Abstract

Single-cell technologies have revolutionized biomedical research by enabling scalable measurement of the genome, transcriptome, proteome, and epigenome of multiple systems at single-cell resolution. Now widely applied to cancer models, these assays offer new insights into tumour heterogeneity, which underlies cancer initiation, progression, and relapse. However, the large quantities of high-dimensional, noisy data produced by single-cell assays can complicate data analysis, obscuring biological signals with technical artifacts. In this review article, we outline the major challenges in analyzing single-cell cancer genomics data and survey the current computational tools available to tackle these. We further outline unsolved problems that we consider major opportunities for future methods development to help interpret the vast quantities of data being generated.

1. Introduction

Cancers are invasive neoplasms derived from single cells that undergo successive acquisition of cell-specific properties through somatic mutation. The clonal diversity that results can lead to metastasis, histologic transformation, and treatment resistance (Greaves and Maley 2012). The trellis of phenotypes generated from this process of branching evolution is subject to the selective pressures of the tumour microenvironment (TME, see table 1, McGranahan et al 2017, Zhang et al 2018). High-throughput approaches such as bulk whole-genome sequencing, RNA-seq, flow cytometry, mass cytometry, and immunohistochemistry have been extensively applied to establish the foundations of our current understanding of carcinogenesis and the cancer cell-microenvironment interface (Weinstein et al 2013, Shah et al 2012, Curtis et al 2012). However, these approaches are generally limited by tradeoffs between the breadth and granularity of genotypic and phenotypic information obtained.

Recent technological advances have enabled multimodal profiling of the genomes, transcriptomes, and proteomes from thousands to millions of single cells. The advent of single-cell RNA-sequencing (scRNA-seq) at scale has resulted in an explosion of recent studies capturing not only malignant phenotypes but also cellular states of immune and stromal cells in the tumour microenvironment (Zhang et al 2019a). Studies using scRNA-seq have highlighted unexpected heterogeneity in the transcriptomes of single tumour cells, which is likely driven by a multitude of underlaying processes including genetic variation (Patel et al 2014), the presence of rare cancer stem cell populations (Nassar and Blanpain 2016), aberrant development hierarchies (Tirosh et al 2016, Gojo et al 2020), and epigenetic plasticity in the context of therapies (Bell et al 2019).

In addition to gene expression profiling, advances in DNA sequencing technologies that allow for single-cell measurements of point mutations or large scale copy number aberrations have enabled the clonal decomposition of primary tumours, cell lines, and...
Table 1. Definitions of major concepts relevant to computational modelling of single-cell cancer genomics data.

| Concept                  | Definition                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| Tumour microenvironment  | The collection of immune, stromal, and vascular cells that may surround or infiltrate a tumour |
| Ploidy                   | The number of sets of chromosomes in a cell                                  |
| B-allele frequency       | Ratio of intensities between two alleles at a specified locus (heterozygous diploid = 0.5) |
| Phylogenetic inference   | Reconstruction of the ancestral mutation tree that gave rise to the observed mutational profile in a tumour |
| Clone                    | Group of tumour cells that share a similar mutation profile                 |
| Infinite sites assumption| A given site or loci will be mutated at most once in the life history of a tumour |
| Flow cytometry           | A method to measure physical properties of cells including protein expression via fluorescence and the scattering of light |
| Overfitting              | Phenomena in a statistical or machine learning model where parameter estimates or predictions are overly influenced by the training dataset used and do not generalize well to new datasets |
| Copy number aberrations  | Regions of the genome that have been amplified or depleted due to double strand breakage |
| Markov Chain Monte Carlo | A set of algorithms to sample from a probability distribution that is impossible to calculate exactly |
| Approximate Bayesian computation | A method of parameter inference when the probability of the data given parameters is difficult to compute, but data can be simulated given the parameters |

Figure 1. Technologies and computational methods for single-cell cancer genomics. (A) Experimental technologies frequently used in cancer genomics for profiling single-cells discussed in this review. (B) Examples of common computational challenges and current solutions in the analysis of single-cell cancer genomic data.
patient-derived xenografts in unprecedented detail, leading to new insights into aneuploidy (Laks et al 2019), oncogenic processes (Yu et al 2014), and treatment resistance (Chen et al 2018, Kim et al 2018). These same methods can be used to detect and functionally characterize circulating tumour cells (CTCs), which have the potential to be used as early diagnostic markers or markers of metastasis. Similarly, single-cell proteomic technologies such as mass cytometry allow scalable measurement of multiple protein markers while retaining information about the spatial origins of cells, revealing the high-dimensional architecture of cancers (Jackson et al 2020).

However, the emergence of these technologies has created a deluge of noisy, high-dimensional data, the analysis and interpretation of which is key to understanding cancer pathogenesis and aetiology. To address this, a range of computational methods have been developed specifically for the analysis of cancer genomics data. In this review, we outline the challenges and methods associated with the computational analysis of single-cell genomic, transcriptomic and proteomic data in the context of cancer research (figure 1(a) and (b)). We further point to unsolved problems and future directions in the development of computational tools for single-cell cancer genomics, concerning (i) inferring clonal dynamics over time and space, (ii) understanding the impact of TME heterogeneity on cancer evolution, (iii) machine learning approaches to make single-cell assays predictive of therapeutic response, and (iv) methods to uncover interactions and signatures from spatially resolved data. We intend this review to serve as a point of reference of the current state of the field as well as an opportunity to encourage discussion on future computational methods necessary to realize the potential of single-cell cancer genomics.

2. Computational methods for single-cell cancer genomics

2.1. Mutation profiling and phylogenetic inference

The noise inherent in detecting mutation events from picograms of DNA at the single-cell level—both point mutations and copy number aberrations—has led to a number of methods for mutation identification and the subsequent clustering of cells into clones. SNV detection from targeted single-cell DNA sequencing is complicated by high false negative rates due to amplification failures at heterozygous variants. To address this, probabilistic models such as Monovar (Zafar et al 2016) and the single cell genotypier (Roth et al 2016) have been developed to assign variants by pooling strength across multiple cells. Monovar assumes each locus is independent and calls global variants present in any cell before re-computing cell-specific variant probabilities. In contrast, the single cell genotypier models population structure by assuming cells belong to discrete clusters or clones, and uses a variational Bayes approach to compute the posterior probabilities of single-cell genotypes and cluster assignments, allowing for potential doublets.

Another set of methods performs CNV detection from near-uniform coverage single-cell whole genome sequencing (WGS) data. Preamplification-free approaches, such as direct library preparation (DLP, Zahn et al 2017) and DLP+ (Laks et al 2019) produce low depth WGS data (<~0.05X) with low amplification bias compared to alternative approaches (Hellani et al 2004), making them amenable to copy number inference but not single-cell level variant calling. To analyze the outputs of these assays, existing methods that have been applied to bulk WGS data such as HMMCopy (Ha et al 2012) as well as newer single-cell specific methods that leverage information from multiple cells simultaneously CHISEL (Zaccaria and Raphael 2020) have been used. HMMCopy computes copy number profiles based on a range of ploidy assumptions while correcting for GC content and mappability effects, and then selects the solution that minimizes non-integer copy number predictions. CHISEL uses B-allele frequency information to compute allele- and haplotype-specific copy number, which can be applied to identify cancer-associated loss-of-heterozygosity events (McGranahan et al 2017).

Once single-cell genotypes have been defined, a common secondary analysis is to find a mutation tree that explains the observed data well—a process known as phylogenetic inference—for which a multitude of methods have been proposed. OncoNEM (Ross and Markowetz 2016) uses a probabilistic scoring function to measure how well a given tree fits observed mutation data accounting for false positive and negative rates under the infinite sites assumption, and searches for an optimal tree configuration using a heuristic local search procedure. SCITE (Jahn et al 2016) similarly employs a probabilistic model under the infinite sites assumption, but uses Markov Chain Monte Carlo (MCMC) to sample trees from the posterior probability distribution using prune and reattach moves. Other phylogenetic models have been created to relax some modelling assumptions in previous work, such as SiFit (Zafar et al 2017), which extends the OncoNEM and SCITE to allow for violations of the infinite sites assumption and consequently models recurrent evolution at each site. Refining this idea, Scarlett (Satas et al 2019) only allows for mutation losses in regions for which there is evidence of copy number losses to account for the fact that mutation loss in other regions (with neutral copy number or copy number gains) is highly unlikely. Further methods in this space allow for more complex experimental designs, such as CALDER (Myers et al 2019), which reconstructs phylogenetic trees from longitudinally acquired samples and builds in temporal constraints
to improve the accuracy of the inferred trees. Finally, this inherent tree structure of single-cell genomic data may be exploited to refine single-cell variant calling using methods such as SCIΦ (Singer et al 2018).

2.2. Gene expression

The advent of high-throughput single-cell RNA sequencing methods have ushered an era of intense computational methods development over the last 5 years. Essential analytical elements—quality control (McCarthy et al 2017), dimensionality reduction (Piersen and Yau 2015, Risso et al 2018), and clustering (Kiselev et al 2017, Duò et al 2018)—were the focus of initial studies, and are covered in other reviews (Rostom et al 2017). Early single cell RNA-seq cancer studies leveraged these techniques to determine cell type composition (Li et al 2017), intratumoural heterogeneity (Chen et al 2018), and signatures of therapeutic resistance (Kim et al 2018).

However, some of these elements necessitate revisiting in the context of cancer samples. Many pioneering single-cell RNA-sequencing studies used largely homogeneous cell lines as the substrate, and thresholds on mitochondrial gene percentage and transcribed gene count used to filter out low quality cells were designed for those scenarios (Ilicic et al 2016, O’Flanagan et al 2019). In contrast, cancer cells are associated with higher mitochondrial content (Oorio and Cai 2020), and the process of mechanical or enzymatic disaggregation of solid tumour samples prior to library preparation may result in increased expression of cellular stress markers, including mitochondrial genes. Additionally, plasma cells in the tumour microenvironment express a restricted profile of genes dominated by immunoglobulins, and are subject to being filtered out by transcribed gene thresholds.

Beyond preprocessing, cell type classification enables a broad survey of the tumour microenvironment. Many analyses (Shih et al 2018, Kim et al 2018) employ ad hoc approaches following unsupervised clustering to assign clusters to cell types. Bespoke approaches for cell type assignment that perform comparisons to bulk or single-cell RNA-seq data of purified populations (either experimentally or in silico) have also been developed (Zheng et al 2017), including correlation-based methods such as scmap (Kiselev et al 2018) and singleR (Aran et al 2019). However, in the context of perturbed cell states in the tumour microenvironment, data for comparable purified populations is often unavailable or not reflective of altered expression patterns (Shiga et al 2015, Wherry 2011). To solve this problem, several approaches such as cellassign (Zhang et al 2019a), Garnett (Pliner et al 2019), and SCINA (Zhang et al 2019c) were developed, which leverage broader cell type-specific marker gene data (Zhang et al 2019b) to probabilistically assign single cells to cell types specified in terms of marker genes or an ‘unknown’ category, in the absence of clustering. While these methods perform well when distinguishing between broad cell type lineages, their performance can be limited when distinguishing more closely related cell types, such as regulatory and memory T cells (Grabski and Irizarry 2020).

A further set of methods have been developed to leverage scRNA-seq to link tumour genotype to phenotype (in this case gene expression), both when the genotype is inferred directly from the gene expression or measured using an orthogonal assay. In the former case, the dosage dependence of gene expression on copy number (Venteicher et al 2017, Müller et al 2018) means scRNA-seq data can be leveraged to predict large-scale copy number aberrations, allowing for approximate inference of clonal composition and separation of malignant cells from putatively normal diploid cells. One class of methods to do this infers copy number profiles directly from scRNA-seq data by comparing expression profiles of input cells to a background constructed from normal cells (Fan et al 2018, Serin-Harmanci et al 2020). These methods rely on large contiguous regions of up- or down-regulation relative to background that correspond to copy number-dependent changes in expression rather than alterations in gene regulation. CONICSmat (Müller et al 2018) removes the reliance on a normal (non-tumour) set of cells specified a priori by fitting a bimodal Gaussian to expression values derived from each segment of the genome, but user interpretation of the results is required to distinguish putative non-malignant and malignant cells. CaSpER (Serin-Harmanci et al 2020) additionally incorporates allelic frequency data to correct CNV calls by identifying regions of LOH. All of these methods depend on the reliability of the specified or inferred background cells as a baseline for the euploid state—background cells derived from a different tissue source or patient may lead to over- or under-calling. Alternatively, when ground truth copy number profiles are available from orthogonal data types, such as single-cell whole-genome sequencing, scRNA-seq profiles can be directly aligned to these with CloneAlign (Campbell et al 2019). Somatic SNV information derived from bulk or single-cell DNA-seq data, instead of CNVs, can also be used to align scRNA-seq data (McCarthy et al 2020). In addition to inferring large-scale clonal structure, these methods can also be useful for distinguishing malignant cells from nonmalignant tissue-related cell types when established marker genes are unavailable.

A further promising technological advance in gene expression profiling of tumour cells concerns whole-transcriptome profiling of tissues and single-cell or near-single-cell resolution with the retention of information on the spatial origin of the region profiled. The main technologies in this area—such as the 10X Genomics Visium system—achieve a resolution.
of 100 μm (3–30 cells), with the state-of-the-art attaining 2 μm resolution enabling sub-cellular measurements (Vickovic et al 2019). To interpret these data, methods that attempt to understand the spatial distribution of cell types and states in the tumour microenvironment, either through de novo discovery or integration with disaggregated scRNA-seq, have been developed. These include multimodal intersection analysis (MIA, Moncada et al 2020) which quantifies the overlap between expression cluster- and region-specific marker genes using a hypergeometric statistical test, and spatial transcriptome deconvolution (STD, Maakola et al 2018) which employs a hierarchical Bayesian model to locally deconvolve measured spots into ‘transcriptomic factors’ that serve as a lower dimensional representation of the data.

### 2.3. Protein expression

Mass cytometry approaches (see Spitzer and Nolan 2016 for an overview) extend classical immunohistochemistry to simultaneously interrogate up to 40 distinct markers using heavy metal-tagged antibodies directed to cell surface, cytoplasmic, and nuclear epitopes, at single-cell resolution. Like immunohistochemistry, these techniques can be applied to archival (fixed) material, which is readily available and typically not suitable for single-cell genome or transcriptome characterization. Compared to single-cell RNA-sequencing, mass cytometry costs substantially less per cell and allows users to focus on biologically meaningful proteins rather than primarily highly expressed genes (Spitzer and Nolan 2016), though with the obvious drawback of not capturing full-transcriptome data. There have been many applications of mass cytometry in cancer research, ranging from profiling apoptotic pathways to identifying putative drug targets (Teh et al 2020), building predictive models of patient relapse at diagnosis (Good et al 2018), and quantifying of immune cell infiltration across large patient cohorts (Wagner et al 2019).

Few tools tailored for the analysis of mass cytometry cancer datasets currently exist, though many existing tools can be repurposed to answer cancer-specific questions (see Nowicka et al 2017 for a recommended cytometry workflow). Normalization and signal correction can be applied using tools such as CATALYST (Chevrier et al 2018), which employs bead-based approaches to account for signal spillover due to contaminating channels. A common secondary step is cell type identification, typically performed via clustering algorithms including bespoke models for mass cytometry such as Phenograph (Levine et al 2015). Supervised marker gene-based approaches, comparable to cellassign or Garnett, have yet to be tested in mass cytometry data. Finally, methods such as SPADE (Qiu et al 2011) allow for the ordering of cell populations based on gradual changes in protein expression, with successful applications identifying cellular transitions in breast cancers (Giesen et al 2014).

Recently, multiple mass cytometry-based methods that measure the cellular proteome at single cell-level or subcellular resolution that retain spatial information of cells in situ have been developed. These methods, collectively referred to as mass cytometry imaging (MCI) methods, include imaging mass cytometry (IMC, Giesen et al 2014), which can profile to a spatial resolution of 1 μm, and multiplexed ion beam imaging (MIBI, Angelo et al 2014), which can profile to 200–300 nm. These methods provide new insights into homootypic and heterotypic cell-to-cell interactions, immune infiltration, and physical microenvironment architecture that cannot be inferred from disaggregated data alone.

The analysis of MCI data consists of 4 major steps: (1) cell segmentation, (2) normalization, (3) cell type assignment, and (4) spatial analysis. As steps (2) and (3) use no specific spatial information they largely follow the process for disaggregated mass cytometry data as described above. Multiple methods exist for segmentation, including CellProfiler (McQuin et al 2018), which implements a two-step approach, using markers to first establish nuclear boundaries and then predict cellular boundaries, which are typically more variable in size and shape. Additional approaches such as Ilastik (Berg et al 2019) and DeepCell (Van Valen et al 2016) are supervised approaches that can be used to establish cellular boundaries based on pixel classification by leveraging random forests and convolutional neural networks, respectively.

Currently, there are few computational methods that integrate the available spatial and phenotypic information to provide insights into cell–cell interactions or tumour architecture, with most studies to date employing ad hoc methods to assess broad patterns in spatial proximity and composition. For example, (Jackson et al 2020) use a Louvain algorithm to find clusters of neighbouring cells (‘spatial communities’) and characterize these communities according to cell type abundance. Similarly, Keren et al (2018) used proportions of co-localizing tumour and immune cells to define three archetypal communities: cold, compartmentalized, and mixed. Initial analyses (Keren et al 2018, Jackson et al 2020) show that this type of spatial information may improve prediction of patient survival based on cell type abundance. When larger panels of proteins can be simultaneously profiled at the single-cell level, methods that leverage existing receptor-ligand databases such as CellPhoneDB (Efremova et al 2020) may provide insights into context-dependent cell–cell interactions in the tumour microenvironment.

### 2.4. Epigenetics

Concurrent with the rise of single-cell genomic, transcriptomic, and proteomic technologies has been the creation of single-cell epigenetic technologies. These
3. Unsolved problems and future directions

There has been a remarkable drive by the research community to create computational methods for the analysis of single-cell data, with over 600 tools at time of writing created for scRNA-seq data alone (Zappia et al 2018). However, tools specifically designed for the analysis of single-cell cancer genomic data are still in their infancy. Here, we outline future computational methods across four research domains (summarized in table 2) that we envisage as necessary to answer pressing questions about cancer initiation, progression, and resistance from single-cell data.

4. Uncovering clonal dynamics from single-cell genomic data

Interpreting mutational dynamics at the single-cell level remains one of the major challenges in single-cell genomics. While the robustness of single-cell copy number technologies such as direct library preparation (DLP, Zahn et al 2017) has been recently demonstrated and substantial progress has been made in calling CNVs from these data (Dong et al 2019, Wang et al 2019), little attention has been paid to the problem of constructing phylogenetic trees from single-cell CNV data despite the large number of such methods for point mutation data. Furthermore, systematic methods for cutting single-cell phylogenies to identify clones are lacking, with most studies to date employing phylogeny-naive approaches such as hierarchical clustering (Zahn et al 2017) or density-based clustering (Laks et al 2019). Existing methods developed for clone-level phylogeny construction from bulk genomic data (Malikic et al 2015) cannot currently be deployed on single-cell data due to computational time constraints. The lack of bespoke methods in this space is notable given the extremely large number of comparable methods for clustering scRNA-seq data, though we expect this gap to be filled as commercial platforms for single-cell CNV sequencing become widely available.

With an increasing number of multi-sample cancer datasets being generated, we expect that approaches to understand and reconstruct clonal dynamics from time series or spatially-sampled single-cell mutation data will be developed. Single-cell whole genome data is the ideal substrate for evaluating clonal fitness in the treatment-naive context and after chemotherapeutic intervention, either from longitudinally-collected samples or from cross-sectional data exploiting topological signatures of fitness in phylogenetic trees. Furthermore, the problem of reconstructing clonal migration histories is particularly important, with early studies demonstrating the timing of genomic mutations relative to clonal

Table 2. Future research domains and possible solutions for single-cell computational genomics.

| Research domain/modality | Computational tools necessary |
|--------------------------|-------------------------------|
| Low-depth single-cell whole genome sequencing | - Copy number aware clustering of scDNA-seq (clonal inference) - Phylogeny tree reconstruction from copy number calls |
| Tumour-microenvironment interactions | - Simulation-based or statistical modelling tools to assess impact of TME phenotypes and composition on clonal evolution |
| Single-cell biomarker approaches | - Quantification of proliferative and resistance potential of rare cell populations - Robust probabilistic machine learning models to predict outcomes from gene expression phenotypes and TME composition |
| Spatial tumour dynamics | - Inference of cell-cell signalling networks - Integration of spatially-resolved data with whole transcriptome scRNA-seq - Automated cell type inference - Reconstruction of 3D tumour architecture from 2D measurements |

include scATAC-seq for profiling chromatin accessibility (see Chen et al 2018 for a computational methods overview), scCHIP-seq for profiling various types of histone modifications (Grosselin et al 2019), single-cell bisulfite sequencing (Smallwood et al 2014) for 5-methylcytosine marks, and scAba-seq (Mooijman et al 2016) for 5-hydroxymethylcytosine modifications.

These new technologies represent a promising novel modality for tackling fundamental questions related to cancer heterogeneity and resistance. For example, in an application of high-throughput scCHIP-seq to breast cancer to breast cancer patient-derived xenografts, (Grosselin et al 2019) found a low-prevalence population of cells in a pre-treatment sensitive tumour that shared chromatin features with all post-treatment resistant cells, which was not observable at a bulk level.

However, the analysis of such data remains in its infancy, with few bespoke computational methods for the analysis of cancer single-cell epigenomic data, to our knowledge. Furthermore, it is not apparent that the development of such tools will be necessary if standard workflows for resolving cellular heterogeneity and identification of differential regions of epigenetic marks are sufficient to answer questions in a cancer-specific context.
invasion (Casasent et al 2018). Such methods operating at the single-cell level could leverage phylogenetic placement and would complement those for establishing clonal migration patterns from bulk DNA sequencing data such as MACHINA (El-Kebir et al 2018).

An important emerging area of technology development that will help uncover clonal dynamics concerns technologies that measure modalities in addition to the genome at the single-cell level. Early examples include G & T-seq (Macaulay et al 2015) for the combined measurement of the genome and transcriptome at the single-cell level along with more recent work such as Genotyping of Transcriptomes (Nam et al 2019). The ability to jointly measure the genome and transcriptome at single-cell resolution across many tumour cells would enable the refined tracking of expression changes linked to ongoing clonal evolution, tying possibly targetable phenotypes to waves of clonal expansion. However, to date few studies have applied such technologies at scale and additional clinical applications are as-yet unrealized.

5. Impact of the TME on tumour evolution and phenotypes

As computational tools mature to provide quantitative estimates of clonal fitness from single-cell genome sequencing technologies and similar tools emerge to quantify tumour phenotypic states and microenvironment composition from single-cell expression profiling, there is a major need for the development of methodologies to interrogate the interplay between these two important facets. For example, it is incompletely understood how the composition of the microenvironment shapes the clonal fitness landscape and preferentially allows for the growth of certain tumours. We envisage in the future that given sufficient data there will exist computational tools to make quantitative predictions of clonal fitness given clonal genotypes and possibly phenotypes, in response to perturbations to the local microenvironment composition and expression phenotype. Such models are crucial to enable an era of personalized chemotherapy, where therapeutic interventions will be actioned not only on the basis of (sub-)clonal genotypes, but also on possible interactions with the TME. These tools will be further enabled by advances in immune-related sequencing technologies, such as the ability to sequence T- and B-cell receptor sequences and genotype the HLA loci responsible for MHC class I antigen presentation. The combination of in situ sequencing chemistry with these technologies would allow for direct interrogation of the T- and B-cell clonotypes that are spatially adjacent and likely respond to individual tumour clones.

6. Machine learning and biomarker-based methods to guide therapy

Arguably one of the major goals of modern cancer research is ‘personalized chemotherapy’—the ability to tailor therapies based on patient-specific characteristics with the goal of achieving better outcomes than treatment with histotype-based standard-of-care. Single-cell technologies can enable this goal in two ways: (i) the identification of rare cell populations that are likely to form resistant clones and lead to patient relapse, and (ii) the association of cell type-specific phenotypic states (either in the tumour itself or the TME) that are associated with particular clinical courses and therapy responses.

Realizing these scenarios requires research discovery and clinical implementation phases with computational tools needed to support both. For rare cell population identification, multiple tools exist for scRNA-seq data (Wegmann et al 2019, Jindal et al 2018) but further interpretation of whether such rare populations have proliferative or chemoresistant potential will require the development of new methods that integrate multiple distinct modalities such as time series transcriptomic and genomic data. Translating these inferences to the clinical context will demand methods that can quickly and robustly identify such populations while accounting for multiple levels of technical variability including substantial inter-patient heterogeneity. For the discovery phase of the association of cell-type specific phenotypic states to outcomes, a current limiting factor is the availability of cohort-scale single-cell datasets with follow-up treatment and outcome data, on top of which existing tools such as CellAssign (Zhang et al 2019a) and edgeR (Robinson et al 2010) could be deployed. In the clinical implementation phase, the design of robust probabilistic machine learning models is necessary to predict outcomes with calibrated uncertainty, taking into consideration the ability to easily overfit to the training datasets given their high dimensional nature and relatively low number of samples.

7. Exploiting spatial data to model tissue dynamics

The advent of technologies to profile tumour cells while retaining information about their spatial origin has provided the opportunity to uncover mechanisms crucial to tumour initiation and progression, such as cell-to-cell signalling and tumour infiltration. However, the handful of papers in this field approach these computational analyses in an ad hoc manner, and thus we expect there to be a number of efforts to systematically address these questions. Firstly, methods are
needed to quantitatively infer cell-to-cell signalling networks from spatially resolved data; this may take the form of simulation-based inference using techniques such as Approximate Bayesian computation due to the fundamental difficulty of assessing whether a cell exists in a phenotypic state due to signalling from a proximal cell, or vice versa. Secondly, we expect the development of methods to integrate data from disaggregated whole-transcriptome technologies such as scRNA-seq with panel-based technologies such as IMC. Thirdly, there is a need for methods to automate the descriptions of spatially resolved datasets which can have cohorts of hundreds to thousands of patients, identifying regions of immune cell infiltration in tumours or regions of high proliferation in a systematic manner across patients. Fourthly, as the majority of genomic assays that retain spatial information operate in 2D only, we anticipate the development of methods to reconstruct 3D expression profiles through the integration of serial 2D ‘slices’. Finally, as the data produced by such technologies are shown to be predictive of patient outcomes, the need for machine learning models that the irregularly-sampled high-dimensional spatial data as input will become necessary.

8. Discussion

In this review, we have outlined the existing computational models designed for the analysis of single-cell genomic, transcriptomic, and proteomic data in the context of cancer research, and described four emerging domains in which the development of future computational tools is essential. However, there exist additional measurement modalities and analysis strategies that will be equally crucial to unlocking insights from single-cell cancer data.

A crucial area of computational methods development that permeates cancer research concerns the visualization of single-cell data. For example, E-scape (Smith et al 2017) contains a suite of methods for the visualization of single-cell mutation data, including Cellscape for visualizing single-cell copy number heatmaps and phylogenies, and TimeScape for visualizing clonal prevalence over time. Meanwhile, some existing tools have visualization capabilities built-in such as CaSPER (Serin-Harmanci et al 2020), while other methods such as Millefy (Ozaki et al 2020) will be exceptionally useful for assessing variant detection from scRNA-seq despite not being designed specifically in a cancer context.

Despite the huge technological progress and numerous discoveries enabled by single-cell technologies, the extent to which these technologies (both experimental and computational) will revolutionize cancer detection and therapeutics remains to be seen. With the exception of the ability to capture CTCs, the applications of single-cell technologies for cancer detection may be limited due to the starting requirement of surgically resected material. However, single-cell technologies may act in future as an effective biomarker for patient subtyping and therapy guidance, particularly based on the presence of rare cell populations that harbour chemoresistant phenotypes. The success of such an approach would require (i) systematic computational mapping of cell types and states that is easily standardizable across patients and technologies as opposed to the myriad of ad-hoc approaches currently applied, (ii) a deep understanding of cell-population-specific pathways in order to map therapeutic vulnerabilities, and (iii) large scale controlled clinical trials backed by generous funding given the expense of single-cell assays and the subsequent bioinformatics analysis.

A further pressing barrier to the widespread adoption of single-cell technologies in the clinic is the perturbing nature of current library preparation protocols. For example, most current studies require a dissociation step to create a single-cell suspension from the initial solid tumour, typically achieved via incubation with collagenase. However, existing dissociation protocols have a disruptive effect on the samples under study, preferentially releasing lymphocytes and inducing a stress response in epithelial tumour cells (van den Brink et al 2017, O’Flanagan et al 2019). Consequently, the computational interpretation of cell type composition and phenotype may be biased and additional technologies—either computational or experimental—to correct for such artifacts will be necessary to fully realize the value of single-cell technologies in the clinic.

Finally, the extent to which clinical outcomes can be predicted from modalities currently available remains to be seen. For example, multiple efforts exist to characterize which cancer clones are likely to metastasize, both on the basis of their genome and transcriptome. However, such methods will fail if there is not an underlying genomic or transcriptomic cause, such as inherent stochasticity or even e.g. clonal proximity to a nearby blood or lymphatic vessel. We believe that recent initiatives such as the Human Tumor Atlas Network (Rozenblatt-Rosen et al 2020) that promise to measure tumours in 3D space across multiple technologies will go a long way to answering such questions.

Acknowledgments

We thank Sally Millett for the creation of the visualizations in figure 1. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.
**ORCID iDs**

Allen W Zhang [https://orcid.org/0000-0002-7606-089X](https://orcid.org/0000-0002-7606-089X)

Kieran R Campbell [https://orcid.org/0000-0003-1981-5763](https://orcid.org/0000-0003-1981-5763)

**References**

Angelo M et al 2014 Multiplexed ion beam imaging of human breast tumors Nat. Med. 20 436–42

Aran D et al 2019 Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage Nat. Immunol. 20 163–72

Bell C C et al 2019 Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia Nat. Commun. 10 2723

Berg S et al 2019 Istilik: Interactive machine learning for (bio) image analysis Nat. Methods 16 1226–32

van den Brink S C, Sage F, Vérites A, Spanjaard B, Peterson-Maduro J, Baron C S, Robin C and van Oudenaarden A 2017 Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations Nat. Methods 14 953–6

Campbell K R et al 2019 Clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers Genome Biol. 20 54

Casaasent A K et al 2018 Multiclonal invasion in breast tumors identified by topographic single cell sequencing Cell 172 205–217.e12

Chen K-Y et al 2018 Single-cell transcriptomics reveals heterogeneity and drug response of human colorectal cancer organoids 2018 40th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) (Piscataway, NJ: IEEE) pp 2378–81

Chevrier S, Crowell H L, Zanotelli V R T, Engler S, Robin et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing Nat. Methods 16 207–15

Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T and Navin N E 2018 Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing Cell 173 879–893.e13

Kiselev V Y et al 2017 SC3: consensus clustering of single-cell RNA-seq data Nat. Method 14 483–6

Kiselev V Y, Yiu A and Hemberg M 2018 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Laks E et al 2019 Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing Cell 179 1207–21.e22

Levine J H et al 2015 Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis Cell 162 184–97

Li H et al 2017 Reference component analysis of single-cell transcripts elucidates cellular heterogeneity in human colorectal tumors Nat. Genet. 49 708–18

Maasoka J et al 2018 Charting tissue expression anatomy by spatial transcriptome deconvolution (bioRxiv:362624)

Macalay J C et al 2019 GtF-seq: parallel sequencing of single-cell genomes and transcriptomes Nat. Methods 12 519–22

Mallick S, McPherson A W, Dommez N and Sahinalp C S 2015 Clonality inference in multiple tumor samples using phylogeny Bioinformatics 31 1349–56

McCarthy D J et al 2020 Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes Nat. Methods 17 414–21

McCarty DJ, Campbell K R, Lun A T L and Wills Q F 2017 Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R Bioinformatics 33 1179–86

McGranahan N et al 2017 Allele-specific HLA loss and immune escape in lung cancer evolution Cell 171 1259–71.e11

McQuin C et al 2018 CellProfiler 3.0: next-generation image processing for biology PLoS Biol. 16 e2009978

Moncada R et al 2020 Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas Nat. Biotechnol. 38 333–42

Mooijman D et al 2016 Single-cell ShmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction Nat. Biotechnol. 34 852–6

Müller S, Cho A, Liu S J, Lim D A and Diaz A 2018 CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones Bioinformatics 34 3217–9

Myers M A, Satas G and Raphael B J 2019 CALDER: inferring phylogenetic trees from longitudinal tumor samples Cell 8 514–22.e5

Greaves M and Maley C C 2012 Clonal evolution in cancer Nature 481 306–13

Grosselin K et al 2019 High-throughput single-cell ChiP-seq identifies heterogeneity of chromatin states in breast cancer Nat. Genet. 51 1090–6

Ha G et al 2012 Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer Genome Res. 22 1995–2007

Hellani A, Coskun S, Benkhalifa M, Tbakhi A, Sakati N, Al-Odaib A and Ozand P 2004 Multiple displacement amplification on single cell and possible PGD applications MHR: Basic Science Reprod. Med. 10 847–52

Illic T et al 2016 Classification of low quality cells from single-cell RNA-seq data Genome Biol. 17 29

Jackson H W et al 2020 The single-cell pathology landscape of breast cancer Nature 578 615–20

Jahn K, Kuipers J and Beerwenkinkel N 2016 Tree inference for single-cell data Genome Biol. 17 86

Jindal A et al 2018 Discovery of rare cells from voluminous single cell expression data Nat. Commun. 9 4719

Keren L et al 2018 A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging Cell 174 1373–87.e19

Kiselev V Y et al 2017 SC3: consensus clustering of single-cell RNA-seq data Nat. Method 14 483–6

Kiselev V Y, Yiu A and Hemberg M 2018 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)

Kiselev V Y et al 2020 scmap: Projection of single-cell RNA-seq data across data sets Nat. Methods 15 359–62

Kiselev V Y et al 2019 SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing (bioRxiv:355807)
Nam A S et al 2019 Somatic mutations and cell identity linked by genotyping of transcriptomes Nature 571 355–60
Nassar D and Blanpain C 2016 Cancer stem cells: basic concepts and therapeutic implications Annu. Rev. Pathol. Mech. Dis. 11 47–76
Nowicka M et al 2017 CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets FI000res. 6 748
O’Flanagan C H et al 2019 Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses Genome Biol. 20 210
Osorio D and Cai J 2020 Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA sequencing data quality control (bioRxiv: 2020.02.20.958793)
Ozaki H et al 2020 Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets BMC Genom. 21 177
Patel A P et al 2014 Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma Science 344 1396–401
Pierson E and Yao C 2015 ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis Genome Biol. 16 241
Pliner H A, Shendure J and Trapnell C 2019 Supervised classification enables rapid annotation of cell atlases Nat. Methods 16 983–6
Qi P, Simonds E F, Bendall S C, Gibbs K D, Bruggner R V, Linderman M D, Sachs K, Nolan G P and Pe’er D 2011 Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE Nat. Biotechnol. 29 886–91
Risso D et al 2018 A general and flexible method for signal extraction from single-cell RNA-seq data Nat. Commun. 9 284
Robinson M D, McCarthy D J and Smyth G K 2010 edgeR: A bioconductor package for differential expression analysis of digital gene expression data Bioinformatics 26 139–40
Ross E M and Markowetz F 2016 OncoNEM: inferring tumor evolution from single-cell sequencing data Genome Biol. 17 69
Rostom R, Svensson V, Teichmann S A and Kar G 2017 Computational approaches for interpreting scRNA-seq data FEBS Lett. 591 2213–25
Roth A et al 2016 Clonal genotype and population structure inference from single-cell tumor sequencing Nat. Methods 13 573–6
Rozenblatt-Rosen O et al 2020 The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution Cell 181 236–49
Satas G et al 2019 Single-cell tumor phylogeny inference with copy-number constrained mutation losses (bioRxiv:840355)
Serin-Harmanci A, Harmanci A O and Zhou A X 2020 CaSpER identifies and visualizes CNN events by integrative analysis of single-cell or bulk RNA-seq data Nat. Commun. 11 89
Shah S P et al 2012 The clonal and mutational evolution spectrum of primary triple-negative breast cancers Nature 486 395–9
Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H and Takeyama H 2015 Cancer-associated fibroblasts: their characteristics and their roles in tumor growth Cancers 7 2443–58
Shih A J, Menzin A, Whyte J, Lovechko J, Liew A, Khalili H, Bhuia T, Gregersen P K and Lee A T 2018 Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq Plos One 13 e0206785
Singer J et al 2018 Single-cell mutation identification via phylogene tic inference Nat. Commun. 9 5144
Smallwood S A et al 2014 Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity Nat. Methods 11 817–20
Smith M A, Nielsen C B, Chan F C, McPherson A, Roth A, Farahani H, Machev D, Steif A and Shah S P 2017 E-scape: interactive visualization of single-cell phylogenetics and cancer evolution Nat. Methods 14 549–50
Szécsi M H and Nolan G P 2016 Mass cytometry: single cells, many features Cell 165 780–91
Teh C E et al 2020 Deep profiling of apoptotic pathways with mass cytometry identifies a synergistic drug combination for killing myeloma cells Cell Death Differ. 27 2217–33
Tirosh I et al 2016 Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma Nature 539 309–13
Van Velden D A et al 2016 Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments PLoS Comput. Biol. 12 e1005177
Venteicher A S et al 2017 Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq Science 355 eaai8478
Vickovic S et al 2019 High-definition spatial transcriptomics for in situ tissue profiling Nat. Methods 16 987–90
Wagner J et al 2019 A single-cell atlas of the tumor and immune ecosystem of human breast cancer Cell 177 1330–45.e18
Wang R, Lin Y D and Jiang Y 2019 SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing (bioRxiv:594267)
Wegmann R et al 2019 CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data Genome Biol. 20 142
Weinstein J N et al 2013 The Cancer Genome Atlas Pan-Cancer analysis project Nat. Genet. 45 1113–20
Wherry E J 2011 T cell exhaustion Nat. Immunol. 12 492–9
Yu C et al 2014 Discovery of bcl-2oncoprotein A and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing Cell Res. 24 701–12
Zaccaria S and Raphael B J 2020 Characterizing the allelic and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL Nat. Biotechnol. (https://doi.org/10.1038/s41587-020-0661-6)
Zafar H, Wang Y, Nakhe L, Navin N and Chen K 2016 Monovar: single-nucleotide variant detection in single cells Nat. Methods 13 505–7
Zafar H et al 2017 SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models Genome Biology 18 178
Zahn H, Steif A, Laks E, Eirew P, VanNordenhe M, Shah S P, Apricio S and Hansen C I 2017 Scalable whole-genome single-cell library preparation without preamplification Nat. Methods 14 167–73
Zappia L, Phipson B and Oshlack A 2018 Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database PLoS Comput. Biol. 14 e1006245
Zhang A W et al 2018 Interfaces of malignant and immunologic clonal dynamics in ovarian cancer Cell 173 1755–69.e22
Zhang A W et al 2019a Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling Nat. Methods 16 1407–15
Zhang X et al 2019b CellMarker: a manually curated resource of cell markers in human and mouse Nucleic Acids Res. 47 D721–8
Zhang Z et al 2019c SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples Genes 10 531
Zheng G Y et al 2017 Massively parallel digital transcriptional profiling of single cells Nat. Commun. 8 14049