SINGLE-CELL OMICS

# Computational and analytical challenges in single-cell transcriptomics

*Oliver Stegle[1], Sarah A. Teichmann[1,2] and John C. Marioni[1,2]*

Abstract | The development of high-throughput RNA sequencing (RNA-seq) at the single-cell level has already led to profound new discoveries in biology, ranging from the identification of novel cell types to the study of global patterns of stochastic gene expression. Alongside the technological breakthroughs that have facilitated the large-scale generation of single-cell transcriptomic data, it is important to consider the specific computational and analytical challenges that still have to be overcome. Although some tools for analysing RNA-seq data from bulk cell populations can be readily applied to single-cell RNA-seq data, many new computational strategies are required to fully exploit this data type and to enable a comprehensive yet detailed study of gene expression at the single-cell level.

Cell identity and function can be characterized at the molecular level by unique transcriptomic signatures[1]. At the organismal level, different tissues have distinct gene expression profiles[2,3], and even cells in consecutive stages of embryonic development have highly divergent transcriptomic landscapes[4]. Consequently, mutations that alter these expression profiles have been associated with adverse phenotypes ranging from a delayed immune response[5] to disease[6].

Until recently, molecular 'fingerprints' were generated using profiling of gene expression levels from bulk populations of millions of input cells[7]. These ensemble-based approaches, whether performed using microarrays[8] or the next-generation sequencing (NGS) approach of high-throughput RNA sequencing (RNA-seq)[9–11], meant that the resulting expression value for each gene was an average of its expression levels across a large population of input cells. In many contexts, such bulk expression profiles are sufficient. For example, in comparative transcriptomics, the goal is to study the selection pressures that apply to gene expression levels between samples of the same tissue taken from different species. In this context, a global view of average gene expression levels in each tissue, which can be obtained from bulk RNA-seq, may be sufficient[2,12]. Similarly, gene expression signatures obtained using ensemble approaches have yielded biomarkers that are predictive for disease status and clinical progression[13].

However, there are also important biological questions for which bulk measures of gene expression are insufficient[14]. For instance, during early development, there are only a small number of cells, each of which can have a distinct function and role[15–17]. Moreover, complex tissues, such as brain tissues, are composed of many distinct cell types that are typically difficult to dissect experimentally[18]. Consequently, bulk-based approaches may not provide insight into whether differences in expression between samples are driven by changes in cellular composition (that is, the abundance of different cell types) or by changes in the underlying phenotype. Finally, ensemble measures do not provide insights into the stochastic nature of gene expression[19,20]. In these and many other settings, assaying gene expression at the single-cell level represents a powerful, high-resolution tool for biological discovery.

Historically, measurements of the expression of a gene at the single-cell level were generated using low-throughput approaches. Examples of such methods are reporter constructs[21] or immunohistochemistry coupled with microscopy[22] at the protein level, and single-cell quantitative PCR (qPCR)[23] or single-molecule RNA fluorescence *in situ* hybridization (RNA FISH)[24] at the RNA level. Although experiments using these approaches provided important insights into transcriptional and translational kinetics[19] and the differentiation potential of individual cells[25], they were typically

[1]*European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.*
[2]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.*
*Correspondence to J.C.M.*
*e-mail: marioni@ebi.ac.uk*
doi:10.1038/nrg3833
Published online 28 January 2015

limited to the study of a single gene or a small number of genes. Consequently, a global picture of the active transcriptional landscape within a cell was simply not experimentally accessible.

Driven by these limitations, recent experimental advances have greatly improved the high-throughput generation of cDNA libraries from the polyadenylated fraction of mRNA molecules within a single cell[26–31]. Briefly, these protocols typically begin by isolating individual cells (manually by fluorescence-activated cell sorting (FACS) or by exploiting a microfluidics-based system) before lysing the cell, capturing the polyadenylated fraction of mRNA molecules and obtaining cDNA by reverse transcription. The cDNA is then amplified, using PCR[26,28,29,32] or *in vitro* transcription[31], to obtain enough material for hybridization to a gene expression microarray[15] or, more recently, to be profiled using NGS[26]; in the latter case, the technique is referred to as single-cell RNA-seq (scRNA-seq). scRNA-seq can now be applied to assay the individual transcriptomes of large numbers of cells isolated via microfluidics[33–36] or other microwell plate-based techniques[30]. The combination of a large number of cells and high-throughput profiling of gene expression at the single-cell level is crucial for answering many biologically relevant questions and represents an unprecedented opportunity for new discoveries in important areas of biology.

However, to ensure that scRNA-seq data are fully exploited and interpreted correctly, it is crucial to apply appropriate computational and statistical methods. Given the widespread use of bulk RNA-seq[37], many powerful tools for processing high-throughput transcriptomic data already exist[38]. However, scRNA-seq data analysis poses several unique computational challenges that necessitate the adaptation of existing workflows, as well as the development and application of entirely new analytical strategies.

The goal of this Review is to illustrate when bulk RNA-seq analysis strategies can be safely applied to scRNA-seq data, and to point out new approaches that are already available or remain to be developed. We focus first on challenges that are common to almost all scRNA-seq experiments, before turning to specific analytical methods that are required to capitalize on three key opportunities provided by scRNA-seq: the identification and characterization of cell types and the study of their organization in space and/or time; inference of gene regulatory networks and their robustness across individual cells; and characterization of the stochastic component of transcription.

## Incorporating quantitative standards

Before addressing these computational challenges, we first outline a fundamental aspect of scRNA-seq experimental design: namely, the incorporation of standards that (in conjunction with appropriate analysis strategies discussed below) facilitate quantitative comparisons of the expression level of each gene between cells. One approach, which is strongly recommended for all scRNA-seq experiments, is to use extrinsic spike-in molecules. Specifically, a whole-transcriptome spike taken from a different species from the cells of interest or a specially designed set of artificial spike-in molecules is added to the lysate extracted from each cell. The most widely used artificial spike-in mix is the External RNA Control Consortium (ERCC) set of 92 synthetic spikes based on bacterial sequences[39]. As an approximately constant volume of the spike-in mix is added to each cell extract (in some instances, variation in cell size may slightly vary the spike-in volume), the fact that the number of molecules of each spike-in RNA species should be the same across all single-cell libraries can be exploited to normalize gene expression levels and to estimate technical sources of variation (see below).

Additionally, although some protocols fragment and then sequence the amplified cDNA fragments, it is also possible to sequence reads derived solely from the 3′ or 5′ end of the amplified transcript. In this case, unique molecular identifiers (UMIs) have been used to barcode individual molecules. In conjunction with spike-ins (which are themselves barcoded before amplification), this protocol yields an estimate of the number of transcribed molecules that is independent of amplification biases[40], which are a major source of technical variability.

## Transcript quantification and quality control

The analysis of scRNA-seq data requires the careful execution of different computational steps, including read alignment and the basic generation of gene expression counts, quality control steps, normalization and downstream modelling. For several of these tasks, pipelines and tools that have been developed for RNA-seq data sets generated from bulk cell populations can be reused. However, there are important single-cell-specific aspects and pitfalls, which need to be considered (FIG. 1).

***Read alignment and generation of counts.*** Read alignment and the quantification of expression values is the first step in the analysis of RNA-seq data sets. In general, most of the methodology developed for bulk RNA-seq, including insights for how to best map the raw sequencing reads, can be reused for scRNA-seq[41–43]. Similar to bulk data sets, it is important to consider biases such as incomplete knowledge of the target genome or transcriptome annotation[44,45]. If synthetic spike-ins are used, then the reference sequence should be augmented with the DNA sequence of the spike-in molecules prior to mapping. When a UMI protocol is used, the barcode attached to each read should be removed for this read alignment step. Moreover, when using spike-ins in conjunction with UMIs, care must be taken to ensure that the sequences at the ends of the synthetic transcripts are complete. Otherwise, it may seem that some spike-ins have lower levels of expression than expected.

Subsequently, the mapped reads can be summarized to generate expression levels using the same approaches that are applied in conventional RNA-seq experiments (for example, by applying tools such as HTSeq[45], a

Spike-in
A few types of RNA with known sequence and quantity (generated either artificially or from a pool of RNA from a distantly related species) that are added as internal controls in RNA sequencing experiments.

Unique molecular identifiers
(UMIs). Tens of thousands of short DNA sequences (6–10 nucleotides in length), which are incorporated in molecules of interest before amplification, thus allowing biases to be accounted for.

Technical variability
Variability in gene expression levels between cells that arises through technical effects.

Read alignment
The alignment of short reads generated from a next-generation sequencing experiment to a reference genome or transcriptome.

Gene expression counts
The number of sequencing reads or unique molecular identifiers that map to a particular gene. These raw data form the basis of gene expression level quantification approaches.

computational processing package). When UMIs are used, these expression counts can be collapsed by summing the number of unique barcodes associated with all reads mapped to a given gene. When performing this analysis, care must be taken to account for sequencing errors in the UMIs that might result in the appearance of 'ghost' molecules. To account for this, error correction of the barcodes and/or removal of singleton barcodes may be required[40]. Finally, although scRNA-seq data can, in principle, be used to quantify the expression of individual exons or to resolve isoform abundances, such analyses are currently challenging owing to the large proportion of technical variability and biases compared to conventional RNA-seq protocols.

***Quality control: how to check the quality of each library and discard poor-quality cells.*** Quality control needs to be applied initially to the raw sequencing reads and, perhaps more importantly in the context of scRNA-seq, after alignment when the initial read counts are obtained in order to identify poor-quality libraries of individual cells.

At the level of the raw sequence files, bulk RNA-seq quality control tools, such as fastqc or kraken[46], can be applied (TABLE 1). Additionally, the data can be visualized using tools such as the Integrative Genomics Browser[47,48]. These steps will help to identify potential sample mix-ups and external contamination, or whether there was a problem with the sequencing itself as opposed to

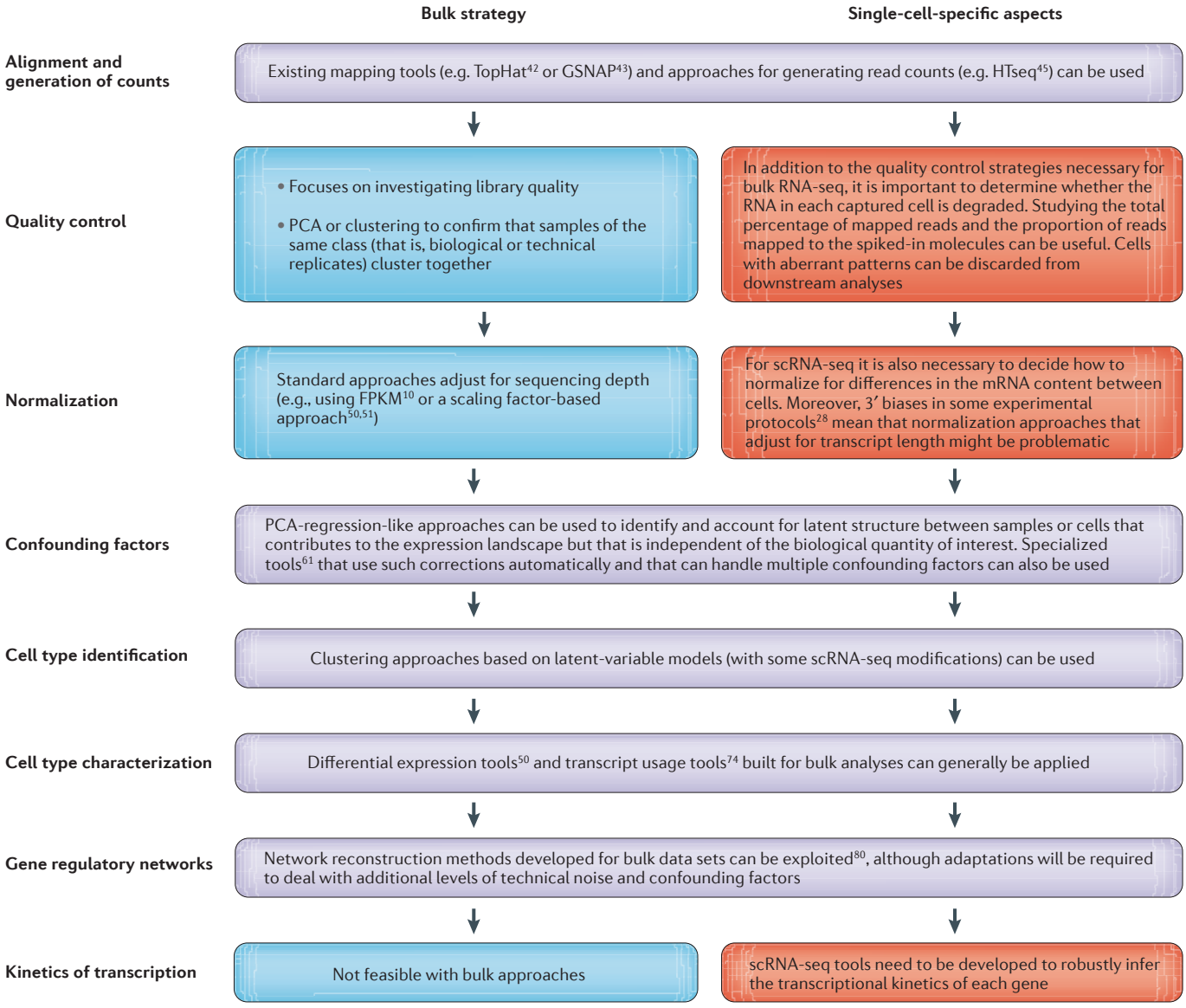



Figure 1 | **Comparison of bulk and scRNA-seq analytical strategies.** A flow chart of the steps in analysis of high-throughput RNA sequencing (RNA-seq) data from bulk cell populations and from single cells is shown. Methods that are common to both approaches are shown in purple, whereas key differences in analysis methods between bulk-based RNA-seq and single-cell RNA-seq (scRNA-seq) are shown in blue and red, respectively. FPKM, fragments per kilobase of exon per million fragments mapped; PCA, principal component analysis.

Table 1 | **Tools for scRNA-seq analysis**

| Name | For bulk cell populations or single cells? | Function | Ref |
|---|---|---|---|
| Fastqc | Bulk population | Mapping quality control | – |
| Kraken | Bulk population | Mapping quality control | 46 |
| GSNAP | Bulk population | Alignment | 43 |
| TopHat | Bulk population | Alignment | 42 |
| HTSeq | Bulk population | Obtaining expression counts | 45 |
| Single-cell normalization | Single cells | Normalization | 33 |
| Monocle | Single cells | Mapping transcripts on differentiation cascade | 66 |
| DESeq | Bulk population | Testing for differential expression | 50 |
| scLVM | Single cells | Accounting for confounding variation in scRNA-seq | 61 |
| Single-cell differential expression | Single cells | Testing for differential expression | 55 |
| Kinetics of transcription | Single cells | Identifying kinetic parameters | 81 |

scRNA-seq, single-cell RNA sequencing. In this table, some common tools for the analysis of scRNA-seq data are described. We note that this list is not exhaustive, especially in relation to the suggested tools for analyses of bulk RNA-seq data sets, but instead is meant to give some examples of tools that can be used at all stages of scRNA-seq analysis. See Further Information.

the single-cell capture and amplification (for example, by examining the proportions of duplicated reads or sequences mapped to bacterial genomes).

After establishing that there are no problems with the quality of the raw sequencing reads, the next step is to ascertain how well RNA was captured and amplified from each cell. This is an extremely important part of the analysis of scRNA-seq data, as many of the cells captured may contain degraded RNA (for example, because the cell is stressed[33]) and should therefore be discarded before downstream analysis. This is a more serious problem for primary tissue samples, as the process of extracting a tissue and then isolating individual cells can affect the quality of the RNA obtained.

A first metric that can indicate whether there is a problem with the sequencing library generated from an individual cell is the fraction of reads that map back to the genome of the organism of interest; this can be obtained directly from the fastqc output. If this value is low, it might indicate that RNA has degraded (possibly because the cell has entered apoptosis), that there is external contamination, or that the cell was inefficiently lysed.

A second metric, which can be computed from the fastqc output or directly from the table of counts generated by HTSeq, is the ratio of the number of reads mapped to the endogenous RNA (that is, the genome of the organism of interest) to the number of reads mapped to the extrinsic spike-ins: a high proportion of reads mapped to the spike-ins would be indicative of a low quantity of RNA in the cell of interest and might be a reason to discard cells (FIG. 2a). However, this ratio can vary from cell to cell for biologically relevant reasons: if the cells under study vary substantially in the amount of RNA contained (for example, if they are captured at different stages of the cell cycle), then this ratio would be expected to vary noticeably (see below). Nevertheless, cells for which the ratio of spike-ins is extremely discordant from the remaining population are strong candidates for exclusion.

Finally, a third useful approach for identifying problematic cells is to apply principal component analysis (PCA) to the gene expression matrix. The expectation when applying PCA is that good-quality cells cluster together and poor-quality cells are outliers. However, in some instances, poor-quality cells may also form a second distinct cluster. For example, it has been observed that poor-quality cells are often enriched in the expression of mitochondrial genes[40] (perhaps because the cells are undergoing apoptosis), which can cause them to cluster separately. This emphasizes that outlier analyses must be performed carefully to ensure that cells with physiologically relevant differences are not inadvertently discarded. To prevent this, one useful observation is that poor-quality cells typically display extreme values of the two other metrics described above.

***Normalization: from counts to expression levels.*** One important computational challenge in scRNA-seq is to appropriately normalize the data. In bulk RNA-seq data, the counts between different libraries are standardized by calculating quantities such as the fragments per kilobase of exon per million fragments mapped[10] (FPKM, which is obtained by standardizing transcript length and library size) and transcripts per million[49], or by using size factors to make counts comparable between libraries obtained from different samples[50,51]. However, approaches for normalizing bulk RNA-seq data make an implicit assumption that the total amount of RNA processed in each sample is approximately the same or that the variation is technical. This motivates the use of normalization strategies that generate relative expression estimates. In some contexts, this assumption has been shown to be misleading (for example, upregulation of *MYC* leads to a two-fold increase in the number of transcripts[52,53]) but, in general, such approaches are still widely used.

In scRNA-seq, the normalization procedure can substantially affect the interpretation of the data. Below, we consider separately normalization strategies for data generated with or without UMIs.

***Normalization of scRNA-seq data without UMIs.*** First, we briefly consider how data can be normalized in the absence of both UMIs and extrinsic spike-ins. A starting point is to apply a bulk-based normalization strategy that standardizes the amount of RNA contained in each cell — this assumes that the total amount of RNA in each cell is the same. However, without external spike-in controls, it is difficult to determine how much RNA is present in a cell.

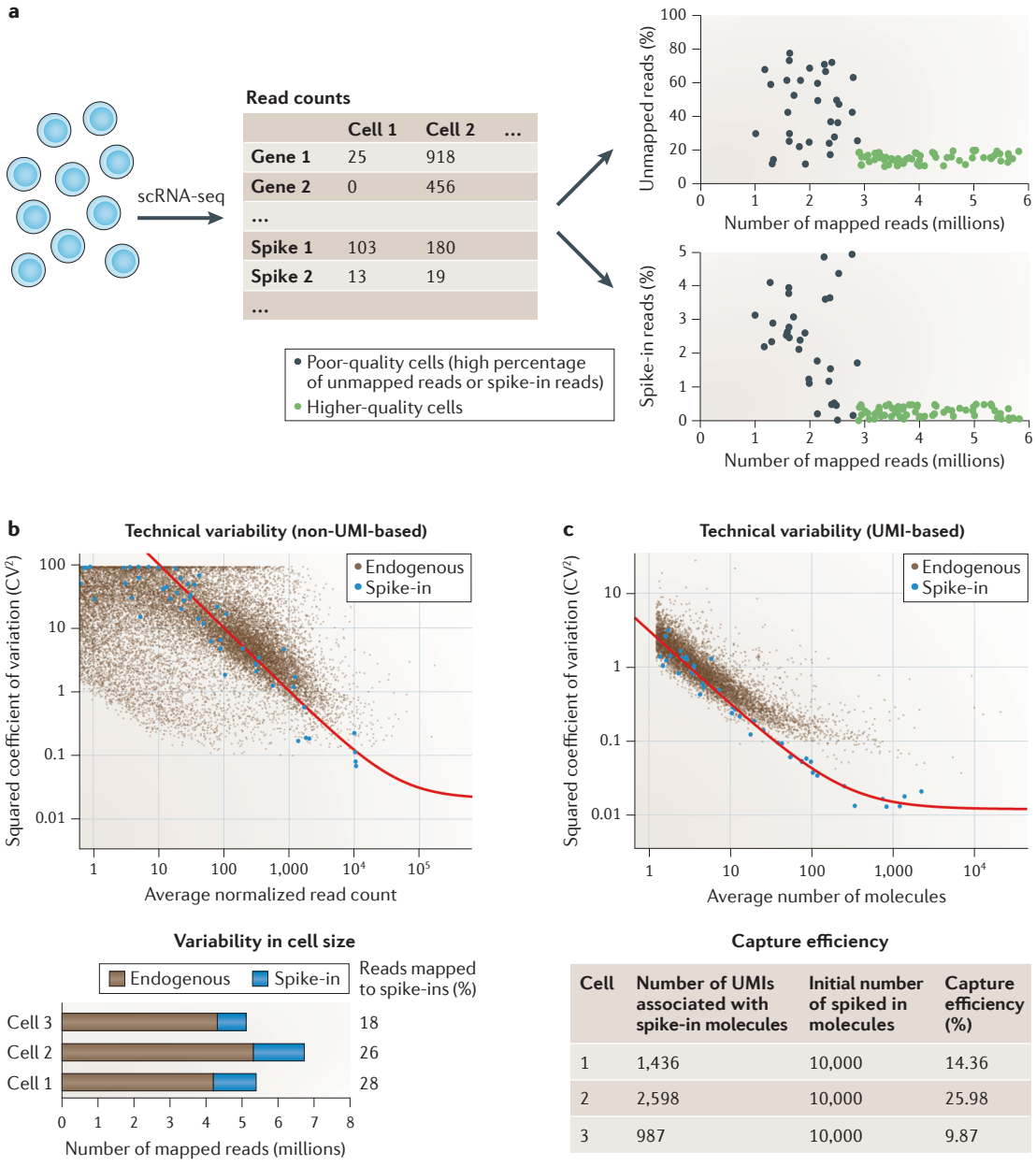
When extrinsic spike-ins are used, it is possible to accurately estimate relative differences in the total RNA content between cells. In particular, as the amount of

Duplicated reads
Identical copies of a sequencing read generated by the PCR amplification process.

Principal component analysis
(PCA). A statistical method to simplify a complex data set by transforming a series of correlated variables into a smaller number of uncorrelated variables called principal components.

Fragments per kilobase of exon per million fragments mapped
(FPKM). A method for quantifying gene expression levels from RNA sequencing data that normalizes for sequencing depth and transcript length.

Size factors
Quantities used to normalize gene expression levels between independently generated RNA sequencing libraries; they account for differences in sequencing depth.




| Cell | Number of UMIs associated with spike-in molecules | Initial number of spiked in molecules | Capture efficiency (%) |
|---|---|---|---|
| 1 | 1,436 | 10,000 | 14.36 |
| 2 | 2,598 | 10,000 | 25.98 |
| 3 | 987 | 10,000 | 9.87 |

Figure 2 | **Quality control and normalization.** **a** | Basic quality control steps are shown. After generating single-cell RNA sequencing (scRNA-seq) data, a key first step is to assess the quality of the data. In addition to quality metrics developed for bulk RNA-seq, it is important to determine whether cells have been captured efficiently and the mRNA fraction amplified faithfully. Two simple but important criteria are to compare the percentage of unmapped reads and the percentage of reads mapped to the external spike-in molecules across cells. Cells in which either of these values is high (grey) are of poor quality and should be discarded, leaving only the higher-quality cells (green) for downstream analyses. **b** | Spike-ins can be used to model technical variability and examine relative variability in cell size for non-unique molecular identifier (UMI)-based scRNA-seq data. If external spike-in molecules are added at the same volume to the RNA mixture from each cell before processing, they can be used to quantify the degree of technical variability across cells and to examine the relationship between technical variation and gene expression (upper panel). The *x* axis shows average expression levels across cells, and the *y* axis shows the squared coefficient of variation; blue points are extrinsic spike-in molecules. The red line indicates the fitted relationship between technical noise and gene expression strength. Additionally, by calculating the ratio between the numbers of reads mapped to the spike-in sequences and to the genes from the organism of interest, the relative amount of mRNA contained in each cell can be estimated (lower panel). **c** | Spike-ins can also be used to model technical variability and to examine relative variability in cell size for UMI-based scRNA-seq data. Similar to part **b**, the upper panel illustrates the relationship between technical noise and expression strength — the difference is that the expression level of each gene is now quantified as the number of unique cDNA molecules. Additionally, spike-ins can be used to quantify the capture efficiency and thus infer the number of mRNA molecules contained in the lysate of each cell (lower panel). Upper panels of parts **b** and **c** adapted from REF. 33 and REF. 40, respectively, Nature Publishing Group.

spiked-in material is assumed to be constant across cells, it is easy to calculate the ratio of the number of reads mapped to the genome of interest to the number of reads mapped to the spike-ins[33]. When compared between cells, this ratio allows differences in the amount of RNA within a cell to be inferred (FIG. 2b).

Given such data, the counts associated with each gene can be converted into absolute numbers of mRNA molecules based on the levels of the spike-ins, which have been added at a known concentration. A caveat is that the most common set of spike-ins (the 92-spike ERCC set) are 500–2,000 nucleotides in length, which is shorter than an average human mRNA (~2,100 nucleotides including untranslated regions[54]). Given the 5′-to-3′ length bias that is inherent to many scRNA-seq protocols, where the 3′ bias is more pronounced for longer transcripts[28], a conversion based on the shorter ERCC spike-ins is potentially problematic. Additionally, the ERCC spike-ins have comparatively short poly(A) tails and lack 5′ caps, which may result in different efficiency of their reverse transcription relative to the endogenous RNA molecules.

Consequently, it is challenging to devise a universally applicable normalization strategy for scRNA-seq data that properly accounts for variability in sequencing depth and in cell size. In many cases, a sensible and pragmatic approach is to calculate two alternative size factors: one for the spike-ins and one for the endogenous mRNA molecules[33]. The size factor for the spike-ins accounts solely for sequencing depth (as the spike-ins are present at the same quantity in all cells), whereas the size factor for the endogenous mRNAs normalizes for relative differences in cell size.

This twofold normalization means that the normalized spike-ins (which are adjusted for library size) can be used to estimate the degree of technical variability across the whole dynamic range of expression (see below). As mentioned above, the ratio of these two size factors can be used to estimate the total mRNA content of each cell, which is an informative additional molecular readout that can be used in downstream analyses (FIG. 2b).

As discussed above, normalizing for transcript length is challenging with current scRNA-seq protocols. In particular, although improvements have been made recently[32], there is still a noticeable 3′ bias to several scRNA-seq protocols, including the SMART-seq protocol used by the popular Fluidigm technology. As a result, normalizing for transcript length (for example, by applying an FPKM-type approach) is potentially problematic because it may underestimate the expression of long transcripts and overestimate the expression of short transcripts. In summary, until protocols allow an unbiased sampling of reads from across the whole transcript length, using FPKMs to compare the expression of transcripts with different lengths must be done with caution.

***Normalization of scRNA-seq data with UMIs.*** When UMIs are used, and assuming that the sample is sequenced to saturation (that is, the user has sequenced the library to a sufficient depth to ensure that each cDNA molecule is observed at least once), the number of UMIs linked to each gene is a direct measure of the number of cDNA molecules associated with that gene. Therefore, it is tempting to use these raw molecular counts — which, unlike expression estimates from non-UMI protocols, are independent of amplification biases — in all downstream analyses. However, despite this advantage of UMI-based protocols, technical sources of expression variability between cells cannot be fully excluded. Differences in the efficiency of the reverse transcription reaction between cells, as well as other cell-specific technical effects independent of amplification, mean that differences in the number of UMIs associated with each gene can vary between cells for technical rather than biological reasons.

One strongly recommended approach that can help to overcome this problem, as for non-UMI-based protocols, is to add extrinsic spike-in molecules to the cell extract before reverse transcription and amplification (FIG. 2c). As the number of mRNA spike-in molecules is theoretically the same across cells, systematic variability in the number of UMIs associated with spike-in genes across cells is indicative of differences in reaction efficiency (that is, technical variability). Consequently, a normalization step can be applied to convert the number of cDNA molecules associated with a gene to the number of mRNA molecules.

Furthermore, if the goal is to compare relative differences in expression, rather than absolute differences (for example, differences can arise in the total number of molecules depending on the stage of the cell cycle at which a cell is captured, or as a result of random fluctuations in the total RNA content within a cell), an additional normalization step can be applied, similar to that proposed for non-UMI-based scRNA-seq data. Finally, as UMI-based scRNA-seq protocols currently sequence only a fragment of each molecule (from either the 5′ or the 3′ end of the transcript), correcting for transcript length during normalization is unnecessary. One consequence of this tag-based protocol is that UMIs cannot be used to study isoform usage or allele-specific expression. However, UMI-based approaches are useful for obtaining accurate quantification of the expressed set of molecules within a cell, which can then be used for downstream analyses, such as cell type identification and characterization.

***Estimating technical variability.*** Once normalized gene expression levels or molecular counts have been generated, it is important to incorporate technical variability estimates. This applies to any downstream analysis but is particularly important when comparing expression levels between cells or when assessing the variability of individual genes (see below). Because of the typically low capture efficiency of current scRNA-seq protocols, even moderately expressed genes are frequently undetected. Consequently, methods to accurately estimate the extent of this technical variability are crucial in order to differentiate between genuine gene expression changes and experimental artefacts[33,55] (FIG. 2b,c).

Allele-specific expression
Gene expression levels measured separately for each of the two parental alleles. RNA derived from each allele can be quantified and assessed separately when RNA sequencing reads overlap with heterozygous sites in the genome.

Capture efficiency
The percentage of mRNA molecules in the cell lysate that are captured, amplified and sequenced. This is normally quantified using spike-in molecules.

Box 1 | **scRNA-seq experimental design considerations**

The most obvious experimental design questions related to single-cell RNA sequencing (scRNA-seq) experiments are the number of cells that need to be sequenced and the depth to which each individual cell should be sequenced. Both of these questions depend, inevitably, on the biological problem of interest, as well as on technical and financial constraints.

As a general rule, it is necessary to generate data from hundreds of cells to identify and characterize subpopulations of cells (especially rare populations) or to study the kinetics of transcription. Of course, if the goal is to answer specific questions relating to a small population of cells (for example, cells from the early developing embryo), such restrictions may not apply.

Recent studies using ~$10^2$–$10^3$ cells[30,36,62,65] have discovered previously uncharacterized transcriptional states, corresponding to new cell types or novel positions along differentiation cascades in various experimental systems. However, there are substantial differences in the depth to which the transcriptome of each individual cell was sequenced. Jaitin *et al.*[30] generated an average of 22,000 aligned sequence reads from 1,536 cells. By contrast, Mahata *et al.*[65] generated 12–20 million sequence reads from each of 93 cells. The key factor when determining the required sequencing depth is the number of reads that is necessary to sequence each reverse-transcribed and amplified cDNA molecule (or fragment thereof, if the protocol used does not involve unique molecular identifiers (UMIs)) at least once. When this depth is reached, further sequencing will not yield additional information. This value will depend on capture efficiency, with a higher capture efficiency meaning that deeper sequencing is necessary to fully capture the complexity of the captured transcriptome. However, this increased ability to fully characterize the transcriptome of a cell directly from deep scRNA-seq data is compensated for by the fact that fewer cells are typically able to be analysed by deep scRNA-seq than for shallow scRNA-seq; therefore, it may be that fewer cell types can be confidently identified when using deep scRNA-seq coupled with a small input population of cells.

Therefore, a balance between these two experimental parameters must be achieved. However, given the insights already obtained from relatively small numbers of cells, sequencing on the order of hundreds of cells at moderate depth would seem to suffice for many applications.

***Other confounding factors.*** Depending on the biological question of interest, other confounding factors may also have to be accounted for. Perhaps the most obvious confounding factor is batch effects. Unlike conventional RNA-seq experiments, batch variation is much more difficult to address at the experimental design stage (BOX 1). In a bulk RNA-seq experiment, the expression of genes in two or more conditions (for example, pre- and post-stimulation) is typically of interest. Assuming that there are multiple biological replicates per condition, the libraries for RNA-seq can be prepared in parallel and randomized among lanes and flow cells, both of which mitigate against batch effects[12]. By contrast, in scRNA-seq, cells from one condition are typically captured and prepared for sequencing independently from cells in a second condition.

Consequently, batch effects may be confounded with the biological covariate of interest (for example, condition) and are thus difficult to remove using regression analysis even with the presence of extrinsic spike-in molecules. One way to overcome this problem is to increase the number of biological replicates. As it is currently challenging to capture cells from multiple conditions (for example, pre- and post-stimulation) in parallel, an alternative is to independently repeat the experiment multiple times (that is, run multiple replicates of cells in the same condition), thus facilitating estimation of the technical variability; this will probably be necessary irrespective of the experimental technique used to capture the cells. In the context of microfluidics or microwell plate experiments, this corresponds to isolating and processing samples of single cells multiple times using independent microfluidic chips or microwell plates in order to properly model such effects.

In addition to technical factors such as batch effects, there may be other biological factors that lead to correlated changes in gene expression between cells, thereby obscuring the biological signal of interest. For example, in differentiating populations of cells, such cell-to-cell heterogeneity in gene expression can be caused by differences in the stage of the cell cycle at which a cell is captured (FIG. 3a). Although some of this variability may be accounted for by adjusting for cell size during the normalization, this will probably not capture all of the underlying variability. Moreover, there may be other latent (that is, hidden) variables that lead to heterogeneity in expression (including those related to core cellular processes or apoptosis). In the context of gene expression studies on bulk cell populations, approaches such as PCA[56], surrogate variable analyses[57], probabilistic estimation of expression residuals[58,59] or removal of unwanted variation[60] have been used to account for such latent factors. Many of these methods can, in principle, also be applied to scRNA-seq studies. As for bulk-based studies, it is important to check whether the latent variables identified correspond to the biological process of interest or whether they can be treated as confounders. Once these factors have been identified and their interpretation established, the data can be corrected to remove the effect of such confounding variables (FIG. 3b). For example, in a recent study[61], a latent-variable model based on Gaussian processes was used to account for confounding variation due to the cell cycle in scRNA-seq data sets. This method modelled and then used linear regression to remove variability in gene expression across cells that is attributable to cell cycle phase, thus allowing other biological components (for example, a differentiation process) to emerge more clearly from the data.

## Obtaining biological insights

We now return to the three specific biological questions for which scRNA-seq can provide insights that are not obtainable via bulk, ensemble-based RNA-seq. In all three cases we assume that the input is a matrix of gene expression counts that have been normalized and that have had confounding variables removed using the approaches described above.

***Identification of cell type and cellular state.*** A major and popular application of single-cell transcriptomics is to characterize a sample in terms of the known and novel cell types it contains[17,30,61–65]. Previous studies have shown that tissues can be clustered by their bulk expression profiles[3]. For example, when examining expression patterns in multiple tissue samples obtained from different primates, clustering analyses have been used to show that samples separate first by tissue and only then by species[2].

Confounding factors
Unobserved covariates that affect gene expression levels and that can obscure the interpretation if not accounted for.

Batch effects
Systematic differences in gene expression levels between independent cells from the same population, which arise as a result of sample preparation.

Biological replicates
Independent replicates from the same population.

scRNA-seq can be used to address hidden tissue heterogeneity: by clustering cells on the basis of their expression profiles, distinct subsets — potentially corresponding to unknown cell types — can be identified. The putative cell types can be characterized by studying the functions of the genes that best distinguish them. Additionally, these approaches can provide insights into differentiation: if a population of cells at different stages of differentiation into a specific cell type are processed in parallel, it is possible to map these cells onto a specific point in the differentiation cascade. One method would be to use unsupervised clustering-based approaches, which do not rely on known marker genes[62,66].

Methods for clustering cells can be split into two groups depending on whether there is established information or an expectation regarding the relationship between the cells. If there is no prior expectation, unbiased clustering methods — such as hierarchical clustering or PCA-like methods — can be applied to group cells according to their position along the differentiation cascade[66]. If prior information is available, PCA-like approaches can be combined with knowledge of the expression patterns of a small set of known marker genes, allowing an approximate spatial map of the tissue under study to be obtained[63]. Accounting for confounding variables can be useful in this context[61] (FIG. 4).

If the spatial location of a cell is known[67,68], in some tissues it can be reasonable to hypothesize that cells located closer to one another are more likely to belong to the same type than more distant cells. To this end, methods to cluster cells using both spatial and quantitative information via a Markov random field (MRF)-based approach show promise[69].

In addition to cell type identification, unsupervised methods such as PCA can also be used to explore cellular state, for example, stage or speed of the cell cycle. Perhaps counter-intuitively, slow-cycling cells tend to have clearer transcriptional signatures of G1/S versus G2/M stages, whereas fast-cycling cells tend to be more homogeneous with respect to expression of cell cycle genes. A recent study of single cells obtained from glioblastomas describes a computational strategy for quantifying the speed of the cell cycle in each cell by comparing expression levels of G1/S versus G2/M genes[35].

***Differential expression and transcript isoforms.*** Having partitioned cells into different clusters using computational approaches as outlined above or, alternatively, using cell-surface markers, one key objective is to define the sets of genes that best discriminate the different clusters.

Perhaps the most obvious way to address this problem is by identifying genes that are differentially expressed between pairs of clusters. For example, by using such a strategy, a recent study discovered that when dendritic cells are challenged by pathogenic stimuli, a set of antiviral genes is upregulated in only a small subset of cells immediately post-stimulation; this set of genes is upregulated in all cells at later time points[36]. From a computational perspective, approaches based on standard differential expression tools for bulk RNA-seq can be used[50,70,71]. Although scRNA-seq data measurements are typically noisier than those generated by bulk RNA-seq[28], this is compensated for, at least to some extent, by noting that the number of cells per study group in scRNA-seq is typically much greater than the number of samples per group in a bulk RNA-seq study. Recently, an alternative approach designed specifically for scRNA-seq data has been described[55], which explicitly accounts for technical variation due to allelic dropout.

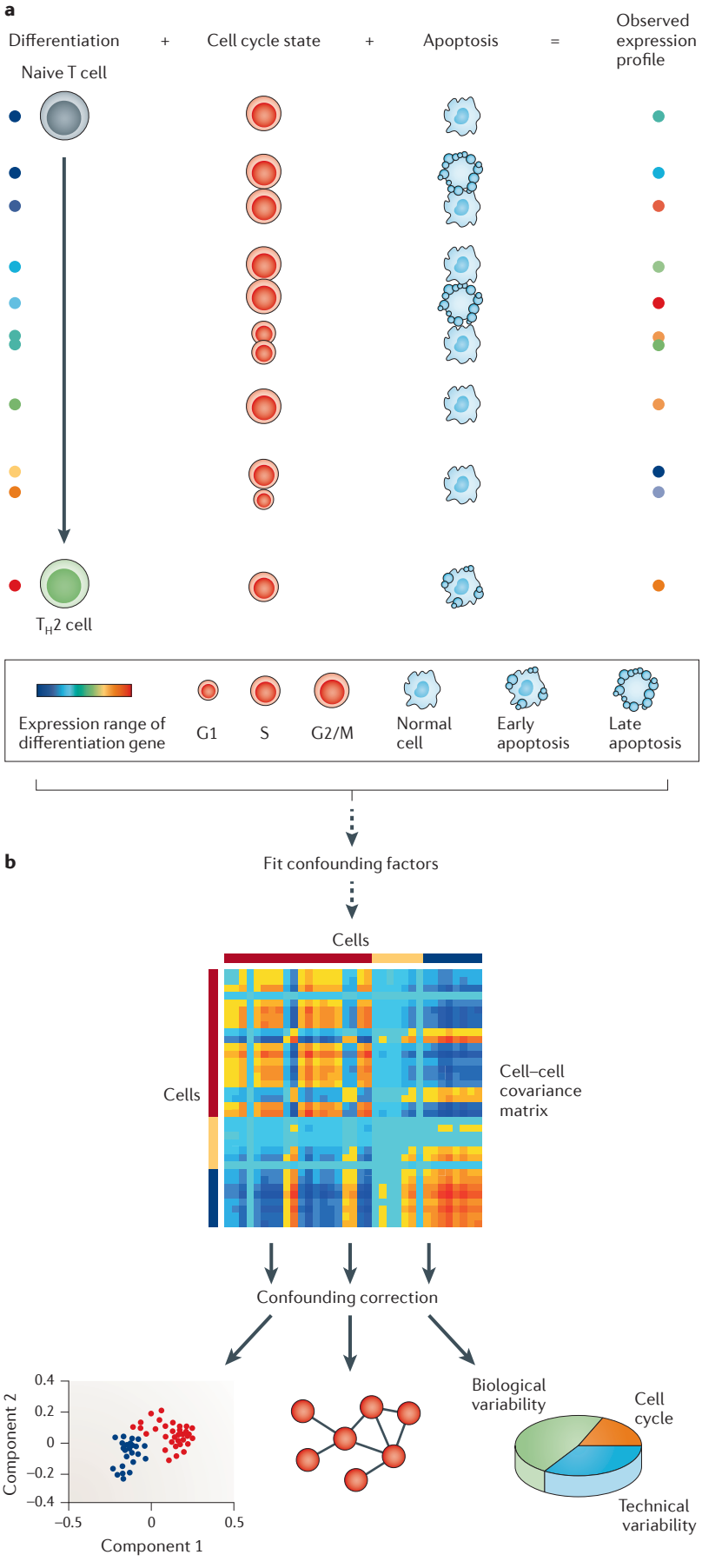
Another way of characterizing the putative clusters is to identify transcripts that display differences in isoform usage[72]. As in studies of differential expression, tools for identifying differentially expressed exons[73,74] can be applied in this case. One potential limitation is the 3′ bias in expression noted above, which will affect the power to identify differential isoform usage. Finally, differences in transcription start sites, which can be straightforwardly identified if a 5′ UMI protocol is used, can also be used to characterize different clusters of cells.

***Identifying highly variable genes.*** In parallel to differential expression analysis, an important challenge is to identify the genes with the most highly variable expression patterns across a population of cells without prior knowledge of the underlying cell types. Gene expression variability can provide clues regarding transcriptional heterogeneity in the sample of interest, giving insights into the robustness of gene expression regulation between cells, as well as cell type characterization. The identification of highly variable genes requires the application of statistical approaches that account for technical sources of variation, such that biological variability in gene expression levels can be

Figure 3 | **Confounding variables and how to account for them.** **a** | For each gene, the observed expression profile generated from single-cell RNA sequencing (scRNA-seq) is caused by a combination of factors. For example, if cells are being sampled randomly from a mixed population containing naive (that is, undifferentiated) cells and cells that are closer to being fully differentiated, then for each cell, the expression profile is a combination of a variety of factors (including position on the differentiation cascade, cell cycle state and apoptotic state). Factors such as the cell cycle or apoptotic state can be considered confounders that prevent the signal of interest (the differentiation state of a cell) from being uncovered. **b** | Confounding factors need to be identified and corrected for in downstream analyses. Latent-variable models, which are built on approaches applied in bulk RNA-seq studies to infer and correct for hidden factors that cause gene expression heterogeneity[56,57,59], can be used to deduce the correlation between cells due to factors such as the cell cycle or apoptotic state. Subsequently, the extent of variance in the expression of each gene across cells that is attributable to this factor (and other factors) can be inferred. Additionally, the scRNA-seq data can be corrected by using regression analyses to remove the confounding factor, thus facilitating downstream analyses such as clustering or network analyses. Figure from REF. 61, Nature Publishing Group.

Markov random field
(MRF). A particular class of statistical model that can exploit smoothness of measurements in a spatial grid, thereby improving the accuracy of parameter estimates.

Dropout
The false quantification of a gene as 'unexpressed' due to the corresponding transcript being 'missed' during the reverse-transcription step. This leads to a lack of detection during sequencing.




quantified. In addition, it is important to realize that high variability of gene expression can also be caused by a confounding factor that is not accounted for, such as the cell cycle (FIG. 3).

One approach is to compute the coefficient of variation (that is, the empirical variance divided by the squared mean) for each gene across the population of cells under study and to rank the genes accordingly. Unfortunately, technical variability, which is intrinsic to the experimental protocol and not associated with genuine biological variability, is greater for lowly expressed genes than for highly expressed genes[72]. Consequently, a null estimate of the expected technical variability needs to be computed. This can be done using the extrinsic spike-in molecules — the extent of variability in their expression across cells can be used as an estimate of the null variance. This information allows the expected technical variation to be modelled across the whole dynamic range of expression, which forms the basis of a statistical test to determine the set of genes that show more variability in expression than would be expected by chance[33]. Recently, extrinsic spike-in molecules have been used to further decompose technical variability into two terms that correspond to sampling noise and heterogeneity in sequencing efficiency across cells[75].

***Regulatory networks and their robustness.*** One important and widespread application of bulk RNA-seq studies has been the identification of co-regulated modules of genes and gene regulatory networks. Typically, RNA-seq is applied to measure mRNA expression levels across a compendium of samples corresponding to, for example, multiple individuals. From these data the sample gene–gene covariance matrix can be computed, which forms the basis of a large number of network reconstruction methods[76]. Pairs of genes with highly correlated expression levels across samples are then assumed to be co-regulated.

To improve the robustness of the inference of co-regulation, methods that group genes into regulatory modules[77] or that account for unmeasured regulators[78] have been considered. Networks inferred using such an approach are typically undirected; that is, it is unknown which gene is upstream in the regulatory cascade[79]. To ascertain directionality, measurements of gene expression levels in perturbed conditions (for example, upon knockdown of a given gene of interest) are typically necessary[80].

In the context of scRNA-seq, one can, in principle, replace samples with cells before performing an analysis very similar to that described for bulk data to identify co-regulated genes. Importantly, the inferred gene–gene correlation network will depend on the cell heterogeneity in the sample. For example, single cells derived from a heterogeneous sample composed of several cell types are unlikely to yield the same network as those derived from one of the constitutive cell types. Thus, combining gene regulatory network inference with cell type identification as an initial step may be important. Despite these pitfalls, there are already promising applications of co-expression analysis in scRNA-seq data[65].

Monoallelic expression
The expression of only one of the two parental alleles.

Assuming that the technical challenges can be overcome, it is important to realize that the regulatory networks derived from scRNA-seq data can provide insights that are not easily obtainable from bulk RNA-seq data networks. For example, if a set of genes can be activated independently by two transcription factors and only one of these two factors is expressed in any given cell, then when constructing a network from bulk RNA-seq data, these two factors might seem to be co-expressed, or it might seem as if one regulates the other. By contrast, from the scRNA-seq data, it would be clear that these genes are never co-expressed, thus providing important insights into mutually exclusive regulatory mechanisms that activate the same set of downstream genes. However, scRNA-seq data may also reveal less biologically meaningful correlations between genes. For example, if some cells are in G1 phase while others are in G2 phase of the cell cycle, this can cause widespread correlations between large sets of genes. Thus, depending on the aims of a particular gene network analysis, it may be important to remove confounding factors such as the cell cycle.

***Stochasticity of transcription.*** One important application of scRNA-seq is the study of the kinetics of gene expression: unlike the population-averaged data from RNA-seq on bulk cell samples, scRNA-seq can characterize diversity in transcription between individual cells[4,81].

Previous experimental approaches, such as time-lapse microscopy studies[24,82], have labelled individual genes and then tracked their expression over time, which has allowed estimation of the rate per unit time at which a gene transitions from the off state to the on state and vice versa, along with the rates per unit time of gene transcription and mRNA decay[24,28,82,83]. These four kinetic parameters can then be used to study transcriptional bursting, which defines the expression profile of the gene under study. Transcriptional bursting is characterized by two quantities: the burst size, which describes the average number of mRNA molecules synthesized when a gene is in the active state; and the burst frequency, which is the number of bursts per unit time[82]. However, studies of transcriptional bursting using approaches such as time-lapse microscopy can typically be applied to only a small number of genes at any one time.

By contrast, scRNA-seq facilitates expression profiling of large numbers of genes across many individual cells[81]. However, as current scRNA-seq protocols require cells to be lysed before library preparation, the expression of a gene cannot be measured over time. Instead, for each gene, the distribution of expression levels across cells can be regarded as a sample from a stationary distribution generated using time-indexed measurements of the expression of that gene in a single cell (FIG. 5a). Subsequently, it is possible, independently for each gene, to use computational approaches to estimate the kinetic parameters[81]. However, as these parameters are measured in units of time and we observe only the stationary distribution, the rate of decay is typically set to one, which makes the kinetic parameters independent of time[19,84]. This is a strong assumption, as the rate of transcript degradation is known not to be constant across genes[85]; therefore, the inferred parameters must be interpreted with this caveat in mind. One additional obstacle that needs to be overcome is the incorporation of technical variability into the procedure for estimating the kinetic parameters.

***Using allele-specific expression to study the regulation of gene expression.*** One interesting application of scRNA-seq is to study allele-specific expression, including random monoallelic expression. Like bulk RNA-seq studies[86], allele-specific expression can be measured and used to determine the extent of allelic bias in gene expression[4] (FIG. 5b). By exploring the degree of allele-specific expression, stochastic transcription of each allele and the degree of co-ordination of expression between alleles can be investigated. For example, scRNA-seq has been used to study stochastic allelic expression during early embryogenesis[4]. Specifically, using first-generation intercrosses between two different inbred strains of mice, the extent of stochastic allele-specific expression during early embryogenesis has been quantified transcriptome-wide.

One concern with using scRNA-seq to study random monoallelic expression is that allelic dropout during library preparation might lead to erroneous measurements of monoallelic expression. Previous approaches have addressed this by splitting cell lysates into two and then repeating the experiment to provide a background estimate of allelic dropout[4]. However, this is an area in which more work is required to develop computational methods that can accurately model this feature of scRNA-seq library preparation, such that accurate measures of allele-specific expression can be obtained.

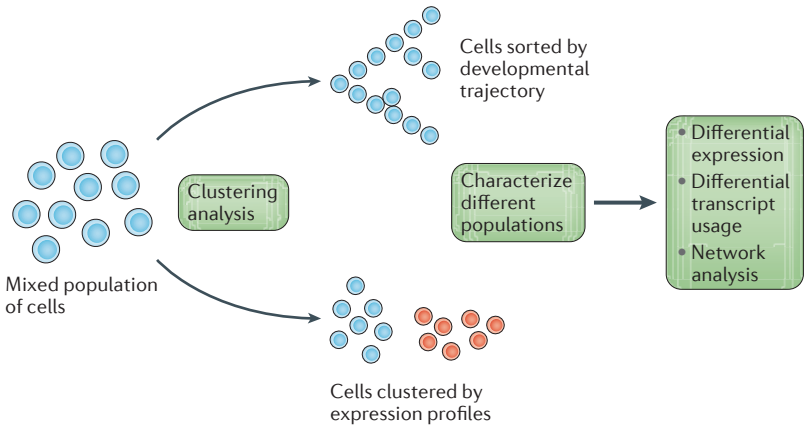



Figure 4 | **Finding new cell types and allocating cells along a differentiation cascade.** Unbiased clustering approaches based on principal component analysis (PCA)-like methods[62,63,66] can be used on a mixed population of cells, to either map them along a differentiation cascade or cluster them into new cell types[55,58]. Subsequently, the newly identified cascades or populations can be characterized, and new marker genes can be found by identifying genes or transcript isoforms that are differentially expressed between the populations.

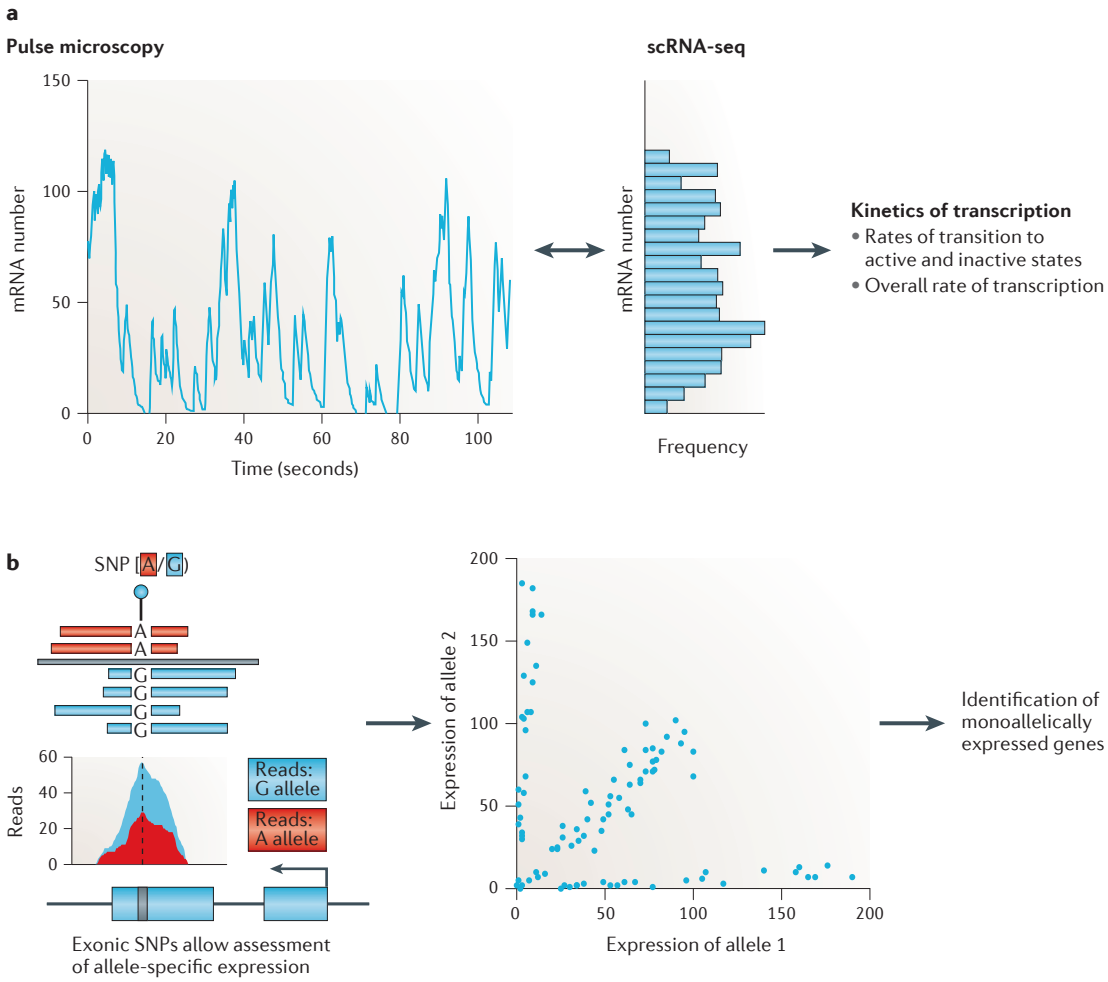



Figure 5 | **The kinetics of transcription.** **a** | Single-cell RNA sequencing (scRNA-seq) can be used to study the kinetics of transcription. RNA labelling followed by pulse microscopy (left panel) can be used to track the expression of a gene over time[82]. scRNA-seq can be used to obtain an instantaneous snapshot of this distribution by measuring the expression of an individual gene across many cells (middle panel)[81]. Subsequently, these data can be used to draw inferences about the kinetics of transcription. **b** | Allele-specific expression can be studied using scRNA-seq. Allele-specific expression can be assayed using single-nucleotide polymorphisms (SNPs) in the sequence of a transcript to allocate reads to alternative alleles. Subsequently, the number of cells in which both alleles are expressed and the numbers of cells in which allele 1 or allele 2 is exclusively expressed can be counted. This allows the identification of genes that display evidence of monoallelic expression[4]. One important challenge is to address technical issues, especially allelic dropout during sample preparation, which can bias the results.

## Conclusions and perspectives

Recent progress in the development of experimental methods for scRNA-seq has been rapid and exciting, with a plethora of unexpected and profound new insights emerging in a short period of time. These include the identification of new cell types, identification of gene expression patterns that are predictive of cellular state, and opportunities for studying the functional implications of stochastic transcription. These results are built on the solid foundations of computational methods that have been developed for sequencing of bulk cell populations, which have proved extremely powerful.

However, to ensure that scRNA-seq data can be properly analysed, it is crucial to develop computational methods that are tailored specifically for processing single-cell data and that keep pace with advances in experimental techniques. New methods have already been, and continue to be, developed for normalization and cell type identification, which focus more heavily on dissecting variability in expression levels across cells. Additionally, there are numerous areas in which new tools remain to be developed.

First, normalization of scRNA-seq data must properly account for differences in the total amount of RNA transcribed within a cell and, for non-UMI-based protocols, differences in sequencing depth. Second, methods for modelling confounding variables and/or using regression-based analysis to remove them will be required if the biologically relevant signal in scRNA-seq data sets is to be robustly uncovered. Third, accurately

modelling technical variability (even after removing confounding variables) is vital because without a basic understanding of the underlying noise that is inherent to scRNA-seq data, downstream interpretation can be seriously compromised. It is expected that such methods will be developed in the next few years, leading to new discoveries in areas ranging from the physiology of tissues to systems biology.


1. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
3. Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K. & Gilad, Y. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* **4**, e1000271 (2008).
4. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
5. Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proc. Natl Acad. Sci. USA* **109**, 1204–1209 (2012).
6. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
7. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14**, 618–630 (2013).
**This is a related review discussing challenges and analysis opportunities of single-cell sequencing, for example, to reconstruct lineages in cancer.**
8. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
9. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
10. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621–628 (2008).
11. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
12. Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* **22**, 602–610 (2012).
13. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
14. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* **11**, 22–24 (2014).
15. Ohnishi, Y. *et al.* Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nature Cell Biol.* **16**, 27–37 (2014).
16. Skamagki, M., Wicher, K. B., Jedrusik, A., Ganguly, S. & Zernicka-Goetz, M. Asymmetric localization of *Cdx2* mRNA during the first cell-fate decision in early mouse development. *Cell Rep.* **3**, 442–457 (2013).
17. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
18. Diez-Roux, G. *et al.* A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.* **9**, e1000582 (2011).
19. Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
20. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
21. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
22. Coons, A. H., Creech, H. J. & Jones, R. N. Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.* **47**, 200–202 (1941).
23. Taniguchi, K., Kajiyama, T. & Kambara, H. Quantitative analysis of gene expression in a single cell by qPCR. *Nature Methods* **6**, 503–506 (2009).
24. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5**, 877–879 (2008).
25. Faddah, D. A. *et al.* Single-cell analysis reveals that expression of nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. *Cell Stem Cell* **13**, 23–29 (2013).
26. Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).
27. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
28. Ramskold, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotech.* **30**, 777–782 (2012).
29. Sasagawa, Y. *et al.* Quartz-seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, R31 (2013).
30. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
31. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
32. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096–1098 (2013).
**Recent protocol developments, such as the development of Smart-seq2, have helped to substantially reduce biases and improved the sensitivity of scRNA-seq.**
33. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**, 1093–1095 (2013).
**This paper reports a statistical approach that estimates and accounts for technical sources of variation in scRNA-seq experiments. This method exploits spike-ins to separate technical and biological variability of individual genes (see also reference 75).**
34. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11**, 41–46 (2014).
35. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
**This paper provides an example in which sequencing the transcriptomes of a large number of single cells provided important insights into intra- and inter-tumour heterogeneity.**
36. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **509**, 363–369 (2014).
37. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
38. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol.* **11**, 220 (2010).
39. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
40. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**, 163–166 (2014).
**UMIs allow individual molecules to be barcoded. This protocol enables the absolute number of transcribed molecules to be estimated independently of amplification biases.**
41. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169–3177 (2012).
42. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
43. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
44. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotech.* **28**, 503–510 (2010).
45. Anders, S., Pyl, P. T. & Huber, W. HTseq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
46. Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49 (2013).
47. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotech.* **29**, 24–26 (2011).
48. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192 (2013).
49. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
50. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
**This seminal paper describes statistical methods to test for differential gene expression using RNA-seq data. Although developed in the context of RNA-seq studies on bulk cell populations, this work has laid the foundation for a large family of normalization procedures, including recent methods that are dedicated to scRNA-seq data (see reference 33).**
51. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
52. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated *c-Myc*. *Cell* **151**, 56–67 (2012).
53. Loven, J. *et al.* Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
54. Krebs, J. E., Goldstein, E. S. & Kilpatrick, S. T. *Lewin's Genes XI* (Jones & Bartlett Publishers, 2014).
55. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740–742 (2014).
**This paper presents a Bayesian approach to test for differential gene expression in scRNA-seq studies. This approach extends methods for bulk RNA-seq (for example, reference 50) by accounting for single-cell-specific noise, such as dropout events and amplification biases.**
56. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
57. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
58. Stegle, O., Parts, L., Durbin, R. & Winn, J. A. Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
59. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protoc.* **7**, 500–507 (2012).
60. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotech.* **32**, 896–902 (2014).
61. Buettner, F. *et al.* Accounting for cell-to-cell heterogeneity in single-cell RNA-seq data reveals novel structure between cells. *Nature Biotech.* http://dx.doi.org/10.1038/nbt.3102 (2015).
**Confounding factors such as the cell cycle can obscure biologically relevant molecular signatures in scRNA-seq data sets. This work describes a computational approach to account for confounding factors. Related methods developed for bulk RNA profiling experiments are described in references 57–60.**

62. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
63. Durruthy-Durruthy, R. *et al.* Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).
64. Moignard, V. *et al.* Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biol.* **15**, 363–372 (2013).
65. Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids *de novo* to contribute to immune homeostasis. *Cell Rep.* **7**, 1130–1142 (2014).
**This paper provides an example from T cell biology that shows how gene–gene correlations in scRNA-seq studies can be used to reveal novel mechanistic insights.**
66. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotech.* **32**, 381–386 (2014).
**This paper describes a computational approach to reconstruct a pseudotemporal order from multiple scRNA-seq snapshot experiments, for example, along a differentiation trajectory.**
67. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343**, 1360–1363 (2014).
68. Lovatt, D. *et al.* Transcriptome *in vivo* analysis (TIVA) of spatially defined single cells in live tissue. *Nature Methods* **11**, 190–196 (2014).
69. Pettit, J. B., Tomer, R., Achim, K., Azizi, L. & Marioni, J. C. Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput. Biol.* **10**, e1003824 (2014).
70. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
71. Hardcastle, T. J. & Kelly, K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
72. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
73. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
74. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009–1015 (2010).
75. Grun, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**, 637–640 (2014).
76. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
77. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.* **34**, 166–176 (2003).
78. Liao, J. C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA* **100**, 15522–15527 (2003).
79. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3**, 78 (2007).
80. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17** S215–S224 (2001).
81. Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, R7 (2013).
82. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
83. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Rev. Genet.* **6**, 451–464 (2005).
84. Larson, D. R. What do expression dynamics tell us about the mechanism of transcription? *Curr. Opin. Genet. Dev.* **21**, 591–599 (2011).
85. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
86. McManus, C. J. *et al.* Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* **20**, 816–825 (2010).



### Acknowledgements
The authors acknowledge members of the Marioni, Stegle and Teichmann groups for comments on the manuscript. They also acknowledge S. Linnarsson for advice on how to present computational challenges relating to scRNA-seq data generated using UMI-based protocols.


### Competing interests statement
The authors declare no competing interests.

### FURTHER INFORMATION
**DESeq:** http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
**Fastqc:** http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
**GSNAP:** http://research-pub.gene.com/gmap/
**HTSeq:** http://www.huber.embl.de/users/anders/HTSeq/doc/overview.html
**Kinetics of transcription:** http://genomebiology.com/content/supplementary/gb-2013-14-1-r7-s4.zip
**Kraken:** http://www.ebi.ac.uk/research/enright/software/kraken
**Monocle:** http://monocle-bio.sourceforge.net/
**scLVM:** http://github.com/PMBio/scLVM
**Single-cell normalization:** http://www.nature.com/nmeth/journal/v10/n11/extref/nmeth.2645-S2.pdf
**Single-cell differential expression:** http://www.nature.com/nmeth/journal/v11/n7/extref/nmeth.2967-S2.zip
**TopHat:** http://ccb.jhu.edu/software/tophat/index.shtml

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**