
Article:
Dudchenko, O., Ordovas-Montanes, J. and Bingle, C.D. orcid.org/0000-0002-5405-6988 (2023) Respiratory epithelial cell types, states and fates in the era of single-cell RNA-sequencing. Biochemical Journal, 480 (13). pp. 921-939. ISSN 0264-6021

https://doi.org/10.1042/bcj20220572

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

PORTLAND PRESS

Review Article

Respiratory epithelial cell types, states and fates in the era of single-cell RNA-sequencing

Oleksandr Dudchenko[1], Jose Ordovas-Montanes[2,3] and Colin D. Bingle[1]

[1]Department of Infection, Immunity and Cardiovascular Disease, The Medical School, University of Sheffield, Sheffield, South Yorkshire, U.K.; [2]Division of Gastroenterology, Hepatology and Nutrition, Boston Children's Hospital, Boston, MA, U.S.A.; [3]Programme in Immunology, Harvard Medical School, Boston, MA, U.S.A.

Correspondence: Colin D. Bingle (c.d.bingle@sheffield.ac.uk)

Standalone and consortia-led single-cell atlases of healthy and diseased human airways generated with single-cell RNA-sequencing (scRNA-seq) have ushered in a new era in respiratory research. Numerous discoveries, including the pulmonary ionocyte, potentially novel cell fates, and a diversity of cell states among common and rare epithelial cell types have highlighted the extent of cellular heterogeneity and plasticity in the respiratory tract. scRNA-seq has also played a pivotal role in our understanding of host–virus interactions in coronavirus disease 2019 (COVID-19). However, as our ability to generate large quantities of scRNA-seq data increases, along with a growing number of scRNA-seq protocols and data analysis methods, new challenges related to the contextualisation and downstream applications of insights are arising. Here, we review the fundamental concept of cellular identity from the perspective of single-cell transcriptomics in the respiratory context, drawing attention to the need to generate reference annotations and to standardise the terminology used in literature. Findings about airway epithelial cell types, states and fates obtained from scRNA-seq experiments are compared and contrasted with information accumulated through the use of conventional methods. This review attempts to discuss major opportunities and to outline some of the key limitations of the modern-day scRNA-seq that need to be addressed to enable efficient and meaningful integration of scRNA-seq data from different platforms and studies, with each other as well as with data from other high-throughput sequencing-based genomic, transcriptomic and epigenetic analyses.

Introduction

The ability to study gene expression at a transcriptome-wide scale in organisms, tissues and single cells was historically limited by the lack of quantitative, high-throughput strategies. Only at the turn of this century has the field of transcriptomics started realising its true potential, driven by the development of two revolutionary technologies [1,2]. Initially, transition from techniques such as Northern blotting and quantitative PCR to oligonucleotide microarrays increased the number of genes profiled per assay by several orders of magnitude [3,4], whereas roughly a decade later, RNA-sequencing (RNA-seq) not only enabled detection of novel transcript isoforms, but also allowed absolute quantification of mRNA species in a sample [5]. The issue of averaged gene expression values for the studied population of cells, which concealed cellular heterogeneity in both microarray and bulk RNA-seq data, was subsequently addressed by the development of RNA-seq at a single-cell resolution (scRNA-seq) [6–10].

Advances in single-cell isolation methods, which increased the throughput to thousands of cells per experiment [11], and reduction in cost per sequenced cell from around 10 USD [9] to approximately 0.5–1 USD on 10× Genomics Chromium platform [12], contributed to the global adoption of scRNA-seq over recent years. Increased need and funding for respiratory research during the coronavirus disease 2019 (COVID-19) pandemic further accelerated the implementation of scRNA-seq in the

Received: 19 November 2022
Revised: 19 June 2023
Accepted: 20 June 2023

Version of Record published:
6 July 2023

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

respiratory field [13], effectively positioning it at the forefront of developments in scRNA-seq and rendering the airways one of the most profiled systems in the human body. In addition to facilitating identification of cell types that are infected by severe acute respiratory syndrome coronavirus clade 2 (SARS-CoV-2) [14,15] and improving the understanding of factors affecting the outcome of SARS-CoV-2 infection [16], concurrent applications of scRNA-seq in homeostasis and disease have also led to a discovery of a novel epithelial cell type termed ionocyte [17,18], yielded new insights into the remodelling and dysfunction of the airway epithelium in smokers [19,20], individuals with chronic rhinosinusitis [21], asthma [22,23], idiopathic pulmonary fibrosis [24,25] and cystic fibrosis (CF) [26]. Furthermore, scRNA-seq has enabled creation of cell atlases of the healthy human lung by individual research groups [27–29] and by consortia. For instance, the first iteration of the LungMAP Single-Cell Reference atlas of the human lung, totalling almost 350 000 single-cell transcriptomes, was released in May 2022 [30]. On the initiative of the discovAIR consortium, the core version of the Human Lung Cell Atlas (HLCA) was also launched in early 2022, incorporating transcriptomic data from nearly 600 000 individual cells [31]. In June 2023, the release of the extended version of HLCA signified one of the largest consortia-led attempts to integrate and harmonise scRNA-seq and single-nuclei RNA-seq data from more than 2.4 million cells from 49 individual datasets [32]. scRNA-seq data accessibility through atlas-associated and standalone online tools, such as CZ CELLxGENE web application and UCSC cell browser, and data integration into the existing Lung Gene Expression Analysis web portal [33] provide user-friendly interfaces for exploration of complex scRNA-seq data and their interpretation along with results from other 'omic' and conventional techniques for gene expression analysis, respectively.

Since major findings attributable to the use of scRNA-seq in the respiratory field have already been extensively reviewed [34–37], this review attempts to showcase key developments in scRNA-seq methodology and to elaborate on some of the fundamental challenges arising from differences in experimental designs and computational approaches that need to be explicitly addressed when interpreting and contextualising scRNA-seq data. For instance, two most common analyses scRNA-seq data are subjected to — categorisation of cells into distinct groups and positioning of cells along the likely differentiation axes on the basis of similarity or dissimilarity between levels of select transcripts, raise several questions. First, to which extent is a transcriptomic snapshot on its own sufficient for defining a cell type? The absence of a consensus definition of a cell type [32,38] and of comprehensive models that consolidate various aspects of cellular identity [39] further complicate matters. Second, given the stochasticity of gene expression [40] as well as biological and technical noise inherent to scRNA-seq [41], thresholds used to identify the proportion of transcriptional variation between individual cells that is both true and biologically meaningful play a pivotal role in interpretation of scRNA-seq results. Third, with various types of sequencing platforms available and a rapidly growing number of data analysis tools, the reproducibility of the discoveries and phenotype-associated outcomes of scRNA-seq studies needs to be critically evaluated. Confounding factors, on top of those introduced by specific scRNA-seq protocols, include interindividual differences, variability in sample acquisition and processing, and distinct tissue culture and differentiation methods used in *in vitro* studies. Although these and other challenges are discussed here in the respiratory context, our insights will be applicable to scRNA-seq conducted in additional tissue systems. Due to the complexity of the respiratory tract, we mainly focus on the ways scRNA-seq has influenced the knowledge about heterogeneity and plasticity of epithelial cells in the lower airways of humans, from trachea to alveoli, with brief context added from studies of upper airways and model organisms.

scRNA-seq workflow

Overcoming technical bottlenecks at the level of single-cell capture and early multiplexing [11], numerous scRNA-seq techniques have been developed and revised since the publication of the pioneering paper by Tang et al. [6]. With timeline figures juxtaposing years of introduction and throughput of exhaustive lists of scRNA-seq techniques generated in other publications [11,42] and given that most protocols share core principles (Figure 1), one way of simplifying categorisation of a large variety of diverse scRNA-seq methods can be based on the extent of transcript coverage with short-read Illumina complementary DNA (cDNA) sequencing — full-length, 5′- and 3′-end. One of the most frequently used plate-based methods from the full-length group is SMART-seq [7,43,44], with the latest iteration being SMART-seq 3. SMART-seq-based scRNA-seq experiments are usually conducted on a relatively small number of cells, ranging from a few hundreds to several thousands, which are sequenced at a depth of up to 1–2 M reads per cell. Early 5′-end scRNA-seq methods included STRT-seq [45] and STRT/C1 [46], while common 3′-end methods encompass CEL-seq [47,48],

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

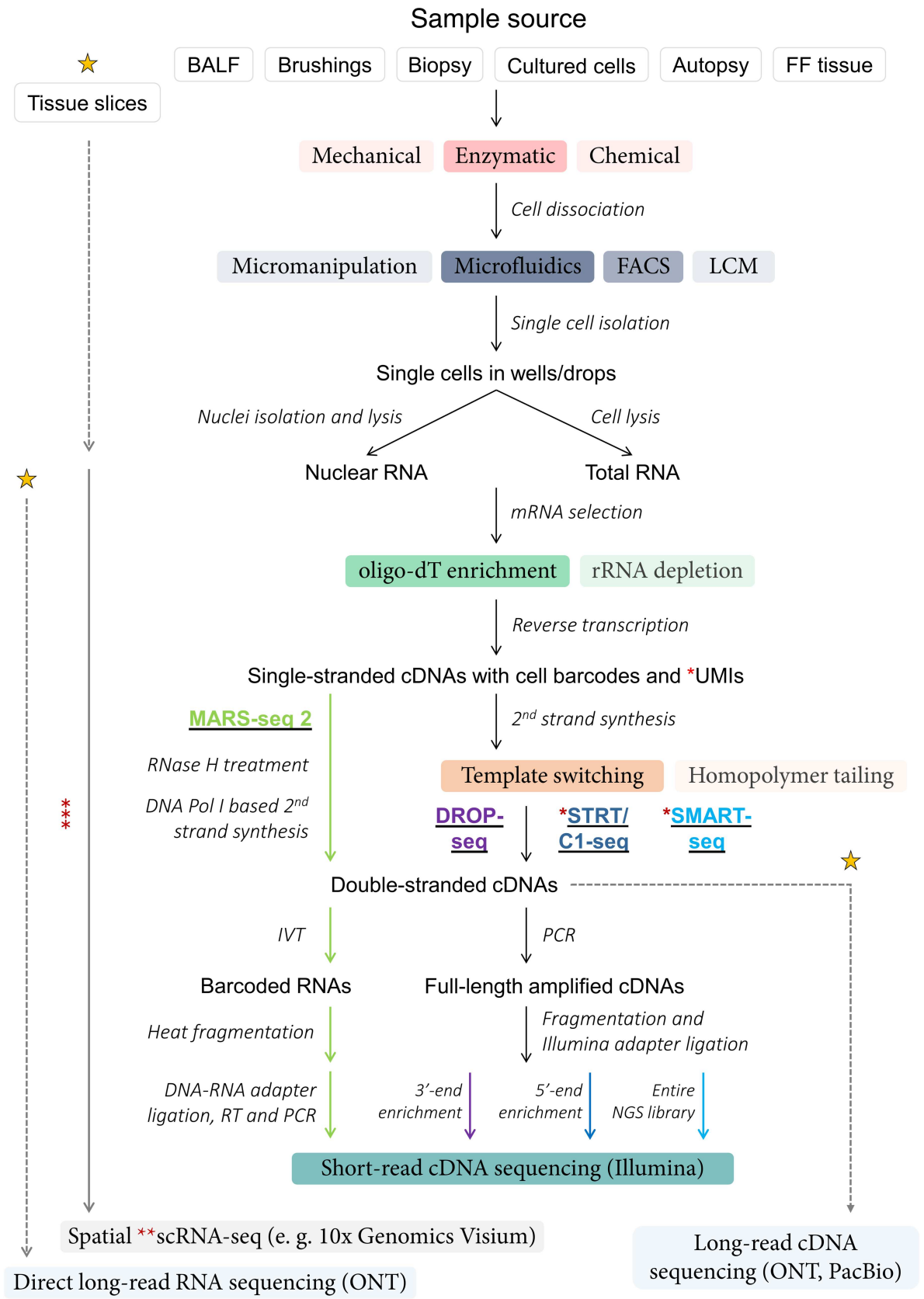


Figure 1. Process of sc-RNA sequencing. Part 1 of 2

Summary of shared sample processing steps and simplified final stages of next generation sequencing (NGS) library

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

Figure 1. Process of sc-RNA sequencing. Part 2 of 2

preparation in established scRNA-seq protocols (MARS-seq 2 — light green font and arrows; DROP-seq — purple font and arrow; STRT/C1-seq — navy font and arrow; SMART-seq — light blue font and arrow). *Key*: black arrows — shared steps; coloured boxes — methods (colour intensity corresponds to relative frequency of method usage in scRNA-seq protocols); italicised text — experimental stages; yellow stars — emerging and/or future scRNA-seq technologies (details not shown); dotted grey arrows — stages of conventional protocols that may be (partially) bypassed with emerging and/or future scRNA-seq technologies; * — unique molecular identifiers (UMIs) were not incorporated into complementary DNAs (cDNAs) in the original STRT/C1 sequencing protocol, SMART-seq v1 and v2 protocols; ** — majority of currently available spatial transcriptomic technologies provide regional rather than single-cell resolution; ***(solid grey arrow) — methodology for emerging spatial transcriptomic techniques not shown; FACS — fluorescence-activated cell sorting; LCM — laser capture microdissection; FF — formalin-fixed; BALF — bronchoalveolar lavage fluid (including endotracheal aspirate); IVT — *in vitro* transcription; ONT — Oxford Nanopore; PacBio — Pacific Biosciences.

MARS-seq [49,50], Cyto-seq [51] and DROP-seq [8] (Figure 1). Droplet-based 3′-end scRNA-seq platforms developed in academic settings or in industry, such as DROP-seq [8] and 10× Genomics Chromium, respectively, generally support higher experimental throughput, but often at the expense of sequencing depth and transcript coverage, resulting in less transcriptomic detail being captured per cell. With thorough descriptions of methodologies provided in the original papers and compared in dedicated reviews [52–54], it is worth highlighting several key features of modern-day scRNA-seq protocols.

First, incorporation of unique molecular identifier (UMI) sequences into the cDNA molecules, which is a step in most 3′-end scRNA-seq methods [55] and in SMART-seq 3 [44], enables absolute quantification of transcripts, identification and elimination of PCR duplicates [56,57]. RNA 'spike-in' molecules, in turn, are widely used for normalisation during data analysis [41,58] and, if modified via addition of internal UMIs, can improve the accuracy of RNA quantification in both droplet- and plate-based scRNA-seq methods [59]. Lastly, in light of the improving accuracy of long-read sequencing technologies from Oxford Nanopore (ONT) and Pacific Biosciences (PacBio) [60], one can envision the increase in their adoption for scRNA-seq in the near future [61] (Figure 1) due to their ability to identify novel and differentially expressed transcript isoforms [62,63]. Introduction of the spatial context into scRNA-seq is also promising [64] (Figure 1), but current commercial platforms, such as 10× Genomics Visium, provide regional rather than single-cell resolution when it comes to *in situ* mRNA capture, barcoding and sequencing.

Similarly to sample preparation and sequencing protocols, many tools for analysis of scRNA-seq data were developed by leveraging insights from bulk RNA-seq [65]. Currently, researchers' options range from proprietary software packages (e.g. 10× Genomics Cell Ranger) to open-source, integrated pipelines such as Seurat [66], Monocle [67], Scanpy [68] and more than 1500 individual scRNA-seq tools [69], which can be mixed and matched to create a highly customised data analysis strategy. With in-depth coverage of general and scRNA-seq-specific bioinformatics being beyond the scope of this review, only the key stages of scRNA-seq data pre-processing, main downstream analyses, common statistical methods and software packages are briefly summarised here (Figure 2). For conceptual overview of each stage and step by step practical guidance, reviews by Wu and Zhang [65], and Luecken and Theis [41], respectively, are recommended.

Although barriers of entry to scRNA-seq data analysis have been substantially lowered both in terms of technical requirements [70] and accessibility to individuals with minimal knowledge of programming, it remains a resource-intensive process, particularly when it comes to analysis and integration of datasets from tens or even hundreds of thousands of cells. Biological hypotheses and method awareness are required for evaluation, selection of filters and parameters at most stages of scRNA-seq data analysis, including quality control, data normalisation, dimensionality reduction and cell cluster annotation. Data re-analysis can also often be hindered by incomplete documentation of conducted bioinformatic analyses in literature. In other words, even if a pipeline or package names are specified in the methods section, re-running them on the same raw data may produce different results unless the exact parameters for each step, e.g. number of UMIs per viable cell or a cell (node) that serves as a starting point for trajectory inference, are provided by the authors. The impact of these factors as well as of the high degree of customisation, large number of zero expression values [71], variable performance of different clustering [72] and trajectory inference methods [73] will be elaborated on in further sections.

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

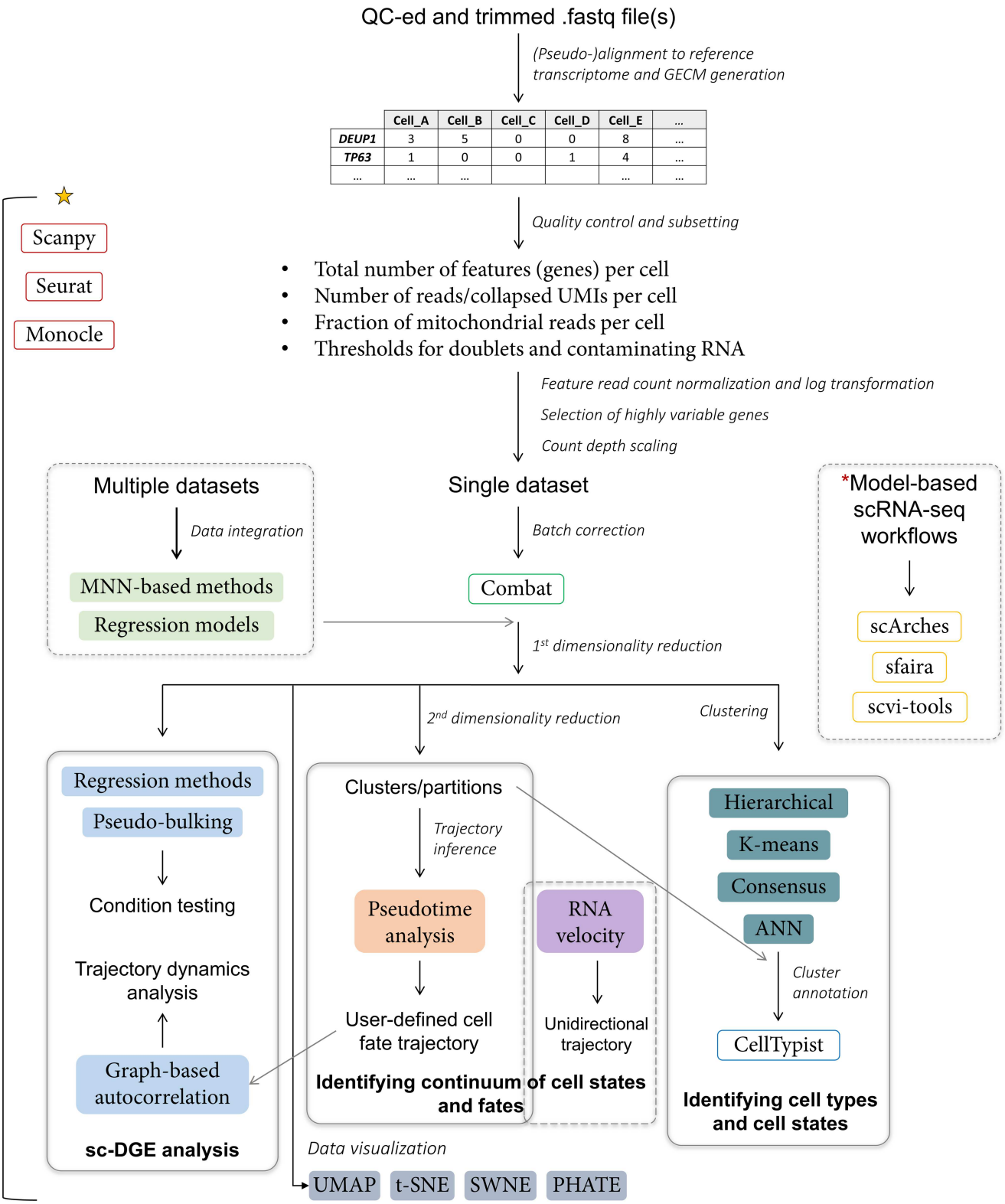


Figure 2. Analysis of sc-RNA-seq data. Part 1 of 2

Summary of key stages in short-read scRNA-seq data analysis with examples of general methods (coloured, filled boxes), common bioinformatic tools (boxes with coloured outline) and key downstream applications (boxes with solid grey outline). Following standard Illumina read quality control (QC) and adapter trimming, paired or single-end reads are mapped to the reference genome or transcriptome and gene expression counts matrix

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

Figure 2. Analysis of sc-RNA-seq data. Part 2 of 2

(GECM) is generated. Afterwards, potentially non-viable cells, cell doublets and contaminating RNAs [74,75] can be filtered out *in silico*. Read counts for each cell passing set criteria are then log-normalised. Having selected highly variable genes for marker identification and optionally, having regressed cell cycle or other classes of genes, log-normalized counts are scaled, giving each gene equal weight in downstream analyses. If a dataset is obtained from an experiment conducted in multiple batches, batch correction can be applied [76]. Analysis of datasets from experiments with different designs will require data integration. Subsequently, the first dimensionality reduction is performed, typically with principal component analysis (PCA). Resulting data can then be visualised; used for single-cell differential gene expression (sc-DGE) analysis; undergo further dimensionality reduction and used for inferring developmental or condition-driven trajectories; clustered and annotated to identify cell states and cell types [77]. *Key*: yellow star — majority of the highlighted analysis steps can be performed using three listed software packages [66–68]; * — alternatively, data analysis workflow based on reference models generated with machine learning algorithms can be selected [78–80], particularly when integrating multiple datasets from the same tissue; UMI — unique molecular identifier; ANN — approximate nearest neighbours; UMAP — uniform manifold approximation and projection; PHATE — potential of heat diffusion for affinity-based transition embedding; SWNE — similarity-weighted non-negative embedding; t-SNE — t-stochastic neighbour embedding.

Evolving definitions

Cell types

Parameter-specific categorisation of cells of multicellular organisms into types is one of the most fundamental concepts in biology [81]. Highlighting the multifaceted nature of this practice, field-specific perspectives on the definition of a 'cell type' across different branches of biological science are not uncommon [38]. Adhering to a more prevalent type of definition, Vickaryous and Hall [82] catalogued cell types only if similar cells in question occurred *in vivo*, were terminally differentiated as well as at the same stage of developmental history and cell cycle. While some of these assumptions hold true for currently accepted cell types, it is not uncommon for some of the cell types to diverge from them, particularly with respect to differentiation status. To exemplify, Tata and Rajagopal [83] demonstrated that upon injury murine club cells were able to dedifferentiate into basal epithelial cells. Looking beyond the classification systems based on resemblances in cellular morphology, marker gene expression etc., evolutionary definitions of a 'cell type' have also been proposed. For instance, according to Arendt et al. [84], cells can be deemed of the same 'type' if they are more evolutionarily related than other cells of the same organism. These examples hint at the difficulty of arriving at a consensus definition of a cell type. However, it is apparent that for the majority of cell types to be definitively recognised as such, not just a single or several properties, but rather — an ensemble of features across multiple sets of parameters constituting a cell (Figure 3), from commonality in the studied population [85] to epigenomes, should be identified, validated and integrated. On top of that, the variability between cells of the same type, which can be observed even at the intra-individual level due to intrinsic stochasticity of nearly all cellular processes, sampling effects or technique artefacts, and can be further amplified during scRNA-seq data analysis, should be factored in. Hence, an optimal consensus definition of a 'cell type' should stipulate both commonly used and acceptable number of cell identity features with corresponding thresholds of variance, thereby providing a tentative framework, while balancing potential for discovery and promoting data exploration.

Two of the most comprehensive standalone single-cell atlases of the healthy human lung to date were constructed by Deprez et al. [28] and Travaglini et al. [27], with transcriptomes of nearly 80 000 cells profiled by each group. Atlas generated by Deprez et al. [28] is mainly comprised of lung epithelial cells from dozens of different locations along the airway tree, whereas the latter atlas also contains large fractions of vascular and immune cell types from the lung. Among major findings of Travaglini et al. [27], a wide range of intra- and intercellular signalling pathways was elucidated at a greater resolution, expression of key lung-disease genes were localised to cell types and differences between gene expression patterns of human lung cell types and murine counterparts were highlighted. Importantly, the authors also increased the cell type count in the lung to 58 individual cell types. This figure includes 14 newly reported cell types, many of which belong to stromal, endothelial and immune cell lineages that are not discussed here. As for the epithelial cells, Travaglini et al. [27] listed populations of 'proximal ciliated', 'proximal basal' and 'alveolar type II signalling' cells as 'novel' cell types. Deprez et al. [28], in turn, described two 'novel' cell types — 'multiciliating-goblet' and 'undefined rare' cells. Whether and to what extent these cell populations, alongside 'proliferating/cycling basal' and

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

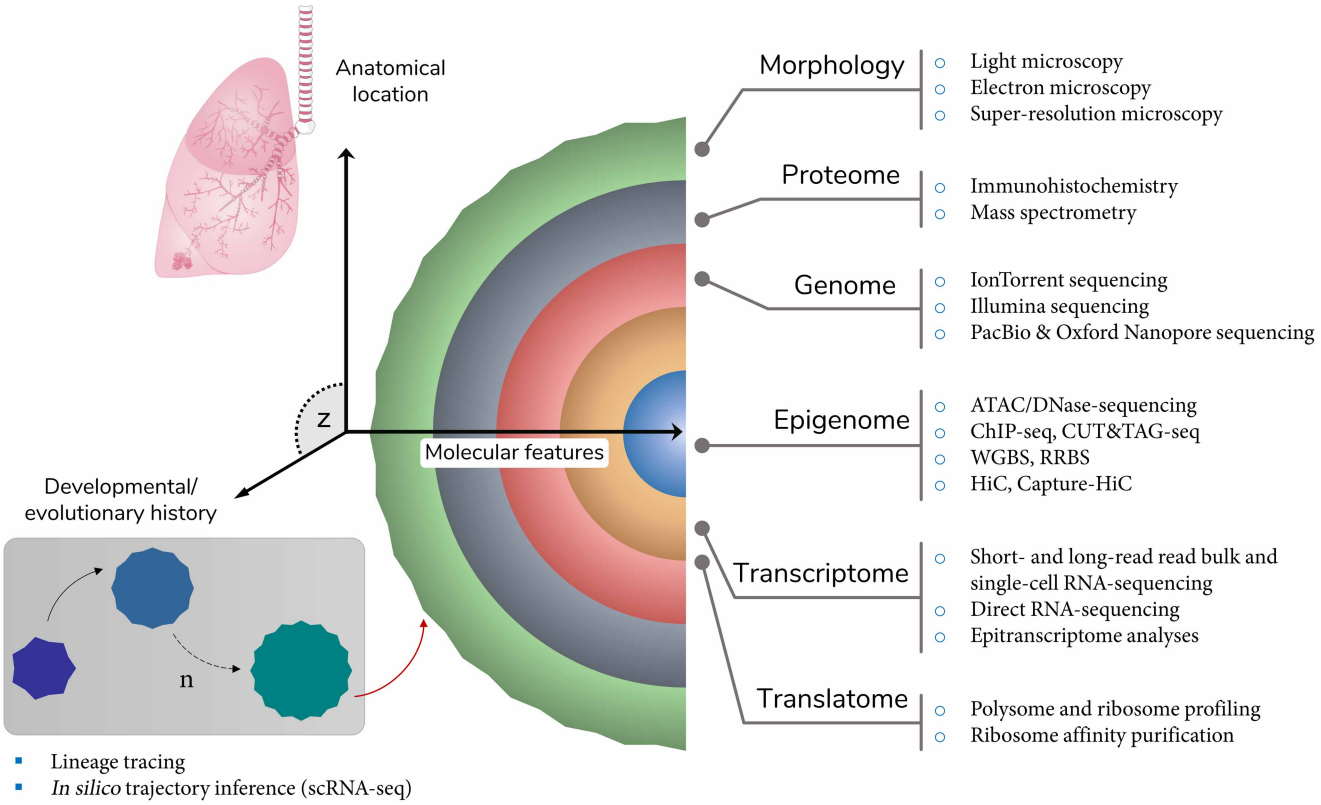


Figure 3. Features of cellular identity.

Common features of cellular identity that can be used individually or collectively to define a cell type, with examples of conventional and 'omics' techniques that can be used to investigate each set of features. *Key*: z space — other sets of features (not shown) that are independent or derived from molecular and evolutionary parameters and can be incorporated into the definition of a cell type; red arrow — directional outcome of the sequence of evolutionary/developmental events that led to the current form of the cell; sc — single cell; seq — sequencing; PacBio — Pacific Biosciences; ATAC — assay for transposase accessible chromatin; ChIP — chromatin immunoprecipitation; WGBS — whole genome bisulphite sequencing; RBBS — reduced representation bisulphite sequencing; HiC — high-throughput chromosome conformation capture.

'differentiating basal' cells, can be classified as 'novel' and/or as 'cell types' requires large-scale integrative efforts for many reasons outlined in the next section of this review. However, inclusion of ionocytes into lists of epithelial cell types that are found in the human lungs has been widely accepted and can be observed in all major publications on this topic.

Discovered concurrently by Montoro et al. [17] and Plasschaert et al. [18], pulmonary ionocytes can arguably be considered the first novel airway epithelial cell type identified using scRNA-seq by capitalising on one of the main strengths of the technique — the ability to detect rare cells with unique gene expression signatures. Distinctive cell clusters, which were highly enriched in transcripts of genes encoding a plasma membrane chloride ion transporter, cystic fibrosis trans-membrane conductance regulator (CFTR), a transcription factor required for *CFTR* expression, FOXI1, subunits of V-type ATPase, ATP6V1C2 and ATP6V0D2 and other proteins conducive to the regulation of properties of the airway surface liquid, were first detected in mouse tracheal and human bronchial tissues as well as among human bronchial epithelial cells (HBECs) differentiated at the air–liquid interface (ALI) [17,18]. With the exception of *CFTR*, similar gene expression patterns were previously reported for intercalating non-ciliated cells in the skin of *Xenopus laevis* [86] and for ionocytes in the gills of fish [87], but had not been reported in the mammalian respiratory tract. The presence of ionocytes was also confirmed using immunostaining for the aforementioned markers by Montoro et al. [17] and Plasschaert et al. [18], and subsequently — with scRNA-seq, immunostaining and/or fluorescence *in situ* hybridisation in more recent studies of both healthy [20,27,28] and diseased human airways [19,22,23,26].

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

Provided sufficient power, in terms of the number of profiled cells, is achieved in a given scRNA-seq study, expression data are often used not only to determine the heterogeneity and plasticity of common and rare epithelial cell types in different sections of the respiratory tract, but also to illustrate their relative tissue distributions. For example, in some publications, relative frequencies of cell types are derived by dividing the number of cells mapping to a particular cluster by the total number of cells that were annotated from sequenced samples [22,88,89]. These may not always correlate with observations from conventional histological analyses, as illustrated by a relatively low abundance of multiciliated cells (MCCs) in the tracheal section of the proximal airway epithelium identified with scRNA-seq in contrast with immunofluorescence staining (Figure 4A). Nonetheless, such type of comparison is vital for highlighting potential challenges with cell dissociation protocols and other experimental factors that may lead to variable efficiencies in survival, capture and sequencing of particular cell types.

Cell states

Identification of canonical cell types via annotation of automatically generated or pre-determined number of clusters, which can be followed by subdivision of select clusters into smaller groups based on the expression levels of several to a few dozens of genes, commonly referred to as tiered clustering, is a recurring theme in a majority of large-scale scRNA-seq studies in the respiratory field. In some papers, author-defined subpopulations of established cell types are numbered, e.g. ciliated 1, ciliated 2, ciliated 3 [20], whereas in others — they are given names that attempt to characterise their general transcriptome states, e.g. proliferating, differentiating and proteasomal basal cells [19], anatomical locations, e.g. proximal and distal basal cells [27] or morphological appearance, e.g. hillock club and basal cells [17]. In addition to potential confusion that may be caused by subpopulation names that vary from publication to publication, especially when a simple numbering strategy is chosen, such approach for classification of cells is frequently accompanied by inconsistent use of terms that describe aspects of cellular identity or by lack of reproducibility. For instance, a subpopulation of MCCs, also known as 'deuterosomal cells', which is enriched in transcripts of early ciliogenesis genes (e.g. *CDC20B*, *DEUP1*, *FOXN4*), is simultaneously categorised as a subtype of MCCs [89], recognised as an independent cell type [28] and is not even found in the first place [27]. Further exacerbating the issue, a comparison of expression profiles of the aforementioned populations and two other subpopulations identified in three different studies, yields mixed results (Figure 5), with mere 4–16 markers being shared out of top 99 differentially expressed genes (DEGs). This number of DEGs was selected because positive log fold change values were available only for 99 genes in two studies for the suprabasal population [20,89]. It is worth noting that a more thorough comparison of clusters between studies would require complete transcriptome integration or correlation analyses, and any conclusions based on whether a particular gene is transcribed or not do not fully reflect observed variation in gene expression between cells. This superficial analysis, however, indicates that only a few genes, which are not necessarily at the top of the DEG list, are identified as shared markers of subpopulations of cells with similar or identical names that were detected in multiple scRNA-seq experiments. Interestingly, sample source in selected scRNA-seq studies and cell populations appears to have little to no impact. For example, proliferating (cycling) basal cells identified among HBECs differentiated at the ALI [19,89] bear greater individual resemblance the same cell population identified in scRNA-seq conducted directly on donor cells [27] than to each other in terms of the number of shared top 99 DEGs (Figure 5B). To what extent this observation and general discrepancies between compared cell populations are affected by variability in cell throughput or sequencing depth remains unclear as key quality metrics were only fully specified by two out of five papers (Figure 5D).

Although the requirement for consensus nomenclature, when it comes to differentiating between 'cell type', 'cell state' and similar terms that are often used interchangeably, has been highlighted before [36,91], there are multiple ways in which it can be satisfied. Under ideal circumstances, reference values and their scope outlined in publications such as meta-analyses and reviews should provide sufficient guidance for researchers on how to improve scRNA-seq data interpretation and gradually standardise the use of key definitions. In practice, experiment- or hypothesis-dependent aspects that may not be covered by such guidance as well as the absence of certain data quality and processing criteria, which can be set by publishers, reviewers, peers etc., may not result in widespread adoption of the guidance. An alternative and more likely solution, given the growing number and size of large scRNA-seq cell atlases, can be the use of reference cell type annotations provided by consortia, which in the respiratory research field are mainly represented by HLCA [31,32] and LungMAP [30]. For example, in the recently released extended HLCA, a new cell identity reference framework was proposed

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

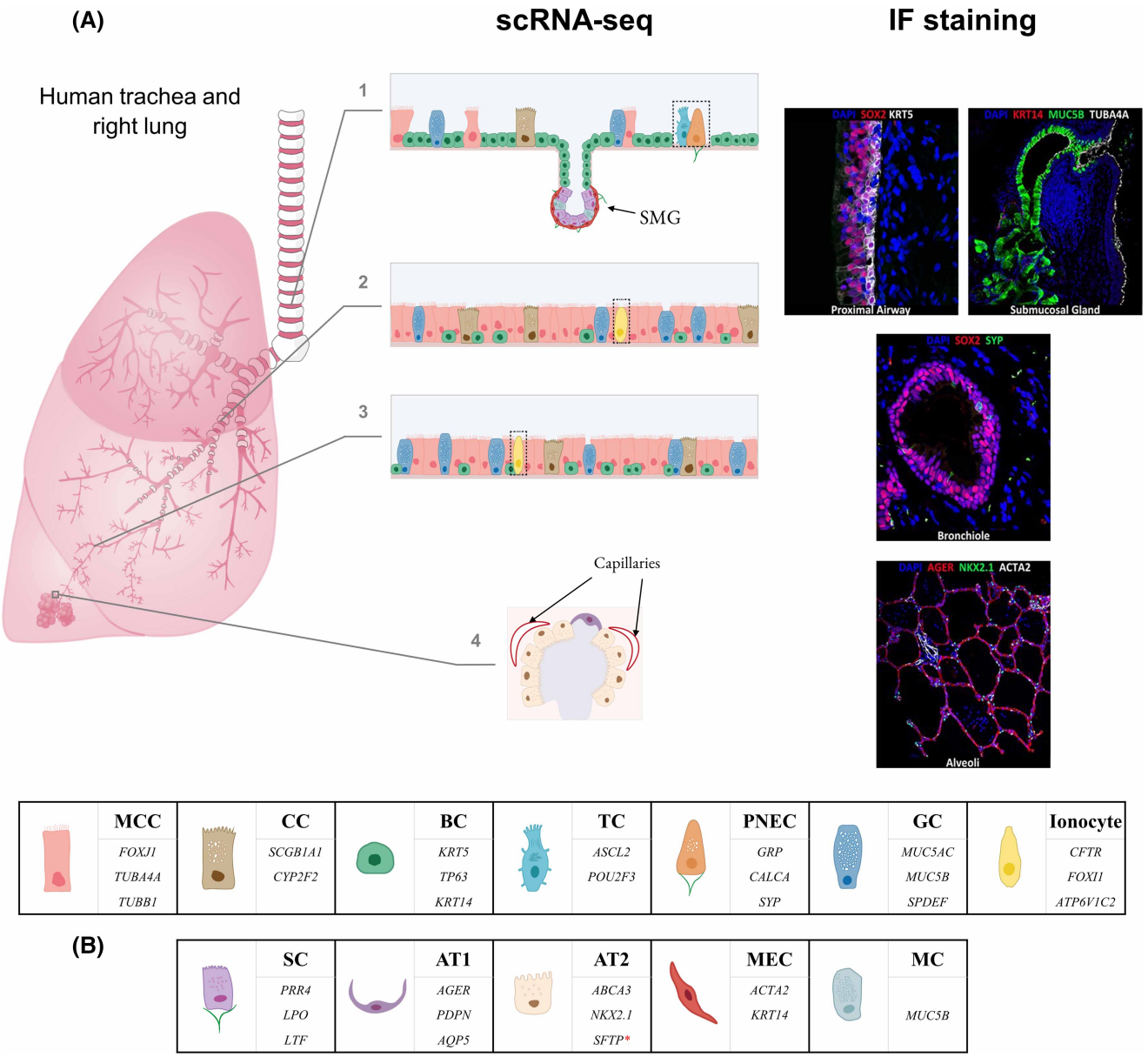
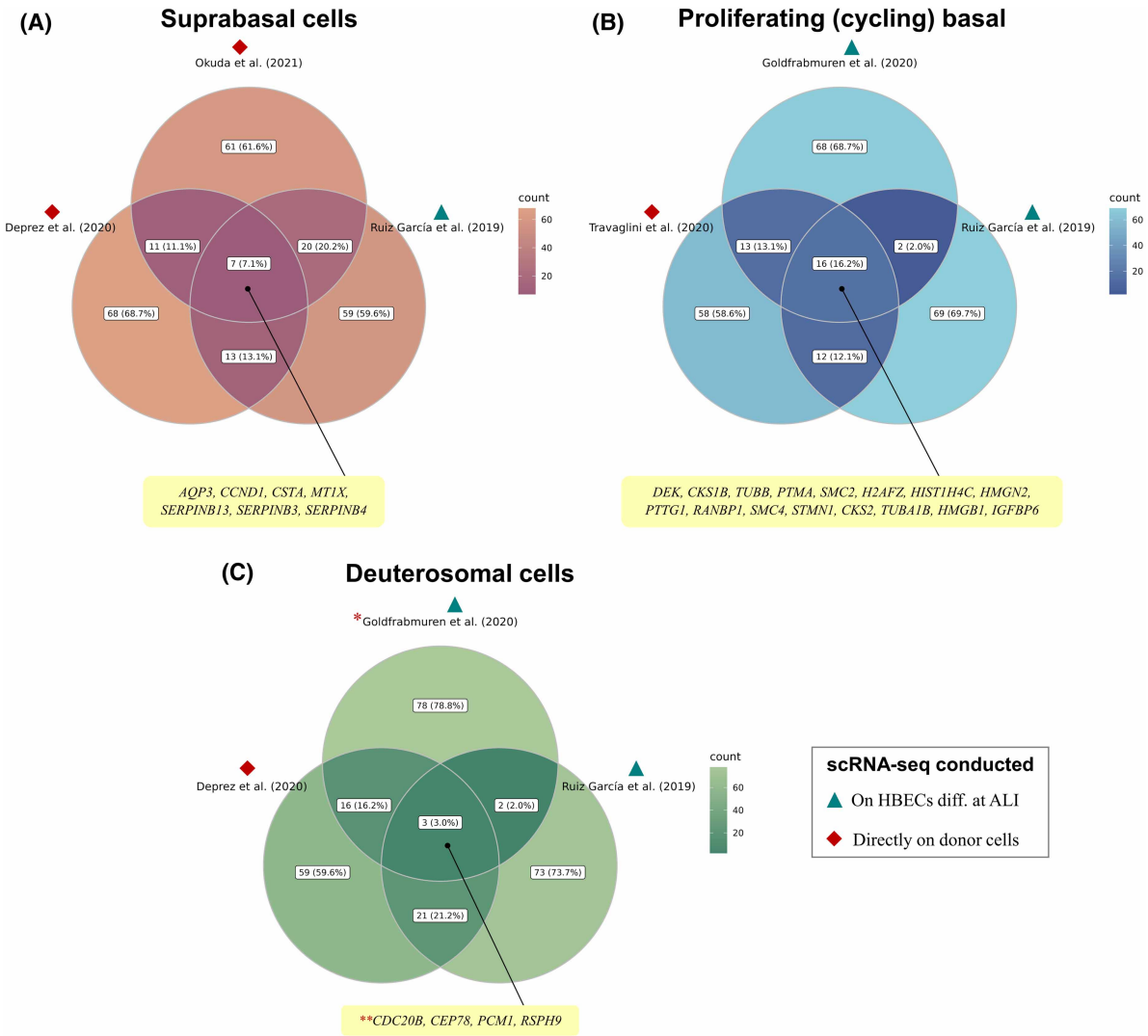


Figure 4. Cell populations in the human lower airways.

(**A**) Epithelial cell landscape of the human lower airways based on select scRNA-seq data and immunofluorescence (IF) staining (undefined and novel 'cell types' that were either identified only in a single study or may better fit the description of 'cell states' or 'subtypes' of canonical epithelial cell types are not shown; where applicable, fractions of author-defined subpopulations of a canonical cell type were added up to represent each canonical cell type as a single category); 1 — trachea (tracheal cell type fractions roughly estimated from Deprez et al. [28], Figure 1F, tracheal biopsy stacked column; number of displayed cells (*n*) = 50); 2 — bronchi (bronchial cell type fractions as provided by Okuda et al. [20], Figure 1C; *n* = 37); 3 — bronchioles (bronchiolar cell type fractions as provided by Okuda et al. [20], Figure 1C; *n* = 38); 4 — alveoli (alveolar cell type fractions roughly estimated from Vieira Braga et al. [22], Figure 1D, parenchyma pie chart, *n* = 10; small fraction of identified multiciliated cells is not shown). IF images taken with permission from LungImage, Cincinnati Children's Hospital Medical Center. (**B**) Some of the main marker genes of canonical lung epithelial cell types [27,36]. *Key*: black dotted outline — cell types constitute less than 1% of the total sampled cell population but were included for illustration purposes; *SOX2* — marker of tracheal, bronchial and bronchiolar epithelial cells; BC — basal cell; GC — goblet cell; MC — mucous cells; MCC — multiciliated cell; SMG — submucosal gland (not shown in bronchi); CC — club (Clara) cell; TC — tuft (brush) cell; PNEC — pulmonary neuroendocrine cell (green lines — nerves); GC — goblet cell; AT1 — alveolar type 1 cell; AT2 — alveolar type 2 cell; MEC — myoepithelial cell; MC — mucous cell; SC — serous cell (green lines — nerves); *SFTP** — *SFTPB*, *SFTPC*, *SFTPD*.

PORTLAND PRESS

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572



(D)	Study	Deprez *et al.* (28)	Okuda *et al.* (20)	Travaglini *et al.* (27)	Ruiz Garcia *et al.* (89)	Goldfarbmuren *et al.* (19)
	Cell dissociation reagent	*B. licheniformis* protease	Accutase, DNase I, Collagenase IV	Liberase DL, elastase, DNase I	*S. griseus* type IV protease, DNase I	Accutase
	scRNA-seq technique	3'-end, droplet-based	3'-end, droplet-based	1) 3'-end, droplet-based 2) Full length, plate-based	3'-end, droplet-based	3'-end, plate-based
	Platform/method	Chromium (10X)	1) Chromium (10X) 2) DROP-seq	1) Chromium (10X) 2) SmartSeq2	Chromium (10X)	WaferGen
Average value per sample (unless specified otherwise)	*Samples*	All	All	All	Pneumacult_ALI28 (D275) Pneumacult_ALI28 (D389)	All
	Estimated cell number	2350.09	NA	1) Total: 65 662 2) Total: 9 774	3319.00	Total: 5976
	Mean reads per cell	103480.69	NA	NA	71273.00	NA
	Median genes per cell	1747.97	***1) 3500-6500 ***2) 400-3000	*** > 500	3073.00	*** > 1500
	Median UMI counts per cell	5738.51	NA	***1) > 1000 UMIs	14623.00	NA

Figure 5. Correlations of gene expression profiles in specific cell populations. Part 1 of 2

Extent of the overlap in gene expression between populations of suprabasal (**A**), proliferating basal (**B**) and deuterosomal (**C**) cells among top 99 marker genes (based on positive log fold change values) identified in multiple scRNA-seq studies, with metadata available on these cell populations

Figure 5. Correlations of gene expression profiles in specific cell populations. Part 2 of 2

provided for each paper (**D**). *Key*: (**A**) (suprabasal cells): Okuda et al. [20] — suprabasal cluster visualised on https://cells.ucsc.edu/?ds=lung-airway+boucher-epithelium; Deprez et al. [28] — Supplementary table E7; Ruiz García et al. [89] — Supplementary table S1; (**B**) (proliferating basal cells): Travaglini et al. [27] — Supplementary table 4; Goldfarbmuren et al. [19] — 'Source data table 28 for Figure 5b' in the original paper; Ruiz García et al. [89] — Supplementary table S4; (**C**) (deuterosomal cells): Deprez et al. [28] — Supplementary table E7; Goldfarbmuren et al. [19] — 'Source data table 28 for Figure 5b' in the original paper; Ruiz García et al. [89] — Supplementary table S4; * — cell population was referred to by authors [19] as '*FOXN4* early ciliating'; ** — expression of *CDC20B* in deuterosomal cells was reported by Deprez et al. [28] and Goldfarbmuren et al. [19], while Ruiz García et al. [89] only reported expression of *CDC20Bshort*; *** — only cells with specified number of genes or unique molecular identifiers (UMIs) were used in downstream scRNA-seq analyses; HBEC — human bronchial epithelial cells; diff. — differentiated; ALI — air liquid interface. Figure produced with R package *ggVennDiagram* [90].

following re-annotation of clusters from integrated datasets and input from experts in the respiratory research field [32]. If implemented as an additional quality control (QC) step, e.g. the number and identities of clusters from scRNA-seq experiments can be cross-checked against a known number of cell types, their relative proportions and expression profiles identified with scRNA-seq within the tissue of interest, such reference annotations will be of great value. The long-term success of this approach, however, will depend on the capacity of consortia to incorporate more datasets as they are published, to regularly evaluate the toolkit available for data analysis and integration, and if necessary re-run pipelines, as well as to set and dynamically adjust, with minimum bias from expert panels, the fundamental criteria for defining 'cell types', 'cell states' etc. For instance, given that all cells from a particular lineage share expression of several marker genes, e.g. transcription factor SOX2 is expressed by all epithelial cells in the proximal airways [92], a threshold needs to be established for the number of genes that need to be expressed exclusively in a subpopulation of a canonical cell type for it to be recognised as a novel cell type rather than a more specialised subtype or a transient state induced by a particular cellular or experimental condition. In addition, the relative weights of parameters such as the extent of differential expression for genes expressed in both of the compared cell populations and the frequency threshold of these populations among individuals will need to be provided. Lastly, datasets from spatial transcriptomics and other high-throughput single-cell analyses will introduce additional layers of complexity, but their incorporation into cell atlases is expected to increase the confidence in cellular identities determined with scRNA-seq.

Cell fates

Since the first use of microscopy-based direct observation methods on organisms with determinate patterns of cell fate, the primary toolkit for studying the history of cell divisions from an ancestor cell to its terminally differentiated descendants, prospective lineage-tracing, has advanced from techniques involving tissue transplantation, usage of tracing dyes and transgenic fluorescent reporters [93] to high-throughput approaches facilitated by next generation sequencing (NGS). For instance, genomes of each cell in a population can now be uniquely barcoded via transgenic integration, *in vivo* recombination or live editing, thereby enabling reconstruction of cell lineages following DNA or RNA-sequencing [94–96].

Notwithstanding the rapid progress in prospective analyses, none of them can be conducted in humans and until recently, most of the information about human cell lineages has been either inferred from genetic lineage-tracing experiments in model organisms or obtained from *in vitro* differentiation of primary cells and retrospective lineage-tracing studies, in which phylogenetic lineages of cells are reconstructed on the basis of somatic mutations [97], particularly in microsatellite regions [9]. With the advent of scRNA-seq, not only a novel method of clonal tracking based on mutations and chromatin accessibility changes of mitochondrial DNA has been proposed [98], but also several *in silico* approaches, known as trajectory inference methods, have been developed for the determination of the relative developmental history of cells in a studied population [73] (Figure 2). As for the principles of key methods, pseudotime analysis attempts to replicate the temporal aspect of differentiation or any event resulting in changes in cell states or types by positioning cells or clusters along a pseudo-temporal axis or axes based on the degree of similarity in gene expression [67,73]. It should be noted, however, that pseudotime is not a chronological or unidirectional measure and akin to many other trajectory analysis tools, it is only intended to provide a framework for inferring directionality [99] on the basis of a biological hypothesis. A more recent type of analysis — RNA velocity, which is often classified as a trajectory inference method despite being an inherently unidirectional measure, identifies transitional cell states and

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

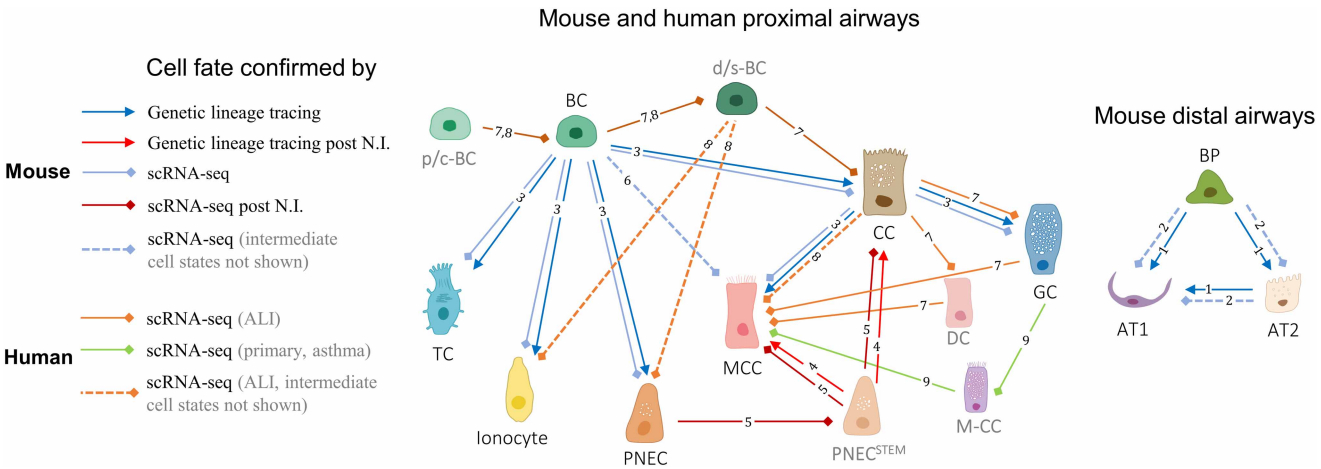


Figure 6. Cellular relationships in the lower airways.

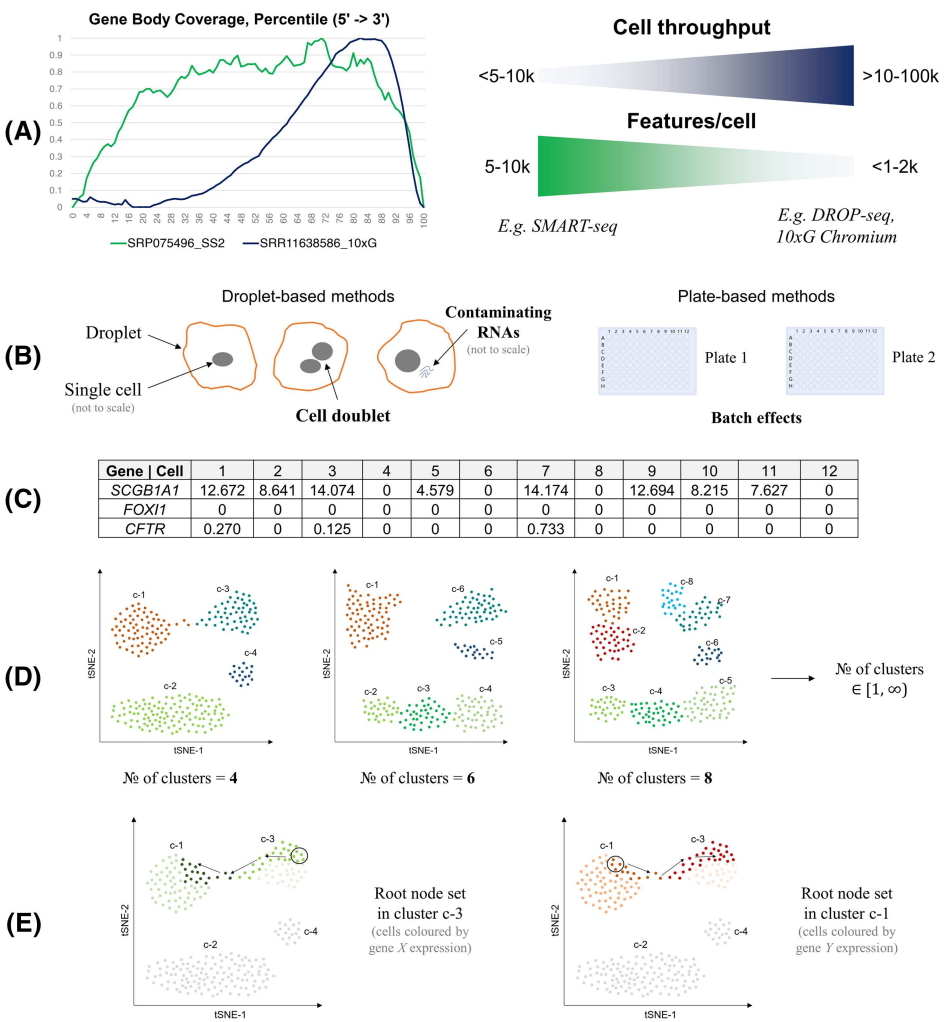
Simplified tree of the epithelial cell fates in the lower airways of mice and humans based on the recent lineage-tracing and scRNA-seq data (submucosal gland cell types and some de-differentiation pathways are not shown). *Key*: black labels — cell types; grey labels — putative cell states; N.I. — naphthalene-induced injury; p/c-BC — proliferating/cycling basal cell; BC — basal cell; d/s-BC — differentiating basal or suprabasal cell; BP — bipotent progenitor cell; AT1 — alveolar type 1 cell; AT2 — alveolar type 2 cell; CC — club cell; GC — goblet cell; DC — deuterosomal cell; M-CC — mucous ciliated cell; MCC — multiciliated cell; TC — tuft cell; TC-like — tuft-like cell; PNEC — pulmonary neuroendocrine cell; PNEC^{STEM} — pulmonary neuroendocrine stem cell. References: 1 — Desai et al. (2014) [104]; 2 — Treutlein et al. (2014) [105], 3 — Montoro et al. (2018) [17]; 4 — Song et al. (2012) [106]; 5 — Ouadah et al. (2019) [107]; 6 — Byrnes et al. (2022) [108]; 7 — Ruiz García et al. (2019) [89]; 8 — Goldfarbmuren et al. (2020) [19]; 9 — Vieira Braga et al. (2019) [22].

positions them along a temporal axis based on the ratios of spliced to unspliced mRNAs [100]. Efficient integration of these two methods [101] and coupling of scRNA-seq with a dual fluorescent reporter-gene based system that can provide an absolute temporal scale of differentiation have also been reported [102].

Epithelial cell lineages in healthy and injured airways of the most common small animal model of the human respiratory system, the house mouse [103], were mainly discovered using genetic lineage-tracing techniques [83] (Figure 6). In the distal mouse lung, alveolar type 1 and 2 epithelial cells were found to arise from a common bipotent progenitor, with the latter cell type being capable of regenerating the former [104], which was subsequently confirmed with scRNA-seq [105]. In a later study, Montoro et al. [17] performed both scRNA-seq of murine tracheal epithelial cells, in which common and rare epithelial cell types were identified, and *in vivo* genetic lineage-tracing providing temporal resolution of differentiation. Their results aligned with the existing knowledge about progenitors of canonical epithelial cell types and showed that populations of ionocytes, tuft and pulmonary neuroendocrine cells are likely derived from basal cells. The ability of mouse pulmonary neuroendocrine cells to replenish club and MCCs after naphthalene-induced injury, which was initially discovered using genetic lineage-tracing [106], was confirmed with scRNA-seq [107]. Byrnes et al. [108] also showed that differentiation of murine basal cells into MCCs may occur through an intermediate precursor cell in the absence of Notch signalling.

Results of both pseudotime and RNA velocity analyses in scRNA-seq studies of the human airways largely confirmed some of the proposed airway epithelial cell lineages [109] and led to discoveries of potentially novel developmental intermediates and differentiation routes [36], the presence of most of which still needs to be experimentally validated *in vivo* (Figure 6). For example, using nasal epithelial cells differentiated at the ALI, Ruiz García et al. [89] showed that differentiation of club cells into MCCs might occur through a developmental intermediate, the deuterosomal cell, and that goblet cells might act as direct precursors of MCCs. Goldfarbmuren et al. [19], in turn, used the same *in vitro* differentiation model established with tracheal epithelial cultures and identified distinct populations of 'early' and 'later ciliating' cells that might precede mature MCCs. As for scRNA-seq conducted directly on human bronchial biopsies, Vieira Braga et al. [22] detected a population of mucous ciliated cells, which were mainly present in asthmatic individuals and expressed markers of both mature multiciliated and goblet cells, that were proposed to eventually acquire a goblet cell fate.

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572



Gene | Cell	1	2	3	4	5	6	7	8	9	10	11	12
SCGB1A1	12.672	8.641	14.074	0	4.579	0	14.174	0	12.694	8.215	7.627	0
FOXI1	0	0	0	0	0	0	0	0	0	0	0	0
CFTR	0.270	0	0.125	0	0	0	0.733	0	0	0	0	0

a1) If isoform analysis or detection of genes expressed at low level required → **full-length scRNA-seq protocols** (e.g. SS3)

a2) Low cell throughput for some scRNA-seq downstream analyses (DGE) can be mitigated → **counts pseudo-bulking**

a3) If cell type/state of interest can be reliably enriched (e.g. with FACS) → **bulk RNA-seq in parallel to scRNA-seq**

b1) Cell doublets → **exclusion of cells with very high UMI counts from analyses**

b2) Contaminating RNAs → ***in silico* prediction and removal**

b3) Batch effects → ***in silico* correction**

c1) Elevated zero expression count values → **validation with sc ISH quantitative imaging and sc-sorted qRT-PCR**

d1) In theory, any number of cell clusters can be set by user or determined in an unsupervised manner → **actual number of clusters will likely be based on prior knowledge or biological hypothesis**

d2) Cell cluster does not imply a cell type, but it can be one → **consensus definitions or reference annotations are needed**

e1) Majority of trajectory analyses are not unidirectional or chronological → **on their own, they do not infer directionality but provide a framework for inferring it**

e2) Any cell can be set as a root node → **actual starting cell type for a temporal process can only be determined experimentally (e. g. lineage tracing)**

Figure 7. Limitations of sc-RNA-seq.

Common limitations of current scRNA-seq protocols and downstream analyses alongside with the general strategies for avoiding pitfalls or mitigating known issues. *Key*: (**A**) (technical limitations of sequencing with examples of protocols) — in contrast with full-length scRNA-seq protocols such as SMART-seq2 and 3 (SS2 and SS3), 3′-end scRNA-seq protocols, e.g. DROP-seq and 10× Genomics Chromium (10×G), generally tend to prioritise cell throughput at the expense of gene body coverage and often, number of features (genes or unique molecular identifies) detected per cell (gene body coverage plots generated from data, which were downloaded using provided accession numbers, with RSeQC package, *geneBody_coverage2.py* [110]; (**B**) technical limitations of droplet- (e.g. DROP-seq) and plate-based scRNA-seq methods (e.g. SMART-seq); (**C**) (post transcript/gene quantification) — some scRNA-seq protocols may underestimate expression and cell type distribution of expression of certain genes whose transcripts are present at low copy number within the cells; (**D**) (examples of tiered clustering on a hypothetical scRNA dataset) — t-distributed stochastic neighbour embedding (t-SNE) plot, with cells coloured by cluster; (**E**) examples of a trajectory analysis on a hypothetical scRNA dataset (arrows indicate potential trajectory of development or differentiation); sc — single cell; ISH — *in situ* hybridisation; FACS — fluorescence-activated cell sorting; DGE — differential gene expression.

Limitations of scRNA-seq and good practice

While duly acknowledging continuous improvements in throughput, sensitivity and ease of data manipulation, which has led to the current state of scRNA-seq (Figures 1 and 2), as well as appreciating insights obtained by analysing transcriptomes at a single-cell resolution, it is worth reiterating that scRNA-seq remains a highly nuanced technology that is still in the active stage of development. The quality of data from scRNA-seq

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

experiments is affected by a wide range of factors, some of which can be bypassed or in contrast — exacerbated in various protocols and adaptations of scRNA-seq, such as single-nuclei RNA-seq and scRNA-seq of formalin-fixed samples which are not discussed here. Broadly speaking, however, limitations of scRNA-seq can be divided into multiple categories depending on the stage of the experiment or data analysis at which they arise (Figure 7).

When discussing issues specific to scRNA-seq, it should be noted that prior to single-cell isolation and RNA extraction, common limitations associated with sample sourcing and preparation resemble those encountered in bulk RNA-seq [111]. When target cells are obtained from a brushing, biopsy or autopsy, parameters such as sites of sampling from a single donor, physiological differences between donors from the same cohort, sample handling and the duration of the period from sampling to sequencing may lead to substantial variability in the results, thereby reducing reproducibility of the experiment and undermining the statistical power. On the other hand, the practice of culture and differentiation of primary cells before sequencing introduces other types of biases since factors ranging from the type of media and plasticware to experimental technique and duration of the experiment can affect both cell type presence and distribution at the point of sampling. For instance, Ruiz García et al. [89] reported that Lonza bronchial epithelial cell growth medium favoured differentiation of MCCs but not goblet cells, whereas the use of PneumaCult-ALI medium resulted in the presence of both cell types in numbers and distributions that were more physiologically relevant. As for the confounders common to samples prepared from any source, the most frequently used cell dissociation protocols involve proteases, which not only can cause cell death of certain cell types or states but can also perturb transcriptomes of profiled cells, leading to the up-regulation of apoptosis- and stress-related genes, e.g. *FOS* and *ATF3* [112]. In bulk RNA-seq this issue can be avoided by bypassing cell dissociation and adding RNA extraction reagents (e.g. TRIzol) directly to the sample or on top of the cell layer, whereas in scRNA-seq, protocol modifications, such as the use of cold active proteases [113], and *in silico* filtering steps have been suggested to mitigate dissociation-induced effects on the gene expression in scRNA-seq.

Technical limitations in commonly used scRNA-seq protocols include cell throughput and transcript coverage by Illumina sequencing (Figure 7A). For example, most of the droplet-based high-throughput scRNA-seq platforms, such as 10× Genomics Chromium, provide almost no information on differential isoform expression as only roughly 100 bp section of the 3′-end of the transcript is sequenced. In addition, it is estimated that only 10–15% of all cellular mRNAs are captured and reverse transcribed in a typical scRNA-seq experiment [65]. Even through variable mRNA capture efficiency can be normalised for with addition of 'spike-in' RNA molecules, amounts and other characteristics of which are known [114], this figure implies that only transcripts of the most highly expressed genes are captured. The impact of this can be even more drastic when read depth in terms of genes or UMIs per cell is low. On top of that, the most frequently used technique of RNA selection, oligo-dT enrichment, is 3′-end biased and leaves out non-polyadenylated RNAs, which include short and long non-coding RNAs. Alternative methods for rRNA depletion in scRNA-seq have been reported [115], but none has been widely implemented to date.

Other technical issues in scRNA-seq are also platform-dependent (Figure 7B). For instance, cell doublets — cases when mRNAs from two different cells are captured and tagged with the same barcode within the same droplet, are common in droplet-based protocols such as DROP-seq. They can lead to unique but artificial expression signatures. Doublets can be removed by using strict upper UMI thresholds or applying filters from *in silico* simulations [65]. Another issue for droplet-based protocols is release of RNA molecules from lysed cells and their colocalization within droplets containing single cells. Currently, contaminating RNA molecules can be predicted and filtered out using various *in silico* tools [74,75]. Data from plate-based scRNA methods, in turn, often requires batch correction due to the impact of variability in timing during individual plate processing. Overall, high degree of technical variability in scRNA-seq experiments can hinder direct comparisons and lower the quality of data integration from different platforms. This holds true not only for highly customisable scRNA-seq platforms established in academic settings, but also for commercial ones. To exemplify the latter, scRNA-seq analysis of the human pancreatic islets conducted by Wang et al. [12] on Fluidigm C1 96/HT, Clontech iCell8 and 10× Genomics Chromium showed that both relative fractions of cells from each cell type that was detected, and their DEGs differed significantly between the evaluated platforms.

With Illumina as an NGS platform of choice in scRNA-seq, analysis of generated data is limited by factors, such as ambiguous mapping of short reads, that are applicable to any method involving short-read sequencing [5], but are not discussed here. As for the issues specific to scRNA-seq, it has been reported that the number of genes with expression values of zero caused by technical reasons varies dramatically between the cells and that

there are more dropouts in scRNA-seq than expected, particularly among genes with lower expression levels [71] (Figure 7C). For example, Okuda et al. [20] reported that two 3′-end scRNA-seq methods underestimated the number of CFTR-expressing cells among bronchial epithelial cells by roughly a factor of 10.

Apart from technical and sequencing aspects, choice of methods from a wide range of available platform-specific and open-source tools for data analysis steps starting from generation of gene expression count matrices to clustering [72] and trajectory inference [73], may cause differences in results, including cases when data from the same scRNA experiment is re-analysed. Even if a custom end-to-end or default data analysis pipeline is chosen, user input is still required at nearly every stage and can have a substantial impact on the outcomes. For instance, by setting strict threshold values for gene or UMI counts and fraction of mitochondrial reads per cell in order to filter out dead cells during the QC step, novel but rare cell types may be missed [65]. Manual input based on a biological hypothesis or prior knowledge is also crucial for clustering, trajectory inference analyses and annotation of cell types (Figure 7D,E). When unsupervised methods are used or when reference annotations are absent, it is important to first evaluate a range of tools and compare the outcomes to datasets generated using the same protocol, from the same tissue or organism.

Conclusions and future outlook

On the cusp of technologies capable of elucidating various facets of cellular identity at a resolution of individual cells, scRNA-seq has not only experienced a phase of rapid development, scaling and adoption, but also vividly showcased the true potential of single-cell transcriptomics in the respiratory science and other research fields. Since its inception, scRNA-seq has been front-running the discovery of novel cell types, states and fates by drastically increasing the scope of transcriptome that can be profiled with less bias and more accuracy by eliminating the barrier of averaged gene expression inherent to both microarray analyses and bulk RNA-seq.

Nonetheless, scRNA-seq is not devoid of constraints and challenges. Higher sensitivity makes it susceptible to numerous sources of biological noise that are found in cellular environments, such as transcriptional bursts [40] and oscillations in cell states driven by cell cycle, responses to regular or sporadic internal and external stimuli [116], which create stochastic variation in gene expression between cells that may not be biologically meaningful in the context of a particular study. The extent of true biological variation may also be hidden or skewed by batch and sampling effects, whereas technical sources of variation can arise from different designs of scRNA-seq platforms and custom protocol modifications. In addition, the introduction of frameworks outlining the scope of experimental and data analysis standards as well as the implementation of consensus nomenclature for naming and characterising cell populations discovered with scRNA-seq are lagging behind the technological progress, which makes direct cross-study comparisons and integration of data difficult. Use of reference annotation data for canonical and rare cell types provided by consortia such as HLCA [31,32] may offer a promising way forward, but will still require answers to fundamental questions raised in this review as values for a broad range of parameters, which are likely to be tissue- or organism-specific, need to be determined.

Apart from using reference annotations, the confidence in conclusions drawn about the observed cell types, states and fates on the basis of any scRNA-seq data can be improved by benchmarking the results against the findings of conventional transcriptomic, proteomic and genetic tools. Continuous method development and evaluation for integration and consolidation of scRNA-seq findings will not only facilitate incorporation of data from other emerging single-cell technologies [117,118], including spatial single-cell transcriptomics [64,119] and long-read scRNA-seq for isoform discovery and differential isoform expression analyses [61], but will also pave the way for the construction of truly comprehensive maps of cellular identities.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Open Access

Open access for this article was enabled by the participation of University of Sheffield in an all-inclusive *Read & Publish* agreement with Portland Press and the Biochemical Society under a transformative agreement with JISC.

CRediT Author Contribution

Colin David Bingle: Conceptualisation, Supervision, Project administration, Writing — review and editing. **Oleksandr Dudchenko**: Conceptualisation, Formal analysis, Investigation, Writing — original draft, Writing — review and editing. **Jose Ordovoaes-Montanes**: Writing — original draft, Writing — review and editing.

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

Abbreviations

ALI, air–liquid interface; cDNA, complementary DNA; CF, cystic fibrosis; CFTR, cystic fibrosis trans-membrane conductance regulator; COVID-19, coronavirus disease 2019; DEGs, differentially expressed genes; HBECs, human bronchial epithelial cells; HLCA, human lung cell atlas; MCCs, multiciliated cells; NGS, next generation sequencing; SARS-CoV-2, severe acute respiratory syndrome coronavirus clade 2; scRNA-seq, single-cell RNA-sequencing; UMI, unique molecular identifier.

References

1 Iida, K. and Nishimura, I. (2002) Gene expression profiling by DNA microarray technology. *Crit. Rev. Oral Biol. Med.* **13**, 35–50 https://doi.org/10.1177/154411130201300105
2 Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 https://doi.org/10.1038/nrg2484
3 VanGuilder, H.D., Vrana, K.E. and Freeman, W.M. (2018) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* **44**, 619–626 https://doi.org/10.2144/000112776
4 Wheelan, S.J., Murillo, F.M. and Boeke, J.D. (2008) The incredible shrinking world of DNA microarrays. *Mol. Biosyst.* **4**, 726–732 https://doi.org/10.1039/b706237k
5 Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 https://doi.org/10.1038/s41576-019-0150-2
6 Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N. et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 https://doi.org/10.1038/nmeth.1315
7 Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R. et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 https://doi.org/10.1038/nbt.2282
8 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M. et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 https://doi.org/10.1016/j.cell.2015.05.002
9 Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 https://doi.org/10.1038/nrg3542
10 Aldridge, S. and Teichmann, S.A. (2020) Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 1 https://doi.org/10.1038/s41467-020-18158-5
11 Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 https://doi.org/10.1038/nprot.2017.149
12 Wang, Y.J., Schug, J., Lin, J., Wang, Z., Kossenkov, A. and Kaestner, K.H. (2019) Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. *bioRxiv* https://doi.org/10.1101/541433
13 Ballestar, E., Farber, D.L., Glover, S., Horwitz, B., Meyer, K., Nikolić, M. et al. (2020) Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. *bioRxiv* https://doi.org/10.1101/2020.11.20.20227355
14 Ziegler, C.G.K., Allon, S.J., Nyquist, S.K., Mbano, I.M., Miao, V.N., Tzouanas, C.N. et al. (2020) SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* **181**, 1016–1035.e19 https://doi.org/10.1016/j.cell.2020.04.035
15 Zhang, H., Rostami, M.R., Leopold, P.L., Mezey, J.G., O'Beirne, S.L., Strulovici-Barel, Y. et al. (2020) Expression of the SARS-CoV-2 ACE2 receptor in the human airway epithelium. *Am. J. Respir. Crit. Care Med.* **202**, 219–229 https://doi.org/10.1164/rccm.202003-0541OC
16 Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F. et al. (2020) COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 https://doi.org/10.1038/s41587-020-0602-4
17 Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E. et al. (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 https://doi.org/10.1038/s41586-018-0393-7
18 Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G. et al. (2018) A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 https://doi.org/10.1038/s41586-018-0394-6
19 Goldfarbmuren, K.C., Jackson, N.D., Sajuthi, S.P., Dyjack, N., Li, K.S., Rios, C.L. et al. (2020) Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat. Commun.* **11**, 2485 https://doi.org/10.1038/s41467-020-16239-z
20 Okuda, K., Dang, H., Kobayashi, Y., Carraro, G., Nakano, S., Chen, G. et al. (2020) Secretory cells dominate airway CFTR expression and function in human airway superficial epithelia. *Am. J. Respir. Crit. Care Med.* **203**, 1275–1289 https://doi.org/10.1164/rccm.202008-3198OC
21 Ordovas-Montanes, J., Dwyer, D.F., Nyquist, S.K., Buchheit, K.M., Vukovic, M., Deb, C. et al. (2018) Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 https://doi.org/10.1038/s41586-018-0449-8
22 Braga F.A., V., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M. et al. (2019) A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 https://doi.org/10.1038/s41591-019-0468-5
23 Jackson, N.D. (2020) Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. *Cell Rep.* **32**, 107872 https://doi.org/10.1016/j.celrep.2020.107872
24 Xu, Y., Mizuno, T., Sridharan, A., Du, Y., Guo, M., Tang, J. et al. (2016) Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 https://doi.org/10.1172/jci.insight.90558
25 Carraro, G., Mulay, A., Yao, C., Mizuno, T., Konda, B., Petrov, M. et al. (2020) Single-cell reconstruction of human basal cell diversity in normal and idiopathic pulmonary fibrosis lungs. *Am. J. Respir. Crit. Care Med.* **202**, 1540–1550 https://doi.org/10.1164/rccm.201904-0792OC
26 Carraro, G., Langerman, J., Sabri, S., Lorenzana, Z., Purkayastha, A., Zhang, G. et al. (2021) Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. *Nat. Med.* **27**, 806–814 https://doi.org/10.1038/s41591-021-01332-7
27 Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V. et al. (2020) A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 https://doi.org/10.1038/s41586-020-2922-4

28 Deprez, M., Zaragosi, L.E., Truchi, M., Becavin, C., Ruiz García, S., Arguel, M.J. et al. (2020) A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 https://doi.org/10.1164/rccm.201911-2199OC
29 Kadur Lakshminarasimha Murthy, P., Sontake, V., Tata, A., Kobayashi, Y., Macadlo, L., Okuda, K. et al. (2022) Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature* **604**, 111–119 https://doi.org/10.1038/s41586-022-04541-3
30 Guo, M., Morley, M.P., Wu, Y., Du, Y., Zhao, S., Wagner, A. et al. (2022) Guided construction of single cell reference for human and mouse lung. *bioRxiv*
31 Luecken, M.D., Zaragosi, L.E., Madissoon, E., Sikkema, L., Firsova, A.B., De Domenico, E. et al. (2022) The discovAIR project: a roadmap towards the human lung cell atlas. *Eur. Respir. J.* **60**, 2102057 https://doi.org/10.1183/13993003.02057-2021
32 Sikkema, L., Ramírez-Suástegui, C., Strobl, D.C., Gillett, T.E., Zappia, L., Madissoon, E. et al. (2023) An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 https://doi.org/10.1038/s41591-023-02327-2
33 Du, Y., Ouyang, W., Kitzmiller, J.A., Guo, M., Zhao, S., Whitsett, J.A. et al. (2021) Lung gene expression analysis web portal version 3: lung-at-a-glance. *Am. J. Respir. Cell Mol. Biol.* **64**, 146–149 https://doi.org/10.1165/rcmb.2020-0308LE
34 Chambers, D.C., Carew, A.M., Lukowski, S.W. and Powell, J.E. (2018) Transcriptomics and single-cell RNA-sequencing. *Respirology* **24**, 29–36 https://doi.org/10.1111/resp.13412
35 Alexander, M.J., Budinger, G.R.S. and Reyfman, P.A. (2020) Breathing fresh air into respiratory research with single-cell RNA sequencing. *Eur. Respir. Rev.* **29**, 200060 https://doi.org/10.1183/16000617.0060-2020
36 Zaragosi, L.E., Deprez, M. and Barbry, P. (2020) Using single-cell RNA sequencing to unravel cell lineage relationships in the respiratory tract. *Biochem. Soc. Trans.* **48**, 327–336 https://doi.org/10.1042/BST20191010
37 Hewitt, R.J. and Lloyd, C.M. (2021) Regulation of immune responses by the airway epithelial cell landscape. *Nat. Rev. Immunol.* **21**, 347–362 https://doi.org/10.1038/s41577-020-00477-9
38 Clevers, H. (2017) What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Syst.* **4**, 255–259 https://doi.org/10.1016/j.cels.2017.03.006
39 Zeng, H. (2022) What is a cell type and how to define it? *Cell* **185**, 2739–2755 https://doi.org/10.1016/j.cell.2022.06.031
40 Raj, A. and van Oudenaarden, A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 https://doi.org/10.1016/j.cell.2008.09.050
41 Luecken, M.D. and Theis, F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, 6 https://doi.org/10.15252/msb.20188746
42 Jia, Q., Chu, H., Jin, Z., Long, H. and Zhu, B. (2022) High-throughput single-cell sequencing in cancer research. *Signal Transduct. Target. Ther.* **7**, 145 https://doi.org/10.1038/s41392-022-00990-4
43 Picelli, S., Faridani, O.R., Björklund, Å.K, Winberg, G., Sagasser, S. and Sandberg, R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 https://doi.org/10.1038/nprot.2014.006
44 Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.J., Larsson, A.J.M. et al. (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 https://doi.org/10.1038/s41587-020-0497-0
45 Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P. et al. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 https://doi.org/10.1101/gr.110882.110
46 Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. et al. (2012) Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 https://doi.org/10.1038/nprot.2012.022
47 Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012) CEL-Seq: single-sell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 https://doi.org/10.1016/j.celrep.2012.08.003
48 Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L. et al. (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 1 https://doi.org/10.1186/s13059-016-0938-8
49 Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I. et al. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 https://doi.org/10.1126/science.1247651
50 Keren-Shaul, H., Kenigsberg, E., Jaitin, D.A., David, E., Paul, F., Tanay, A. et al. (2019) MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nat. Protoc.* **14**, 1841–1862 https://doi.org/10.1038/s41596-019-0164-4
51 Fan, H.C., Fu, G.K. and Fodor, S.P.A. (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 https://doi.org/10.1126/science.1258367
52 Saliba, A.E., Westermann, A.J., Gorski, S.A. and Vogel, J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 https://doi.org/10.1093/nar/gku555
53 Picelli, S. (2016) Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol.* **14**, 637–650 https://doi.org/10.1080/15476286.2016.1201618
54 Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M. et al. (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 https://doi.org/10.1016/j.molcel.2017.01.023
55 Chen, G., Ning, B. and Shi, T. (2019) Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **10**, 317 https://doi.org/10.3389/fgene.2019.00317
56 Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. et al. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 https://doi.org/10.1038/nmeth.1778
57 Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 https://doi.org/10.1101/gr.209601.116
58 Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C. et al. (2017) Power analysis of single-cell RNA-Sequencing experiments. *Nat. Methods* **14**, 381–387 https://doi.org/10.1038/nmeth.4220
59 Ziegenhain, C., Hendriks, G.J., Hagemann-Jensen, M. and Sandberg, R. (2022) Molecular spikes: a gold standard for single-cell RNA counting. *Nat. Methods* **19**, 560–566 https://doi.org/10.1038/s41592-022-01446-x

Biochemical Journal (2023) **480** 921–939
https://doi.org/10.1042/BCJ20220572

60 Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 https://doi.org/10.1038/s41576-020-0236-x
61 Al'Khafaji, A.M., Smith, J.T., Garimella, K.V., Babadi, M., Popic, V., Sade-Feldman, M. et al. (2023) High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01815-7
62 Philpott, M., Watson, J., Thakurta, A., Brown, T., Brown, T., Oppermann, U. et al. (2021) Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat. Biotechnol.* **39**, 1517–1520 https://doi.org/10.1038/s41587-021-00965-w
63 Healey, H.M., Bassham, S. and Cresko, W.A. (2022) Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics* **220**, iyac017 https://doi.org/10.1093/genetics/iyac017
64 Moses, L. and Pachter, L. (2022 May) Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 https://doi.org/10.1038/s41592-022-01409-2
65 Wu, Y. and Zhang, K. (2020) Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol.* **16**, 408–421 https://doi.org/10.1038/s41581-020-0262-0
66 Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A. et al. (2021) Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 https://doi.org/10.1016/j.cell.2021.04.048
67 Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M. et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 https://doi.org/10.1038/nbt.2859
68 Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1 https://doi.org/10.1186/s13059-017-1381-1
69 Zappia, L., Phipson, B. and Oshlack, A. (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 https://doi.org/10.1371/journal.pcbi.1006245
70 Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 https://doi.org/10.1038/nbt.3519
71 Hicks, S.C., Townes, F.W., Teng, M. and Irizarry, R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 https://doi.org/10.1093/biostatistics/kxx053
72 Duò, A., Robinson, M.D. and Soneson, C. (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**, 1141 https://doi.org/10.12688/f1000research.15666.2
73 Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 https://doi.org/10.1038/s41587-019-0071-9
74 Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M. et al. (2020) Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 https://doi.org/10.1186/s13059-020-1950-6
75 Young, M.D. and Behjati, S. (2020) Soupx removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, giaa151 https://doi.org/10.1093/gigascience/giaa151
76 Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A. and Theis, F.J. (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 https://doi.org/10.1038/s41592-018-0254-1
77 Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T. et al. (2022) Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 https://doi.org/10.1126/science.abl5197
78 Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M. et al. (2022) Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 https://doi.org/10.1038/s41587-021-01001-7
79 Fischer, D.S., Dony, L., König, M., Moeed, A., Zappia, L., Heumos, L. et al. (2021) Sfaira accelerates data and model reuse in single cell genomics. *Genome Biol.* **22**, 248 https://doi.org/10.1186/s13059-021-02452-6
80 Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A. et al. (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 https://doi.org/10.1038/s41592-020-01050-x
81 Trapnell, C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 https://doi.org/10.1101/gr.190595.115
82 Vickaryous, M.K. and Hall, B.K. (2006) Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.* **81**, 425–455 https://doi.org/10.1017/S1464793106007068
83 Tata, P.R. and Rajagopal, J. (2017) Plasticity in the lung: making and breaking cell identity. *Development* **144**, 755–766 https://doi.org/10.1242/dev.143784
84 Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H. et al. (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 https://doi.org/10.1038/nrg.2016.127
85 Zheng, H.B., Doran, B.A., Kimler, K., Yu, A., Tkachev, V., Niederlova, V. et al. (2021) A treatment-naïve cellular atlas of pediatric Crohn's disease predicts disease severity and therapeutic response. *medrxiv* https://doi.org/10.1101/2021.09.17.21263540
86 Quigley, I.K., Stubbs, J.L. and Kintner, C. (2011) Specification of ion transport cells in the Xenopus larval skin. *Development* **138**, 705–714 https://doi.org/10.1242/dev.055699
87 Esaki, M., Hoshijima, K., Nakamura, N., Munakata, K., Tanaka, M., Ookata, K. et al. (2009) Mechanism of development of ionocytes rich in vacuolar-type H+-ATPase in the skin of zebrafish larvae. *Dev. Biol.* **329**, 116–129 https://doi.org/10.1016/j.ydbio.2009.02.026
88 Strunz, M., Simon, L.M., Ansari, M., Kathiriya, J.J., Angelidis, I., Mayr, C.H. et al. (2020) Alveolar regeneration through a Krt8 + transitional stem cell state that persists in human lung fibrosis. *Nat. Commun.* **11**, 3559 https://doi.org/10.1038/s41467-020-17358-3
89 Ruiz García, S., Deprez, M., Lebrigand, K., Cavard, A., Paquet, A., Arguel, M.J. et al. (2019) Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* **146**, dev.177428 https://doi.org/10.1242/dev.177428
90 Gao, C.H., Yu, G. and Cai, P. (2021) Ggvenndiagram: an intuitive, easy-to-use, and highly customizable R package to generate Venn diagram. *Front. Genet.* **12**, 706907 https://doi.org/10.3389/fgene.2021.706907
91 Bakken, T., Cowell, L., Aevermann, B.D., Novotny, M., Hodge, R., Miller, J.A. et al. (2017) Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinform.* **18**, S17 https://doi.org/10.1186/s12859-017-1977-1

92 Mollaoglu, G., Jones, A., Wait, S.J., Mukhopadhyay, A., Jeong, S., Arya, R. et al. (2018) The lineage-defining transcription factors SOX2 and NKX2-1 determine lung cancer cell fate and shape the tumor immune microenvironment. *Immunity* **49**, 764–779.e9 https://doi.org/10.1016/j.immuni.2018.09.020
93 Kretzschmar, K. and Watt, F.M. (2012) Lineage tracing. *Cell* **148**, 33–45 https://doi.org/10.1016/j.cell.2012.01.002
94 Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J. and van Oudenaarden, A. (2018) Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 https://doi.org/10.1038/nature25969
95 Baron, C.S. and van Oudenaarden, A. (2019) Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* **20**, 753–765 https://doi.org/10.1038/s41580-019-0186-3
96 Wagner, D.E. and Klein, A.M. (2020) Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 https://doi.org/10.1038/s41576-020-0223-2
97 Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. and Shapiro, E. (2005) Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, e50 https://doi.org/10.1371/journal.pcbi.0010050
98 Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H. et al. (2019) Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 https://doi.org/10.1016/j.cell.2019.01.022
99 Weiler, P., Van Den Berge, K., Street, K. and Tiberi, S. (2023) A guide to trajectory inference and RNA velocity. In *Single Cell Transcriptomics* (Calogero, R.A. and Benes, V., eds), pp. 269–292, Springer US, New York, NY. (Methods in Molecular Biology; vol. 2584)
100 La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V. et al. (2018) RNA velocity of single cells. *Nature* **560**, 494–498 https://doi.org/10.1038/s41586-018-0414-6
101 Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M. et al. (2022) Cellrank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 https://doi.org/10.1038/s41592-021-01346-6
102 Gehart, H., van Es, J.H., Hamer, K., Beumer, J., Kretzschmar, K., Dekkers, J.F. et al. (2019) Identification of enteroendocrine regulators by real-time single-cell differentiation mapping. *Cell* **176**, 1158–1173.e16 https://doi.org/10.1016/j.cell.2018.12.029
103 Bonniaud, P., Fabre, A., Frossard, N., Guignabert, C., Inman, M., Kuebler, W.M. et al. (2018) Optimising experimental research in respiratory diseases: an ERS statement. *Eur. Respir. J.* **51**, 1702133 https://doi.org/10.1183/13993003.02133-2017
104 Desai, T.J., Brownfield, D.G. and Krasnow, M.A. (2014) Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507**, 190–194 https://doi.org/10.1038/nature12930
105 Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H. et al. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 https://doi.org/10.1038/nature13173
106 Song, H., Yao, E., Lin, C., Gacayan, R., Chen, M.H. and Chuang, P.T. (2012) Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. *Proc. Natl Acad. Sci. U.S.A.* **109**, 17531–17536 https://doi.org/10.1073/pnas.1207238109
107 Ouadah, Y., Rojas, E.R., Riordan, D.P., Capostagno, S., Kuo, C.S. and Krasnow, M.A. (2019) Rare pulmonary neuroendocrine cells are stem cells regulated by Rb, p53, and Notch. *Cell* **179**, 403–416.e23 https://doi.org/10.1016/j.cell.2019.09.010
108 Byrnes, L.E., Deleon, R., Reiter, J.F. and Choksi, S.P. (2022) Opposing transcription factors MYCL and HEY1 mediate the Notch-dependent airway stem cell fate decision. *bioRxiv* https://doi.org/10.1101/2022.10.05.511009
109 Hogan, B.L.M., Barkauskas, C.E., Chapman, H.A., Epstein, J.A., Jain, R., Hsia, C.C.W. et al. (2014) Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 https://doi.org/10.1016/j.stem.2014.07.012
110 Wang, L., Wang, S. and Li, W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 https://doi.org/10.1093/bioinformatics/bts356
111 Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y. and Ge, Q. (2021) Bias in RNA-seq library preparation: current challenges and solutions. *BioMed Res. Int.* **2021**, 6647597 https://doi.org/10.1155/2021/6647597
112 van den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S. et al. (2017) Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 https://doi.org/10.1038/nmeth.4437
113 O'Flanagan, C.H., Campbell, K.R., Zhang, A.W., Kabeer, F., Lim, J.L.P., Biele, J. et al. (2019) Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 1 https://doi.org/10.1186/s13059-018-1612-0
114 Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 https://doi.org/10.1038/nrg3833
115 Loi, D.S.C., Yu, L. and Wu, A.R. (2021) Effective ribosomal RNA depletion for single-cell total RNA-seq by scDASH. *PeerJ* **9**, e10717 https://doi.org/10.7717/peerj.10717
116 Wagner, A., Regev, A. and Yosef, N. (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 https://doi.org/10.1038/nbt.3711
117 Lee, J., Hyeon, D.Y. and Hwang, D. (2020) Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 https://doi.org/10.1038/s12276-020-0420-2
118 Ogbeide, S., Giannese, F., Mincarelli, L. and Macaulay, I.C. (2022) Into the multiverse: advances in single-cell multiomic profiling. *Trends Genet.* **38**, 831–843 https://doi.org/10.1016/j.tig.2022.03.015
119 Waylen, L.N., Nim, H.T., Martelotto, L.G. and Ramialison, M. (2020) From whole-mount to single-cell spatial assessment of gene expression in 3D. *Comm. Biol.* **3**, 1–11 https://doi.org/10.1038/s42003-020-01341-1