP-seq without requiring a new set of antibodies to be conjugated for capture with the ATAC-seq kit.

Whereas ASAP-seq and CITE-seq enable proteogenomic measurements via the covalent attachment of an oligo to an antibody, **Phage-ATAC (PAC)-tag** [62] leverages an M13 bacteriophage system in which genetically encoded antibody binders are attached to a PAC-tag (Figure 4C, right), obviating the need for recombinantly expressed antibodies or covalent conjugations. The M13 phagemid system uses nanobodies with known specificity to detect surface antigens and quantify binding in droplets via the barcoding of the CDR3 hypervariable region, which directly encodes the protein binder (Figure 4D and Box 5). Although the barcode-swapping rate in CITE-seq is minimal, the parsimony of the PHAGE-ATAC system upholds this principle of 'getting as close to the source' as possible because the same piece of DNA that encodes the protein binder is barcoded. Given that phages are effectively a renewable resource after infection in bacteria (Figure 4E), these reagents are ideal for large-scale experiments. Similarly, **Apt-seq** utilizes aptamer probes, which are nucleic acids that form a 3D fold and specifically attach to protein epitopes [63]. Unlike conventional antibody-tagging techniques, aptamer binding occurs via nucleic acids, which can be readily identified through DNA sequencing, eliminating the need for tag conjugation (Box 5).

In sum, coupling proteogenomic measurements to single cell genomics assays is the most mature multi-omic technology in droplet-based workflows, featuring several distinct protein detection and quantification modes. While CITE-seq is the most commonly used method, other approaches, including Apt-seq and PHAGE-ATAC, provide a more parsimonious framework to detect the nucleic acids that underlie the binding molecule rather than relying on a covalent attachment of an oligo barcode to an antibody (Box 5).

Lessons learned: new multi-omic technologies can pair orthogonal measurements, including protein abundance or perturbation, alongside standard measurements. When detecting these modalities, barcoding and quantifying the sequence directly (e.g., DNA sequencing encoding the CDR3 of binder or protospacer of gRNA) are the most robust means for quantifying multi-omic features. Future assays must minimize or rigorously benchmark the use of surrogate barcodes to avoid errant interpretation.

Concluding remarks and future perspectives

Following the first demonstration of droplet-based scRNA-seq nearly a decade ago [8,9], the limits of these technologies have been redefined via new biotechnological tools and bioinformatics analyses. In particular, baselines underlying these original developments have been reimagined with developments in the single cell field. We highlight four instances of these innovations and collectively suggest two themes that may catalyze continued innovation in single cell technologies in the coming decades (see Outstanding questions).

First, when new single cell technologies are assessed, explicitly defining and testing the assay assumptions can validate and enable future methodological advancements. In addition to experimental assumptions about what is loaded into a droplet, namely a single cell (Figure 1), no ambient molecules (Figure 2), and an individual bead (Figure 3), revisiting widely held assumptions that assay baselines can enhance the interpretation of single cell genomics data. The interplay of experimental and computational innovation in detecting barcode multiplets highlights this idea (Figure 3). In this case, an algorithm that could detect barcode multiplets, allowed for the superloading of beads into droplets, ultimately increasing the cell capture efficiency in the dscATAC-seq assay [10]. As another example, a typical preprocessing step in scATAC-seq analyses required the binarization of the peak-by-cell matrix upstream of dimensionality reductions [28,64]. Recently, a pair of studies revisited this assumption, instead finding that the peak matrix contained a dynamic range of values that improved performance in downstream analyses, such

Outstanding questions

For a new droplet-based assay, has the assay been verified to measure single cells, and can cell doublets be identified/quantified?

Can common modes of noise, including ambient nucleic acids, be quantified via experimentation and/or accounted for via statistical modeling?

Can internal controls, including those derived from multiple barcodes per droplet, be used to quantify reproducibility and/or performance?

Do the measured molecules directly correspond to the biological feature, or is there a potential failure mode that would mismatch a barcode with what it labels?

as cell clustering [65,66]. Thus, methodological innovation can be driven by new wet-lab protocols or computational approaches in developing new technologies.

Second, because the range of possible analytes measured in single cell genomics technologies is ever-expanding, principled designs can enhance assay performance and reliability. Species-mixing experiments (Figure 1) remain a gold standard for verifying single cell capture. Given that new assays couple multiple measurements from the same cell, we highlight successful applications of multi-omic design (Figure 4) that have catalyzed the detection and quantification of heterogeneous analytes. Although many potential solutions exist, parsimonious methods allow for greater interoperability and future development. For example, the bridge oligonucleotide in ASAP-seq allowed for the reuse of existing CITE-seq [12] reagents despite the capture sequence on the gel beads changing between assays. This economical design enabled rapid extensions, including scCUT&Tag-pro [67], which maps histone modifications and links disparate measurements through a common surface proteome quantification. Similarly, the capture strategy in CROP-seq, which directly associates the cell barcode with the protospacer, allowed for more robust detection of the gRNA compared with a correlated barcode in the original Perturb-seq implementation [14], ultimately mitigating an issue of barcode swapping that can occur during lentiviral packaging [60]. Thus, establishing experimental systems that measure molecules as close to their source as possible can bolster the utility and interpretability of new analytes in single cell assays. In this sense, we highlight the transformative potential of new sequencing chemistries, including long-read technologies [68] and bioinformatics workflows (e.g., scNanoGPS [69] and Blaze [70]) that may provide deeper insights from direct measurements of the source molecules.

In sum, the expansive biological utility of single cell genomics technologies has established these assays as a critical workflow in the molecular biology toolkit. By reflecting on these innovations over the past few years, this review provides a new perspective that may expedite future method development.

Acknowledgments

We are grateful to members of the Lareau lab for helpful discussions. This work was supported by R00 HG012579 (C.A.L.) and the NIH/NCI Cancer Center Support Grant P30 CA008748 (A.C. and C.A.L.).

Declaration of interests

None declared by authors.

References

1. Tang, F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382
2. Tanay, A. and Regev, A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338
3. Chen, H. *et al.* (2021) High-throughput Microwell-seq 2.0 profiles massively multiplexed chemical perturbation. *Cell Discov.* 7, 107
4. Sziraki, A. *et al.* (2023) A global view of aging and Alzheimer's pathogenesis-associated cell population dynamics and molecular signatures in human and mouse brains. *Nat. Genet.* 55, 2104–2116
5. Clark, I.C. *et al.* (2023) Microfluidics-free single-cell genomics with templated emulsification. *Nat. Biotechnol.* 41, 1557–1566
6. Joensson, H.N. and Andersson Svahn, H. (2012) Droplet microfluidics–a tool for single-cell analysis. *Angew. Chem. Int. Ed. Eng.* 51, 12176–12192
7. Zheng, G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049
8. Klein, A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201
9. Macosko, E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214
10. Lareau, C.A. *et al.* (2019) Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* 37, 916–924
11. Satpathy, A.T. *et al.* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* 37, 925–936
12. Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868
13. Peterson, V.M. *et al.* (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939
14. Dixit, A. *et al.* (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866
15. Datlinger, P. *et al.* (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301
16. Mimitou, E.P. *et al.* (2021) Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* 39, 1246–1258

17. Mimitou, E.P. *et al.* (2019) Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* 16, 409–412
18. Hao, Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587
19. Swanson, E. *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* 10, e63632
20. Chen, A.F. *et al.* (2022) NEAT-seq: simultaneous profiling of intranuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* 19, 547–553
21. Ding, J. *et al.* (2020) Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 38, 737–746
22. Huang, M. *et al.* (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542
23. Mylka, V. *et al.* (2022) Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq. *Genome Biol.* 23, 55
24. Stoeckius, M. *et al.* (2018) Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* 19, 224
25. McGinnis, C.S. *et al.* (2019) MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* 16, 619–626
26. McGinnis, C.S. *et al.* (2019) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8, 329–337
27. Wolock, S.L. *et al.* (2019) Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 8, 281–291
28. Granja, J.M. *et al.* (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* 53, 403–411
29. Thibodeau, A. *et al.* (2021) AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.* 22, 252
30. Zhu, Q. *et al.* (2024) deMULTIplex2: robust sample demultiplexing for scRNA-seq. *Genome Biol.* 25, 37
31. Kang, H.M. *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94
32. Heaton, H. *et al.* (2020) Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* 17, 615–620
33. Hartoularos, G.C. *et al.* (2023) Reference-free multiplexed single-cell sequencing identifies genetic modifiers of the human immune response. *bioRxiv*, Published online June 3, 2023. https://doi.org/10.1101/2023.05.29.542756
34. Xu, J. *et al.* (2019) Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.* 20, 290
35. Yazar, S. *et al.* (2022) Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* 376, eabf3041
36. Datlinger, P. *et al.* (2021) Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* 18, 635–642
37. Hwang, B. *et al.* (2021) SCITO-seq: single-cell combinatorial indexed cytometry sequencing. *Nat. Methods* 18, 903–911
38. Yang, S. *et al.* (2020) Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* 21, 57
39. Madissoon, E. *et al.* (2019) scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* 21, 1
40. Janssen, P. *et al.* (2023) The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol.* 24, 140
41. Denisenko, E. *et al.* (2020) Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* 21, 130
42. Slyper, M. *et al.* (2020) Author Correction: a single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* 26, 1307
43. Lun, A.T.L. *et al.* (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20, 63
44. Megas, S. *et al.* (2024) EmptyDropsMultiome discriminates real cells from background in single-cell multiomics assays. *Genome Biol.* 25, 121
45. Anon. (2023) CellBender removes technical artifacts from single-cell RNA sequencing data. *Nat. Methods* 20, 1285–1286
46. Young, M.D. and Behjati, S. (2020) SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9, giaa151
47. Alvarez, M. *et al.* (2020) Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Sci. Rep.* 10, 11019
48. Mulè, M.P. *et al.* (2022) Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nat. Commun.* 13, 2099
49. Lareau, C.A. *et al.* (2020) Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat. Commun.* 11, 866
50. Batson, J. *et al.* (2019) Molecular cross-validation for single-cell RNA-seq. *bioRxiv*, Published online September 30, 2019. https://doi.org/10.1101/786269
51. Tyler, S.R. *et al.* (2023) Self-supervised benchmarking for scRNAseq clustering. *bioRxiv*, Published online July 10, 2023. http://x.doi.org/10.1101/2023.07.07.548158
52. Ma, S. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 183, 1103–1116
53. Mitra, S. *et al.* (2023) Single-cell multiome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *bioRxiv*, Published online June 14, 2023. https://doi.org/10.1101/2023.06.13.544851
54. Kartha, V.K. *et al.* (2022) Functional inference of gene regulation using single-cell multi-omics. *Cell Genom.* 2, 100166
55. Replogle, J.M. *et al.* (2022) Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* 185, 2559–2575
56. Hunt, K.V. *et al.* (2022) scTEM-seq: single-cell analysis of transposable element methylation to link global epigenetic heterogeneity with transcriptional programs. *Sci. Rep.* 12, 5776
57. Rodriguez-Fraticelli, A.E. *et al.* (2020) Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* 583, 585–589
58. Chung, H. *et al.* (2021) Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* 18, 1204–1212
59. Xie, S. *et al.* (2018) Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS One* 13, e0198635
60. Hill, A.J. *et al.* (2018) On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* 15, 271–274
61. Pierce, S.E. *et al.* (2021) High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.* 12, 2969
62. Fiskin, E. *et al.* (2022) Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC. *Nat. Biotechnol.* 40, 374–381
63. Delley, C.L. *et al.* (2018) Combined aptamer and transcriptome sequencing of single cells. *Sci. Rep.* 8, 2919
64. Stuart, T. *et al.* (2021) Single-cell chromatin state analysis with Signac. *Nat. Methods* 18, 1333–1341
65. Miao, Z. and Kim, J. (2024) Uniform quantification of single-nucleus ATAC-seq data with Paired-Insertion Counting (PIC) and a model-based insertion rate estimator. *Nat. Methods* 21, 32–36
66. Martens, L.D. *et al.* (2023) Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat. Methods* 21, 28–31
67. Zhang, B. *et al.* (2022) Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.* 40, 1220–1230
68. Al'Khafaji, A.M. *et al.* (2024) High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat. Biotechnol.* 42, 582–586
69. Shiau, C.-K. *et al.* (2023) High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat. Commun.* 14, 4124
70. You, Y. *et al.* (2023) Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. *Genome Biol.* 24, 66
71. Huang, Y. *et al.* (2019) Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* 20, 273
72. Wu, B. *et al.* (2024) Overloading And unpacKing (OAK) - droplet-based combinatorial indexing for ultra-high throughput single-cell multiomic profiling. *bioRxiv*, Published online January 24, 2024. https://doi.org/10.1101/2024.01.23.576918

73. Caglayan, E. *et al.* (2022) Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* 110, 4043–4056
74. Muskovic, W. and Powell, J.E. (2021) DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* 22, 329
75. Xi, J. *et al.* (2023) SiftCell: a robust framework to detect and isolate cell-containing droplets from single-cell RNA sequence reads. *Cell Syst.* 14, 620–628
76. Marshall, J.L. *et al.* (2020) HyPR-seq: single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes. *Proc. Natl. Acad. Sci. USA* 117, 33404–33413
77. Bell, A.D. *et al.* (2020) Insights into variation in meiosis from 31,228 human sperm genomes. *Nature* 583, 259–264
78. Russell, A.J.C. *et al.* (2024) Publisher Correction: slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* 625, E11
79. Tran, V. *et al.* (2022) High sensitivity single cell RNA sequencing with split pool barcoding. *bioRxiv*, Published online August 27, 2022. https://doi.org/10.1101/2022.08.27.505512
80. Janesick, A. *et al.* (2023) High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.* 14, 8353
81. Salmen, F. *et al.* (2022) High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.* 40, 1780–1793
82. Hornung, B.V.H. *et al.* (2023) Comparison of single cell transcriptome sequencing methods: of mice and men. *Genes* 14, 2226
83. Abate, A.R. *et al.* (2009) Beating Poisson encapsulation statistics using close-packed ordering. *Lab Chip* 9, 2628–2631
84. Lareau, C.A. *et al.* (2023) Single-cell multi-omics of mitochondrial DNA disorders reveals dynamics of purifying selection across human immune cells. *Nat. Genet.* 55, 1198–1209
85. Olsen, T.R. *et al.* (2023) Scalable co-sequencing of RNA and DNA from individual nuclei. *bioRxiv*, Published online February 10, 2023. https://doi.org/10.1101/2023.02.09.527940