Trends in Immunology

CellPress REVIEWS

Opinion

# Immunology Driven by Large-Scale Single-Cell Sequencing

Tomás Gomes,[1] Sarah A. Teichmann,[1,2,3,*] and Carlos Talavera-López[1,2]

**The immune system encompasses a large degree of phenotypic diversity and plasticity in its cell types, and more is to be uncovered. We argue that large, multiomic datasets of single-cell resolution, in conjunction with improved computational methods, will be essential to resolving immune cell identity. Existing datasets, combined with 'big data' methodologies, can serve as a platform to support future studies in immunology. Technical and analytical advances in multiomics and spatial integration can provide a reference for gene regulation and cellular interactions in spatially structured tissue contexts. We posit that these developments may allow guided functional studies of immune cell populations and lay the groundwork for informed cell engineering and precision medicine.**

## Highlights

Cell-type references generated from collections of single-cell RNA sequencing data can accelerate the functional characterization of diseases.

Computational methods process and analyze sequencing data for a detailed characterization of cellular phenotypes.

Single-cell profiling of different molecular layers can give further functional context to cell-type identity.

The addition of spatial information can reveal immune cell function in tissue contexts.

## Unraveling the Immune System One Cell at a Time

The human immune system is one of the most complex; further understanding these complexities can have a significant impact on preventing and curing a variety of diseases. A large number of cell types and states, many of which remain to be further characterized, underlie the many types of immune responses. Many gene products have also been studied over the years; however, owing to the low number of available high-throughput approaches, many more are either unstudied or have undetermined functions.

We discuss here the most recent developments in single-cell technology and analysis, and what they can mean for immunology. Advances in **single-cell RNA sequencing** (scRNA-seq; Glossary) data analyses are poised to result in a complete census of human cell types [1,2]. This growth in datasets has been accompanied by the development of experimental methods that capture the 'states' of different molecules (DNA, RNA, and protein) in individual cells, revealing many of the regulatory underpinnings of cellular immunity, as well as virulence mechanisms of pathogens. Methods for whole-transcriptome spatial mapping are also emerging and reaching single-cell resolution, enabling for the first time the construction of an atlas of cellular interactions in complex tissues during health and disease. Supporting these advances have been developments in computational methods. In particular, the wide adoption of cutting-edge machine learning and artificial intelligence methods are set to improve our modeling and predictive power by deconvoluting gene expression networks and creating informative, integrated models of immune cells in disease. We therefore posit that single-cell approaches will, in the near future, be one of the tools most widely used for characterizing many aspects of the immune system.

## Multiple Windows into the Molecular Machinery of the Cell

The earliest single-cell methods relied on protein expression to determine cell types and discern mechanisms underlying biology and disease [3]. Recently, RNA has been used as a defining molecule for single-cell phenotyping; nonetheless, important information about cellular heterogeneity can still be found at the level of DNA and proteins (Figure 1A).

Methylation patterns govern gene expression [4], and single-cell methylation profiling has been used to distinguish rare hematopoietic stem cell subpopulations [5]. At the single-cell level, however, open chromatin regions are easier to profile, and are associated with regulatory and active elements in the genome, which can also be used to define cell types [6]. These are more efficiently profiled by the **assay for transposase-accessible chromatin** (ATAC-seq) protocol [7,8]. ATAC-seq can effectively separate immune cell populations based on transcription factor motifs detected in open chromatin peaks [9]. It has also been possible to obtain information on the genome

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

[3]Theory of Condensed Matter, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, UK

*Correspondence: st9@sanger.ac.uk

https://doi.org/10.1016/j.it.2019.09.004

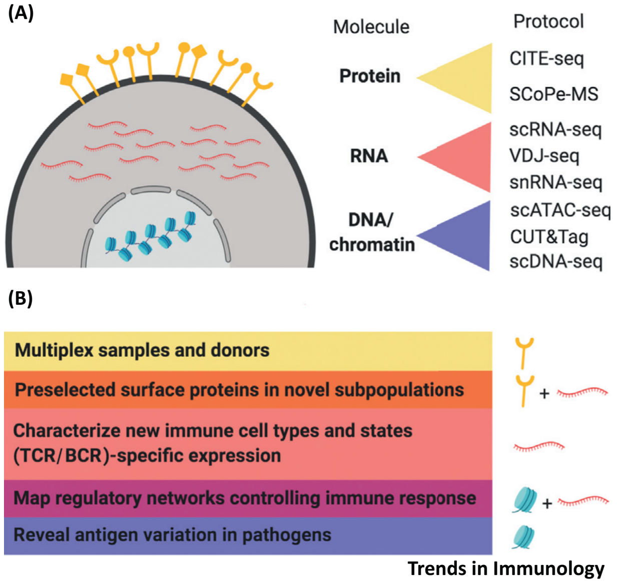
(A)
Molecule
Protocol
Protein
CITE-seq
SCoPe-MS
RNA
scRNA-seq
VDJ-seq
snRNA-seq
DNA/ chromatin
scATAC-seq
CUT&Tag
scDNA-seq
(B)
Multiplex samples and donors
Preselected surface proteins in novel subpopulations
Characterize new immune cell types and states (TCR/BCR)-specific expression
Map regulatory networks controlling immune response
Reveal antigen variation in pathogens
+
+
Trends in Immunology

**Figure 1. An Integrative Approach for Single-Cell Multiomic Data.**

(A) Multiple measurements, such as proteome, genome, transcriptome, methylome, and spatial expression, can be taken from a given tissue and later be integrated using computational methods to gain biological insight. (B) These measurements can be used to answer multiple questions for a given system or to study different aspects of a complex process such as host–pathogen interactions.

conformation of single cells by using **single-cell Hi-C** (scHi-C) [10]. Histone modifications in individual cells have only recently been effectively profiled [11,12]. This development may significantly advance the study of transcriptional regulation in different cell types. We also envisage that other sequencing methods previously performed in cells 'in bulk' might attain single-cell resolution in the not so distant future.

Protein profiling in single cells has seen important advances using mass spectrometry, such as SCoPe-MS [13]; however, more reliable approaches use a panel of barcoded antibodies whose signal can be amplified by sequencing [14,15]. An exciting example is **cellular indexing of transcriptomes and epitopes by sequencing** (CITE-seq), which has greatly improved the identification of known immune subsets by combining scRNA-seq with surface protein profiling [14].

To properly understand cellular mechanics, it is necessary to combine multiple measurements from RNA, DNA, and protein (Figure 1B). Integrating these molecular layers can show how regulatory networks in cells contribute to shaping the immune system. Methods have been developed for using single-cell sequencing data to infer these networks [16–18] and to integrate the different molecular profiles of single cells [19,20]. Combining these with **pseudotime inference** can inform on the regulation of dynamic processes in immunology, such as infections and development.

**Single-cell CRISPR/Cas9 screens** can help us to learn about variation and robustness in cellular responses. These have been used to dissect T cell receptor (TCR) signaling and response to lipopolysaccharide (LPS) in dendritic cells [21,22]. The use of these technologies is still in its infancy, but we predict that they will be key in elucidating the molecular mechanisms behind complex diseases. CRISPR/Cas9 screens have been accompanied by significant breakthroughs in computational analysis, and further combination with DNA or protein profiling should propel the development and validation of causal inference methods [23], yielding interpretable and actionable models of immunobiology.

## Glossary

**Assay for transposase-accessible chromatin (ATAC-seq)**: uses the Tn5 enzyme to detect open chromatin regions.
**Autoencoders**: single-layer neural networks that learn the optimal way to compress and regenerate data. They can be especially useful for non-linear dimensionality reduction and data denoising, a case where variational autoencoders (VAEs) are mostly used.
**Batch alignment methods**: computational algorithms to combine datasets generated with large batch effects, eliminating their technical differences.
**Capsule networks (CapsNet)**: artificial neural networks designed to better model hierarchical relationships between data by adding 'capsules' to reuse the output from different layers of the neural network.
**Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)**: a method to quantify protein presence using sequencing via antibodies carrying a molecular barcode. It can also be used for cell hashing – barcoding the cells to allow multiplexed cell capture from different samples or donors.
**Classifiers**: machine and statistical learning algorithms used, in the context of single cells, to attribute a label (such as cell type and treatment) to a cell according to gene expression.
**Copy-number variation (CNV)**: genomic segments of variable length that differ in the number of copies per cell.
**Generative adversarial networks (GANs)**: neural network algorithms where one neural network that generates outputs of one type has its performance evaluated by another discriminative neural network.
**Generative model**: a machine learning method that models the conditional probability of an observable variable given a target. The generative model can produce random outcomes of either the observation or target. scANVI and scGen are two frameworks developed specifically for single-cell sequencing data.

## Cell-Type References to Study Disease

Since the first scRNA-seq study on five mouse blastomeres [24], the use of single-cell sequencing technologies has seen exponential growth [1]. scRNA-seq is currently the method of preference to define cell states, study developmental trajectories, and characterize unknown cell populations. The rapid acquisition of large datasets surveying multiple organs [25,26], from different organisms [27] and at different stages of development [28], can allow us to perform informed experimental designs to answer outstanding questions in the field of immunobiology. This increase in throughput has been achieved partly thanks to a reduction in sequencing costs, but mostly due to improvements in cost-effective cell isolation (discussed in detail in [1]). Experimental innovations such as cell hashing [29] and **split-pool approaches** [30,31] can enable significant increases in the number of cells and donors profiled with scRNA-seq. Multidonor designs hold the promise of linking cell type-specific expression to specific diseases or variants, as recently reported in human blood cells [32].

As more single-cell studies move towards unraveling cell-specific responses in the immune system, cell-type annotation has been facilitated by computational methods matching cell populations across samples, tissues, and species [33] (Figure 2A); some **classifiers**, such as Moana [33] and Garnett [34,35], have added a layer of hierarchical stratification of cellular identity [34,35]. Recent work [36] has taken the predictive approach a step further by combining variational **autoencoders** and **latent space vector arithmetics** to build computational models that are capable of predicting cell type-specific responses based on how other cells types respond to the same stimulus. This method has accurately predicted the transcriptional responses of different human peripheral blood mononuclear cells (PBMCs) to IFN-β stimulation in culture, based on gene expression variations of the remaining unrelated cell types; it has also predicted species-specific responses of phagocytes to LPS. Strategies based on connectionist systems, such as artificial neural networks (Box 1), might soon provide accurate predictive models that could potentially facilitate large-scale, transcriptome-wide studies of immune responses *in silico* (Figure 2C).

Pairwise correspondence of datasets can be useful to dissect specific immune processes. However, systems-level insights will come from integrated cross-tissue datasets. The vast data collections that will make up the Human Cell Atlas [36] will necessarily include an Immune Atlas of our species [37]. Comparing novel data with inclusive references might also accelerate interpretation, allowing parallels to be immediately drawn across profiled tissues at steady-state or under disease conditions, and can eliminate the need to profile healthy subjects for disease studies. Establishing such references requires the development of global cell-identity models and the adoption of curated hierarchical cell-type annotations [38,39]. Nonetheless, immune cell phenotypes are also reflected in DNA modifications and protein expression, thus requiring computational methods to define cells beyond RNA molecule expression.

## Hidden Molecular Layers of Cellular Phenotypes

Most cellular heterogeneity is reflected at the level of RNA expression, which can be used to characterize cell states based on markers and functional pathways. Nevertheless, multiple efforts have further probed the data for other features that can expand cellular phenotyping.

High-throughput sequencing reads are at the base of expression measurements. Isoform analysis has also been an important parameter in transcriptomics but, aside from a small number of studies [40,41], remains understudied at the single-cell level. Even so, splicing variability can be highly informative in the context of an immune response. For instance, using logistic regression for differential expression analysis of scRNA-seq data has identified different isoforms of CD45 in human T cells [42], and scRNA-seq using long-read sequencing methods has added more detailed information regarding the importance of splicing in cell identity and disease [41]. Differential detection of spliced and unspliced reads can also reflect transcriptional changes in the developmental trajectories of cells, with the assumption that unspliced transcripts are located in the nucleus and are more recently transcribed than those in the cytoplasm. This application of RNA kinetics to scRNA-seq data is termed

**Latent space vector arithmetics**: mathematical operations in a reduced dimension space that, through generative processes, can be translated into new artificially generated data.
**Lineage tracing**: methods to infer cellular lineages by tracking artificial constructs or endogenous sequences.
**Long-/short-term memory (LSTM) neural networks**: artificial neural networks containing both forwards and backwards connections. LSTMs can process sequences of datapoints such as nucleotide sequences or time-series expression datasets.
**Proteomics**: the study of all the proteins produced in a cell. At the single-cell level, this can be achieved using SCoPE-MS, a mass spectrometry-based method.
**Pseudotime inference**: a computational approach to infer a continuous trajectory for single-cell data, often ordering time-course experiments to reflect temporal changes in gene expression.
**RNA velocity**: a concept representing the dynamic change in the transcriptome of a cell, as modeled based on spliced (current) versus unspliced (novel) transcripts.
**Single-cell CRISPR/Cas9 screens**: CRISPR screens at the single-cell level can evaluate the cell type-specific response to a perturbation at the transcriptome level.
**Single-cell Hi-C (scHi-C)**: a high-throughput sequencing and chromatin conformation capture (Hi-C) method to detect chromatin contacts in individual cells.
**Single-cell RNA sequencing (scRNA-seq)**: methods for obtaining the transcriptome of individual cells. They vary in cell capture methods and portions of transcripts sequenced. The most widely used methods are Smart-seq2 (full-length transcripts, plate-based) and Chromium (3′ or 5′ ends, droplet-based).
**Single-nucleotide polymorphisms (SNPs)**: differences in individual genomic bases across populations (and cells).
**Spatial transcriptomics (ST)**: methods to unbiasedly capture gene expression samples while maintaining their spatial resolution.

**RNA velocity** [42] and, among other uses, has been combined with pseudotime inference to confirm the direction of adaptation of murine T regulatory cells from a lymph node to a barrier tissue [43].

Early approaches such as TraCeR have been devised to reconstruct expressed TCRs from scRNA-seq reads and determine cell clonality [44]. This method has further been extended to B cells [45], incorporating an additional lineage reconstruction step to account for somatic hypermutation events at the B cell receptor (BCR) locus. These methods were initially designed for full transcript sequencing approaches such as Smart-seq2 [46], but they can also be applied to droplet-based protocols, including 10X Genomics VDJ-seq. The combination of VDJ and RNA-seq at a large scale has given new insights into the relationship between activation and TCR sequences of clonotypes in the breast tumor microenvironment [47]. Moreover, increased resolution of TCR and BCR clonality has also been achieved by long-read sequencing, providing detailed descriptions of immune repertoires in various cancers [48]. Ultimately, exploration of adaptive immunity repertoires can advance our understanding of the bias and selection of TCR and BCR chain pairs, and, together with single-cell profiling of antigen specificity, aid in inferring the association between sequence motifs and specific antigens, and presumably diseases [49,50].

**V(D)J recombination** at the TCR and BCR loci can also be treated as barcodes for clonally related cells and used to track clonotype expansion or migration [43,51]. In other cell types, however, different **lineage tracing** approaches must be employed. For model organisms, artificial barcoding systems can be combined with single-cell transcriptome profiling to track and characterize cell lineages [52,53]. Tracking cell-type lineages can enable mapping the ontogeny of immune cell types, as well as other phenomena such as cell trafficking to different tissues or tumors.

Heterogeneity in single-cell data can also be found at the genome level. Transcriptomic reads can be used to call transcript variants, such as **single-nucleotide polymorphisms** (SNPs) or fusion genes, although comparing them to the original genome is recommended. This principle has been used to study the human maternal–fetal interface to assign a maternal or fetal origin to immune cells in the placenta [54]. Other methods focus on studying larger **copy-number variation** (CNV) patterns, and deduce these variations from expression data [55]. Under circumstances where genes are highly mutated, such as cancer, a full cell lineage can be reconstructed and directly compared to its expression profile. Recently, computational approaches have converged on leveraging naturally occurring somatic mutations to undertake lineage tracing in unmodified human cells [56,57]. In particular, mitochondrial DNA is present in an elevated number of independent and heterogeneous copies per cell. Thus, a high number of mitochondrial reads obtained from scATAC-seq and scRNA-seq can be used to establish clonal relationships in cells from healthy individuals [58,59]. Lineage tracing in wild-type human cells can add an informative layer about cellular origin to gene expression studies, and be broadly applied to track lineage relationships between any cell types.

Relationships between the presence of specific immune cell types and disease have been demonstrated [60], and disease-associated variants have been linked to immune cell type marker genes in diseases such as asthma [61]. In the short term, combining scRNA-seq and genotyping may enable studies on the impact of genetic diversity on cell-type abundance and on specific immune responses.

## Host–Pathogen Interactions at the Cellular Level

With the latest advances in single-cell omic technologies, we can now study host–pathogen interactions at single-cell resolution. Recent studies identified marker genes that defined populations of human CD4$^+$ T cells that were more prone to HIV-1 infection than others [62]; moreover, transcriptional programs defining HIV-1 latency and reactivation in host cells were analyzed simultaneously, revealing that not all cells were infected in the same way, and that the rate of CD4$^+$ T cell invasion could be affected by the virus genotype [63]. Viral molecules can also be sequenced together with the host transcriptome to quantify viral loads of infected cells, which can have a significant impact in the interpretation of immune responses to intracellular pathogens [64–67].

**Split-pool approaches**: combinatorial barcoding methods used to provide probabilistically unique barcodes to each cell.
**V(D)J recombination**: genomic rearrangements between the variable (V), diversity (D), and joining (J) regions in the B cell receptor (BCR) or T cell receptor (TCR) loci, that generate variability in receptor chain peptides. These rearrangements can captured by full transcript sequencing methods or VDJ-seq, and used as endogenous genetic barcodes for lineage tracing of immune cells. Computationally, they can be deconvoluted from single-cell RNA-seq data using the TraCeR program. BCR sequences have additional variability generated by somatic hypermutation which introduces random mutations in the BCR locus and can generate receptors with increased affinity for specific antigens.

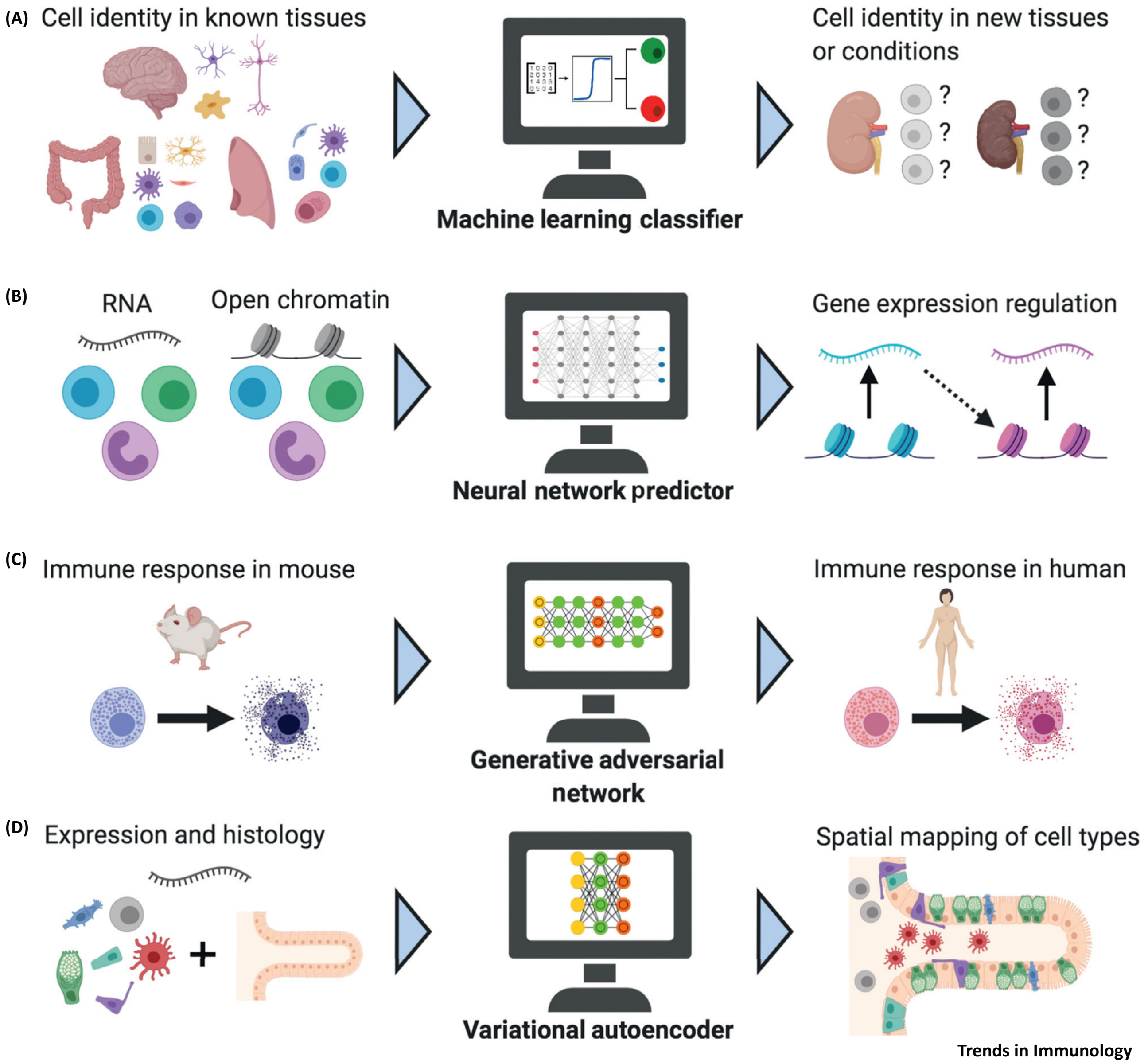
(A)
Cell identity in known tissues
Machine learning classifier
Cell identity in new tissues or conditions
?
(B)
RNA
Open chromatin
Neural network predictor
Gene expression regulation
(C)
Immune response in mouse
Generative adversarial network
Immune response in human
(D)
Expression and histology
+
Variational autoencoder
Spatial mapping of cell types
Trends in Immunology

**Figure 2. Artificial Intelligence Meets Immunobiology.**
(A) Machine learning-based methods of classification – such as logistic regression – can be used to learn the transcriptional features of different cell states and use these learned models to recognize these cell types in new tissues and under multiple conditions. (B) Multilayer neural networks can identify relationships between expression datasets (scRNA-seq) and transcriptional regulation (scATAC-seq) to predict regulatory circuits under different cellular conditions. (C) Generative adversarial networks can model single-cell data produced from animal models and evaluate how different perturbations (i.e., infection) may alter cellular states in the model under study or in other models. (D) Expression data from individual cells and whole tissues can be integrated using an autoencoder to map expression trajectories to tissue coordinates. Abbreviations: scRNA-seq, single-cell RNA deep sequencing; scATAC-seq: single-cell assay for transposase-accessible chromatin combined with deep sequencing.

scRNA-seq has already been applied to studying the transcriptional heterogeneity of parasites from the *Plasmodium* genus [68], and can be used as a method to uncover putative diagnostic markers in other parasitic diseases [69]. Unicellular pathogens, such as *Kinetoplastids*, change their genomes to increase their repertoire of surface molecules to evade the host immune

### Box 1. Artificial Neural Networks for Single-Cell Data Analysis

The large scale of recently generated single-cell datasets [28,97] suggests that traditional analytical methods may not be enough to fully understand a system, and complementary methods to fully exploit these data may be needed. This has led researchers to apply methods from other fields such as physics, artificial intelligence, and machine learning to the study of single-cell multi-omic data (Figure 2). Most methods from artificial intelligence are derived from connectionist systems, and these include autoencoders and deep neural networks [98].

Machine learning uses pattern recognition and statistical inference algorithms for finding relationships or patterns in large collections of data with (supervised learning) or without (unsupervised learning) the use of explicit instructions [99]. Supervised learning methods require a set of examples for use as training data such that the algorithm can later try to fit this model to new datasets. Unsupervised methods are applied without any previous training step and try to learn patterns in the data.

Both autoencoders and neural networks can be used in a supervised or unsupervised way, and this decision should be based on the problem that the algorithm is intended to solve.

One of the main problems of single-cell data is the experimental 'noise' that accompanies them, and one appealing way to deal with this issue is to 'clean' the data using an autoencoder. The autoencoder uses one or multiple datasets to identify features in the data that are common to all datasets, and assigns a probability of which feature does or does not represent the original data [100]. The latent variable obtained by the autoencoder can then be used to reconstruct the data but without the noise or batch effects (Figure 2C).

Artificial neural networks have become an attractive tool to study single-cell omic data. In a standard neural network each artificial neuron is arranged in a layer and connected to other artificial neurons within or between layers. The first layer captures different types of inputs that are then passed on to the underlying layers for data abstraction; the final layer collects these results to produce an output [87]. Depending on the design, the neural network can have only a few or thousands of layers. In this way, an artificial neural network can be used to take multiple data inputs, such as expression values, protein abundance, and tissue localization, to identify a specific cell type (Figure 2D).

response; thus, it is of interest to simultaneously study changes in both pathogen genome and transcriptome, which might be achieved using simultaneous genome and transcriptome sequencing (G&T-seq) [70,71].

Microbiology has relied on the study and characterization of bulk cultured isolates, although these stocks are highly heterogeneous [72]. Single-cell technologies can measure this heterogeneity to gain better insights into pathogen population dynamics and the molecular mechanisms involved in their antigenic variation. CITE-seq can be used to measure pathogen surface virulence factors and the receptors expressed by immune cells to identify clonal, stage-specific pathogen antigens together with immune subpopulations needed for the control of infection. However, the use of single-cell **proteomics** for pathogens is hampered by the limited repertoire of antibodies against conserved regions of many surface virulence factors.

The integration of multiomic datasets is certain to change the field of infectology [73], but first it is crucial that single-cell datasets of well-characterized isolates are generated to explore the level of heterogeneity and plasticity of different pathogens. Because many pathogens rely on genomic polymorphisms to evade the immune system, advancing technologies to capture this feature as accurately as possible is also imperative [74].

## High-Resolution Spatial Tissue Maps

The underlying molecular complexity of cells makes them remarkably adaptable machines that are fully equipped to sense and react to the surrounding environment. This is key for immune cells that not only drive specialized responses to pathogens but also fulfill particular homeostatic roles in tissue development and maintenance.

In sum, context matters. The neighbors of a cell can influence its function via cell–cell interactions, established through receptor–ligand pairs. For scRNA-seq, these relationships can be assessed using CellPhoneDB (www.cellphonedb.org) [54]. Approaches such as this can be used to characterize cell–cell communication not only in homeostatic tissues but also under pathological conditions. It has recently been used in lung-derived scRNA-seq to reveal unique interactions between type 2 T helper (Th2) cells and mesenchymal cells in asthmatic human donors [61]. Despite advances such as this example, the inference of cell interactions from transcriptomic data is still in its infancy, and methods that can integrate known interactions and expression data will be important in understanding coordinated cellular responses in the context of disease; they may also aid in unraveling a wide array of immune cell functions across tissues.

Cell–cell contacts are key to understanding how cells organize into tissues. A bone marrow study used mild tissue dissociation followed by microdissection, and recorded the interactions of cell pairs before sequencing, revealing stable interactions unique to neutrophils [75]. Spatial transcript profiling methods have indeed seen steady progress in resolving RNA map associations and heterogeneities in tissue slices. Approaches with fluorescent probes have been designed and scaled to work with thousands of transcripts [76]. Sequencing-based **spatial transcriptomics** (ST) methods have also been developed to combine histology and unbiased transcriptome profiling [77]. Recently, two protocols achieved single-cell resolution of tissue slices, giving tissue-wide transcription patterns a direct link to the cells generating them [78,79]. Although improvements in cellular spatial profiling are following this direction, it is also computationally possible to match scRNA-seq data with spatial data from different sources. Such approaches have relied on adaptations or improvements of **batch alignment methods** that are used to integrate different scRNA-seq datasets [80–82], but the field is still exploring novel methods to increase accuracy across different platforms and modalities. A recent approach applying multimodal intersection analysis successfully integrated cell populations – identified with scRNA-seq – with tissue architecture – defined using ST [83]. These advances are relevant because spatial profiling of transcriptomes and TCR/BCR sequences could inform how immune cells can regulate tolerance and immune reactions in different contexts – for instance, within tumor microenvironments.

## Artificial Intelligence for Single-Cell Omic Studies

The sheer size of single-cell data indicates that standard data analysis techniques that researchers have previously used for small experiments may not be able to fully take advantage of these datasets. Nevertheless, techniques from the field of artificial intelligence and machine learning can allow not only the processing of large, complex datasets but also the identification of hidden patterns and relationships that are not obvious to the human analyst (Figure 2B).

**Generative models** such as scANVI [84] and scGen [36] allow data integration, clustering, and marker identification of single-cell datasets; however, these deep generative models can also be used to produce data simulations to predict how a given cell population might react to a given stimulus or insult. **Generative adversarial networks** (GANs) have been successful in other tasks in biomedicine, such as the accurate diagnosis of skin cancer [85], and are starting to be used to analyze single-cell datasets [86]. These generative methods are useful in that they provide a powerful tool to study a system by taking advantage of single-cell datasets from healthy individuals, as well as from vast bulk datasets already generated from multiple conditions and perturbations; they can then generate perturbed single-cell datasets for dissection of condition-specific transcriptional circuits that can be subsequently validated *in vitro* [86].

The use of deep learning algorithms [87] for image analysis can facilitate large-scale single-cell transcriptomics integrated with spatial information [88]. Indeed, the integration of these datasets using deep autoencoders has provided important insights into (i) morphological profiling of an entire tissue, (ii) the interrogation of regulatory and transcriptional landscapes of any given tissue [89,90], and (iii) the classification of cell types using their subcellular structures [91]. Some of these methods rely on neural networks, and have been optimized to visualize large-scale single-cell datasets, allowing the mapping of new datasets onto references; they can then be used to visualize millions of cells [92]. A new exciting concept that has been developed recently is that of **capsule networks** (CapsNet); here, neural networks are designed to model hierarchical relationships in data, and can be used directly in single-cell datasets to identify cell types and cell states and how they might interact [93].

The most exciting aspect of artificial intelligence methods is that we can now use a combination of new single-cell reference datasets and the vast amount of data that has been generated in the past decade to gain insight from biological systems. The use of deep learning to classify cell types using multiple types of data, and the potential to use **long/short-term memory (LSTM) neural networks** for text mining using transcriptional signatures, means that we might be able to analyze biological systems in depth using publicly available resources. These could later be used as the input for generative models to identify putative transcriptional circuits that are activated or disrupted following, for example, exposure of an immune cell to a new antigen, or when a pathogen might attempt to evade a host immune system.

## Concluding Remarks

The growth in the size, depth, and breadth of single-cell sequencing experiments over the past 10 years has achieved remarkable proportions. scRNA-seq has allowed us to unbiasedly probe cell identity for the first time, and gain knowledge on sets of transcripts that define a particular cell population. Multimodal data are further expanding the borders of cell identity to the regulatory realm, while spatial approaches are preserving information on the *in vivo* context and contacts of cells.

There is still room for improvement in the throughput of single-cell experiments, which is crucial for profiling cell populations from many individuals. Although new methodologies have focused on a cell throughput increase, this can come at a cost of lower numbers of genes profiled per cell. Expression sparsity can be dealt with computationally [94,95]; nonetheless, improvements in measuring gene numbers per cell should be a medium-term goal for developing improved protocols. In addition, more standardized approaches might be more quickly adopted than custom pipelines, and it is possible that the real impact of cheap high-throughput split-pool methods [30] might come only after these become commercially available.

Most single-cell data produced to date have been analyzed for specific projects, generating publications, but mining such resources to perform more comprehensive meta-analyses might enable the extraction of more biologically relevant information from them. Structuring stored data is a key challenge, not only for modern biological data repositories but also for large consortiums associated with the generation of large datasets – in this particular case, the Human Cell Atlas and similar initiatives.

scRNA-seq offers a static, descriptive snapshot of the transcriptome, and it is in this context that inferred cell identity should be understood. Nevertheless, cell identity can be seen from many perspectives, and perhaps the chief among them relates to cell function. There is a correlation between the transcriptome and cellular function, but other layers of complexity can greatly influence it. This underscores the importance not only of surveying cell heterogeneity at the DNA, RNA, and protein levels but also of how best these should be combined with methods to functionally phenotype cells at a large scale (Figure 2B) [96]. Methods recording the spatial environment are a step in this direction because they may give clues to the effects a cell has on its microenvironment (and vice versa).

The coming challenges in immunology will require acquiring more detailed data on immune cells, with a deeper vision into disease and immune response mechanisms (see Outstanding Questions). The fine resolution of scRNA-seq and related methods requires bridging immune cell biology with systems biology. New data collected and computational approaches developed should focus on enhancing our predictive capabilities of immune reactions, and on assessing how our knowledge of the molecular workings of the immune system can be leveraged to fine-tune novel candidate therapies, moving towards precision medicine for a variety of ailments.

### Outstanding Questions

How can single-cell data be reused and made easily available? Current repositories for sequencing data have not been designed to deal with large numbers of individual files, or associated metadata. New data archives should allow data and metadata to be obtained in a seamless and structured way.

Can data from scRNA-seq, open chromatin, and other elements show us the boundaries of cell identity and plasticity? How flexible are those boundaries in the context of the underlying regulatory networks?

What is the relationship between cell identities inferred from different layers of information?

How much functional information does the combination of data from different layers of cell identity provide?

How much of a cellular phenotype can be explained by genetics?

What populational heterogeneities exist in cell-type composition and plasticity?

What are the distinctive spatial patterns in which immune cells accumulate in tissues? Do they establish specific stable interactions in the steady-state?

## Acknowledgments

T.G. was supported by the EU Horizon 2020 research and innovation program ENLIGHT-TEN under Marie Skłodowska-Curie grant agreement 675395.

## References

1. Svensson, V. *et al.* (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604
2. Svensson, V. and da Veiga Beltrame, E. (2019) A curated database reveals trends in single cell transcriptomics. *bioRxiv* Published online August 21, 2019. https://doi.org/10.1101/742304
3. Hardy, R.R. *et al.* (1982) B-cell subpopulations identified by two-colour fluorescence analysis. *Nature* 297, 589–591
4. Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476
5. Hui, T. *et al.* (2018) High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. *Stem Cell Rep.* 11, 578–592
6. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
7. Buenrostro, J.D. *et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490
8. Cusanovich, D.A. *et al.* (2015) Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914
9. Chen, X. *et al.* (2018) A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* 9, 5345
10. Nagano, T. *et al.* (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64
11. Kaya-Okur, H.S. *et al.* (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* 10, 30
12. Wang, Q. *et al.* (2019) CoBATCH for high-throughput single-cell epigenomic profiling. *Mol. Cell* Published online August 27, 2019. https://doi.org/10.1016/j.molcel.2019.07.015
13. Budnik, B. *et al.* (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* 19, 161
14. Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868
15. Peterson, V.M. *et al.* (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939
16. Aibar, S. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086
17. Pliner, H.A. *et al.* (2018) Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* 71, 858–871
18. Papili Gao, N. *et al.* (2018) SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 34, 258–266
19. Argelaguet, R. *et al.* (2018) Multi-omics factor analysis – a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* 14, e8124
20. Angermueller, C. *et al.* (2017) DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18, 67
21. Datlinger, P. *et al.* (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301
22. Dixit, A. *et al.* (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866
23. Qiu, X. *et al.* (2018) Towards inferring causal gene regulatory networks from single cell expression measurements. *bioRxiv* Published online September 25, 2018. https://doi.org/10.1101/426981
24. Tang, F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382
25. Tabula Muris Consortium *et al.*. (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372
26. Han, X. *et al.* (2018) Mapping the mouse cell atlas by microwell-seq. *Cell* 173, 1307
27. Cao, J. *et al.* (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667
28. Cao, J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502
29. Stoeckius, M. *et al.* (2018) Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* 19, 224
30. Rosenberg, A.B. *et al.* (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182
31. Kang, H.M. *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94
32. van der Wijst, M.G.P. *et al.* (2018) Single-cell RNA sequencing identifies cell type-specific *cis*-eQTLs and co-expression QTLs. *Nat. Genet.* 50, 493–497
33. Kiselev, V.Y. *et al.* (2018) Scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15, 359–362
34. Wagner, F. and Yanai, I. (2018) Moana: a robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv* Published online October 30, 2018. https://doi.org/10.1101/456129
35. Pliner, H.A. *et al.* (2019) Supervised classification enables rapid annotation of cell atlases. *bioRxiv* Published online February 25, 2019. https://doi.org/10.1101/538652
36. Lotfollahi, M. *et al.* (2018) Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. *bioRxiv* Published online December 14, 2018. https://doi.org/10.1101/478503
37. Regev, A. *et al.* (2017) The human cell atlas. *Elife* 6, e27041
38. Bard, J. *et al.* (2005) An ontology for cell types. *Genome Biol* 6, R21
39. Meehan, T.F. *et al.* (2011) Logical development of the cell ontology. *BMC Bioinformatics* 12, 6
40. Song, Y. *et al.* (2017) Single-cell alternative splicing analysis with Expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67, 148–161
41. Gupta, I. *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* 36, 1197–1202

42. Ntranos, V. *et al.* Identification of transcriptional signatures for cell types from single-cell RNA-seq. *bioRxiv* Published online February 14, 2018. https://doi.org/10.1101/258566
43. Miragaia, R.J. *et al.* (2019) Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation. *Immunity* 50, 493–504
44. Stubbington, M.J.T. *et al.* (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* 13, 329–332
45. Lindeman, I. *et al.* (2018) BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* 15, 563–565
46. Picelli, S. *et al.* (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181
47. Azizi, E. *et al.* (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* 174, 1293–1308
48. Singh, M. *et al.* (2019) High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.* 10, 3120
49. Glanville, J. *et al.* (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature* 547, 94–98
50. Dash, P. *et al.* (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547, 89–93
51. Lönnberg, T. *et al.* (2017) Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci Immunol* 2, eaal2192
52. Raj, B. *et al.* (2018) Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442–450
53. Spanjaard, B. *et al.* (2018) Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* 36, 469–473
54. Vento-Tormo, R. *et al.* (2018) Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353
55. Müller, S. *et al.* (2018) CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* 34, 3217–3219
56. Campbell, K.R. *et al.* (2018) Clonealign: statistical integration of independent single-cell RNA & DNA-seq from human cancers. *Genome Biol.* 20, 54
57. McCarthy, D.J. *et al.* (2018) Cardelino: integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. *bioRxiv* Published online November 26, 2018. https://doi.org/10.1101/413047
58. Xu, J. *et al.* (2018) Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *bioRxiv* Published online November 29, 2018. https://doi.org/10.1101/480202
59. Ludwig, L.S. *et al.* (2019) Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* 176, 1325–1339
60. Keren-Shaul, H. *et al.* (2017) A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* 169, 1276–1290
61. Vieira Braga, F.A. *et al.* (2019) A cellular census of healthy lung and asthmatic airway wall identifies novel cell states in health and disease. *bioRxiv* Published online January 23, 2019. https://doi.org/10.1101/527408
62. Rato, S. *et al.* (2017) Single-cell analysis identifies cellular markers of the HIV permissive cell. *PLoS Pathog.* 13, e1006678
63. Golumbeanu, M. *et al.* (2018) Single-cell RNA-seq reveals transcriptional heterogeneity in latent and reactivated HIV-infected cells. *Cell Rep.* 23, 942–950
64. Zanini, F. *et al.* (2018) Virus-inclusive single-cell RNA sequencing reveals the molecular signature of progression to severe dengue. *Proc. Natl. Acad. Sci. U. S. A.* 115, E12363–E12369
65. Russell, A.B. *et al.* (2018) Extreme heterogeneity of influenza virus infection in single cells. *Elife* 7, e32303
66. Wyler, E. *et al.* (2019) Single-cell RNA-sequencing of herpes simplex virus 1-infected cells identifies NRF2 activation as an antiviral program. *bioRxiv* Published online March 4, 2019. https://doi.org/10.1101/566992
67. Drayman, N. *et al.* (2019) HSV-1 single cell analysis reveals anti-viral and developmental programs activation in distinct sub-populations. *Elife* 8, e46330
68. Reid, A.J. *et al.* (2018) Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* 7, e33105
69. Nötzel, C. *et al.* (2018) Single-cell transcriptome profiling of protozoan and metazoan parasites. *Trends Parasitol* 34, 731–734
70. Macaulay, I.C. *et al.* (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522
71. Macaulay, I.C. *et al.* (2017) Single-cell multiomics: multiple measurements from single cells. *Trends Genet* 33, 155–168
72. Ackermann, M. (2015) A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol.* 13, 497–508
73. Woyke, T. *et al.* (2017) The trajectory of microbial single-cell sequencing. *Nat. Methods* 14, 1045–1054
74. Blecher-Gonen, R. *et al.* (2019) Single-cell analysis of diverse pathogen responses defines a molecular roadmap for generating antigen-specific immunity. *Cell Syst.* 8, 109–121
75. Boisset, J.-C. *et al.* (2018) Mapping the physical network of cellular interactions. *Nat. Methods* 15, 547–553
76. Chen, K.H. *et al.* (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348, aaa6090
77. Ståhl, P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82
78. Vickovic, S. *et al.* (2019) High-density spatial transcriptomics arrays for *in situ* tissue profiling. *bioRxiv* Published online March 13, 2019. https://doi.org/10.1101/563338
79. Rodriques, S.G. *et al.* (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467
80. Lopez, R. *et al.* (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058
81. Stuart, T. *et al.* (2019) Comprehensive integration of single cell data. *Cell* 177, P188–1902.E21.
82. Welch, J. *et al.* (2018) Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv* Published online November 2, 2018. https://doi.org/10.1101/459891
83. Moncada, R. *et al.* (2019) Integrating single-cell RNA-seq with spatial transcriptomics in pancreatic ductal adenocarcinoma using multimodal intersection analysis. *bioRxiv* Published online March 13, 2019. https://doi.org/10.1101/254375
84. Xu, C. *et al.* (2019) Harmonization and annotation of single-cell transcriptomics data with deep generative models. *bioRxiv* Published online January 29, 2019. https://doi.org/10.1101/532895

85. Esteva, A. *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118
86. Ghahramani, A. *et al.* (2018) Generative adversarial networks simulate gene expression and predict perturbations in single cells. *bioRxiv* Published online July 30, 2018. https://doi.org/10.1101/262501
87. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
88. Grønbech, C.H. *et al.* (2019) scVAE: variational auto-encoders for single-cell gene expression data. *bioRxiv* Published online May 21, 2019. https://doi.org/10.1101/318295
89. Chan, T.E. *et al.* (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267
90. Meng, N. *et al.* (2018) Large-scale multi-class image-based cell classification with deep learning. *IEEE J. Biomed. Health Inform.* 23, 2091–2098
91. Ozaki, Y. *et al.* (2019) Label-free classification of cells based on supervised machine learning of subcellular structures. *PLoS One* 14, e0211347
92. Cho, H. *et al.* (2018) Generalizable and scalable visualization of single-cell data using neural networks. *Cell Syst.* 7, 185–191
93. Wang, L. *et al.* (2019) scCapsNet: a deep learning classifier with the capability of interpretable feature extraction, applicable for single cell RNA data analysis. *bioRxiv* Published online May 21, 2019. https://doi.org/10.1101/506642
94. Svensson, V. (2019) Droplet scRNA-seq is not zero-inflated. *bioRxiv* Published online March 19, 2019. https://doi.org/10.1101/582064.
95. Hafemeister, C. and Satija, R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv* Published online March 18, 2019. https://doi.org/10.1101/576827
96. Noble, D. (2012) A theory of biological relativity: no privileged level of causation. *Interface Focus* 2, 55–64
97. Pijuan-Sala, B. *et al.* (2019) A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490–495
98. Zou, J. *et al.* (2019) A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18
99. Grabowski, P. and Rappsilber, J. (2019) A primer on data analytics in functional genomics: how to move from data to insight? *Trends Biochem. Sci.* 44, 21–32
100. Doersch, C. (2016) Tutorial on variational autoencoders. *arRxiv* Published online June 19, 2016. https://arxiv.org/abs/1606.05908