Single cell genomics as a transformative approach for aquaculture research and innovation

Rose Ruiz Daniels | Richard S. Taylor | Diego Robledo | Daniel J. Macqueen

Abstract

Single cell genomics encompasses a suite of rapidly maturing technologies that measure the molecular profiles of individual cells within target samples. These approaches provide a large up-step in biological information compared to long-established ‘bulk’ methods that profile the average molecular profiles of all cells in a sample, and have led to transformative advances in understanding of cellular biology, particularly in humans and model organisms. The application of single cell genomics is fast expanding to non-model taxa, including aquaculture species, where numerous research applications are underway with many more envisaged. In this review, we highlight the potential transformative applications of single cell genomics in aquaculture research, considering barriers and potential solutions to the broad uptake of these technologies. Focusing on single cell transcriptomics, we outline considerations for experimental design, including the essential requirement to obtain high quality cells/nuclei for sequencing in ectothermic aquatic species. We further outline data analysis and bioinformatics considerations, tailored to studies with the under-characterized genomes of aquaculture species, where our knowledge of cellular heterogeneity and cell marker genes is immature. Overall, this review offers a useful source of knowledge for researchers aiming to apply single cell genomics to address biological challenges faced by the global aquaculture sector though an improved understanding of cell biology.

KEYWORDS
aquaculture, bioinformatic pipelines, cell isolation, nuclei isolation, single cell genomics, transformative applications

1 | INTRODUCTION

The field of genomics is rapidly transitioning from a state where bulk samples are predominantly studied, providing read-outs averaged across all cells, to a position where the molecular profiles of individual cells can be reliably profiled. As most phenotypes result from the actions and interactions of specific cell types, generating cell-specific rather than tissue-averaged molecular data creates a large up-step in information on the mechanisms underpinning trait expression. Such knowledge will have great value in research applications addressing biological challenges faced by the aquaculture sector, representing ongoing barriers to the sustainable production of many species, including, for example, the threat posed by disease outbreaks caused by diverse parasites and pathogens.

Single cell genomics involves the molecular analysis of individual cells, typically using high-throughput sequencing technologies. The transformative impact of such methods on our understanding of...
cellular and developmental biology, immunology and disease biology in humans and model organisms cannot be overstated. The most common application is single cell transcriptomics, though methods for profiling epigenomic features at single cell resolution, inclusive of methylation (e.g., Ref. 5), chromatin accessibility (e.g., Ref. 6) and chromosome conformation, are maturing rapidly, as is single cell proteomics. In model organisms, it is now common to combine different single cell technologies to generate ‘multi-omic’ profiles from the same cells. The scale of recent knowledge advancement in this field is perhaps best illustrated by the Human Cell Atlas initiative, which aims to deliver comprehensive molecular maps of all human cell types.

This review concerns the uptake and applications of single cell genomics in aquaculture research, a relatively new field that is expanding rapidly. Our primary focus is on single cell transcriptomics (Section 2), performed using single cell (sc) or single nuclei (sn) RNA-Seq (scRNA-seq and snRNA-seq, respectively) on various platforms (reviewed in Ref. 15; also see Section 2) that have increased dramatically in throughput over recent years, such that thousands to hundreds of thousands of cells are being routinely profiled in single studies. Compared to bulk transcriptomics using RNA-seq, which is already widely utilized in aquaculture research (reviewed in Ref. 17), sc/snRNA-seq is a ‘game changer’ due to its ability to identify cell types and their heterogeneity, along with

### TABLE 1 Summary of single cell transcriptomic studies in aquaculture species including major knowledge advances

<table>
<thead>
<tr>
<th>Species</th>
<th>Sample type</th>
<th>Approach</th>
<th>Summary and knowledge advances</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic salmon</td>
<td>Liver</td>
<td>snRNA-seq 10x platform</td>
<td>Liver cell atlas capturing host response to bacterial infection; revealed unique hepatocyte states associated with infection and resolved major immune cell types</td>
<td>24</td>
</tr>
<tr>
<td>Gill</td>
<td>Gill cell</td>
<td>snRNA-seq 10x platform</td>
<td>Gill cell atlas; revealed known and novel cell types and cellular remodelling during smoltification, including evidence for repression of immune system</td>
<td>23</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>Blood</td>
<td>scRNA-seq 10x platform</td>
<td>Atlas of sorted B cell heterogeneity; improved understanding of B cell heterogeneity, including new marker genes for distinct B cell subsets and states</td>
<td>26</td>
</tr>
<tr>
<td>Blood</td>
<td>scRNA-seq 10x platform</td>
<td>Analysis of antibody gene diversity in sorted B cells; supporting hypothesis that B cells produce antibodies with distinct specificities</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Turbot</td>
<td>Multi-tissue</td>
<td>scRNA-seq 10x platform</td>
<td>Immune cell atlas; identified sixteen immune cell sub-types, including major classes of lymphocytes and myeloid cells</td>
<td>64</td>
</tr>
<tr>
<td>Multi-tissue</td>
<td>scRNA-seq 10x platform</td>
<td>Immune cell atlas; identified fourteen immune cell sub-types, including major classes of lymphocytes and myeloid cells; providing evidence that neutrophils are key effectors of trained innate immunity</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>Orange spotted grouper</td>
<td>Midbrain</td>
<td>scRNA-seq 10x platform</td>
<td>Midbrain cell atlas capturing response to RGNNV infection; described thirty-five cell types, both neuronal and non-neuronal, with evidence that microglia transform into macrophages upon infection</td>
<td>27</td>
</tr>
<tr>
<td>Gonads</td>
<td>scRNA-seq 10x platform</td>
<td>Cell atlas of adult male gonad; identified 10 cell types, including germ and somatic cells, with a panel of associated cell-specific marker genes</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Nile tilapia</td>
<td>Head kidney</td>
<td>snRNA-seq 10x platform</td>
<td>Atlas of immune cells; revealing major populations of lymphocytes and myeloid cells, including diverse B and T cell subsets</td>
<td>29</td>
</tr>
<tr>
<td>Head kidney</td>
<td>scRNA-seq 10x platform</td>
<td>Atlas of immune cells after viral stimulation; revealed major populations of lymphocytes and myeloid cells, and subsets of non-specific cytotoxic cells (akin to natural killer cells)</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Asian sea bass</td>
<td>Ovaries</td>
<td>scRNA-seq Drop-seq</td>
<td>Adult cell atlas; identified multiple germ cell and somatic cell types and revealed conserved and divergent ovary development mechanisms via comparative single cell transcriptomics</td>
<td>44</td>
</tr>
<tr>
<td>Kuruma shrimp</td>
<td>Haemolymph</td>
<td>scRNA-seq Drop-seq</td>
<td>Atlas of cell types; described six haemocyte subtypes and their marker genes, and evidence for factors driving haemocyte differentiation</td>
<td>31</td>
</tr>
<tr>
<td>White shrimp</td>
<td>Haemolymph</td>
<td>scRNA-seq 10x platform</td>
<td>Atlas of cell types; described haemocytes, monocytic haemocytes and granulocytes, including macrophage-like haemocytes</td>
<td>65</td>
</tr>
<tr>
<td>Black tiger shrimp</td>
<td>Hepatopancreas</td>
<td>scRNA-seq 10x platform</td>
<td>Atlases of cell types; identified seven haemocyte sub-types including candidate stem cells and seven cell types in hepatopancreas, along with cell-resolved responses to ammonia nitrogen stress</td>
<td>188</td>
</tr>
<tr>
<td>Hong Kong Oyster</td>
<td>Haemolymph</td>
<td>scRNA-seq 10x platform</td>
<td>Atlas of haemocytes; identified thirteen populations including granulocytes, semi-granulocytes, and hyalinocytes, revealing candidate transcription factors driving granulocyte lineage differentiation.</td>
<td>189</td>
</tr>
<tr>
<td>Haemolymph</td>
<td>scRNA-seq 10x platform</td>
<td>Investigated cellular heterogeneity in response to copper exposure, revealing cell-specific sensitivities to and markers for copper exposure</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>
resolving cell-specific gene expression responses to environmental perturbation. Such data are especially informative in species where knowledge of cellular biology is limited, that is, most aquaculture species.

While confined to a few labs in its early applications, extensive development of methods and commercial platforms has brought single cell technologies within the reach of many researchers. A portfolio of single cell transcriptomic studies have recently emerged in nonmodel aquatic taxa, including Atlantic cod, Mexican cavefish, corals, and diverse aquaculture species, including Atlantic salmon, rainbow trout, orange-spotted grouper, Nile tilapia, kuruma shrimp, and Pacific oyster, among others (Table 1). This growing body of work immediately demonstrates the potential of single cell transcriptomics to transform our understanding of cellular biology in aquaculture species (Section 3), in settings that can be exploited to address diverse sustainability challenges, including the pressing need to manage disease outbreaks.

However, single cell genomics is more challenging than equivalent bulk methods in multiple respects, presenting a barrier to broad-scale uptake in aquaculture research. This includes the lab work, requiring high-quality cells or nuclei and specialist library preparation methods, which brings several considerations to achieve optimal results (e.g., Refs. 35,36) (see Section 4). Optimization efforts in single cell genomics have concentrated on mammals, which have major differences in biology with aquatic species. Analysis of single cell data is more complicated than bulk genomics due to its higher dimensionality. Multiple decisions are involved at many steps of data analysis, with a plethora of tools available, but without tried-and-tested standards that fit all studies, species and problems. This challenge is compounded in aquaculture species as analysis pipelines are optimized for model species (see Section 5). Furthermore, while reference genomes are now available for many aquaculture species, they remain comparatively poorly annotated, which is one of several challenges faced when the aim is to transfer knowledge on cell biology across species—a standard practice to classify cell types on the basis of known marker genes from well-characterized organisms.

This review starts with a brief overview of single cell transcriptomics (Section 2), before outlining the potential applications and impacts that single cell genomics can bring to aquaculture research (Section 3). Subsequently, we explore barriers to achieving such impacts, along with potential solutions and considerations when designing/executing experiments relevant to the wet-lab work (Section 4) and also concerning the downstream data analysis and interpretation (Section 5). The overall aim is to provide a state-of-the-art review on single cell transcriptomics relevant to aquaculture researchers, while also offering recommendations and tips for those aiming to uptake single cell methods in their research.

2 WHAT IS SINGLE CELL/NUCLEI TRANSCRIPTOMICS?

In a nutshell, single cell transcriptomics involves the global profiling of gene expression in individual cells or nuclei. It is not our aim to describe the development of this field (see reviews by Refs. 1,4), nor do we comprehensively review the various platforms currently on offer (see Ref. 40). Instead, this section offers a short primer on single cell transcriptomics to provide context for the remainder of the review.

Currently the most popular high-throughput single cell transcriptomics methods are droplet-based, with all studies published to date in aquaculture species using such strategies (Table 1). These approaches employ microfluidic capture of cells/nuclei inside microdroplets that contain an mRNA capture bead that includes a unique barcode, typically next to a unique molecular identifier (UMI). Subsequently, reverse transcription is used to generate a cDNA product with the expressed transcript linked to the barcode and UMI. In downstream analysis, the cell barcode retains the identity of the captured cells or nuclei, while the UMI ensures only unique transcripts are quantified. The resulting cDNA is used to make a library for sequencing, typically on a high-throughput short-read platform, which may be indexed to distinguish different samples. An important consideration at this stage is sequencing depth per cell/nuclei, a product of the (per sample) output of sequencing divided by the number of cells or nuclei captured, with 25,000–100,000 reads per cell/nucleus offering sufficient depth for most applications.

The most widely used single cell transcriptomics platform is the 10x Genomics Chromium, which is user-friendly with demonstrated capability to detect a high number of transcripts per cell in diverse taxa and sample types. The popularity of this platform extends to most studies performed with aquatic species to date (e.g., Table 1). However, it is also the most expensive droplet based approach on the market, with other available platforms being more affordable, including DropSeq (commercialized by Dolomite Bio) and InDrop. Studies have validated DropSeq in several commercially important aquatic species. There also exist non-droplet based approaches including Smart-seq2, where single cells are separated with a micro-capillary pipette or via FACS and then individually sequenced, along with the microplate-based method SPLIT-seq, where cells or nuclei are uniquely labelled through multiple barcoding rounds. This strategy has been commercialised by Parse Biosciences, with the advantage that no platform or FACS sorting is required.

3 APPLICATIONS OF SINGLE CELL GENOMICS IN AQUACULTURE RESEARCH

The majority of knowledge on animal cell biology derives from a small group of well-characterized organisms that benefit from advanced research ‘toolboxs’, allowing known cell types to be routinely isolated and manipulated (e.g., Ref. 47). Farmed fish and shellfish, by contrast, are a highly diverse group of >500 species presenting enormous variation in cellular biology and tissue organization, yet have limited species-specific research tools available to target or enrich particular cell types (e.g., monoclonal antibodies and marker genes for well-characterized cells). Consequently, our understanding of cell biology in aquaculture species remains in its infancy compared to the most characterized animal species. The uptake of single cell genomics, which allows for the unbiased high-throughput characterization of single cells,
offers an unprecedented and immediate opportunity to fast-track our understanding and exploitation of cell biology across the great diversity of aquaculture species (Figure 1; Table 1). This section is not intended to be exhaustive of all applications of single cell transcriptomics in aquaculture research and innovation expected to arise in the coming years. Instead, our goal is to provide some illustrative directions in which single cell transcriptomics will advance different fields, which can be built up in the future.

### 3.1 Up-step in fundamental cellular biology and molecular toolboxes

In aquaculture species, major knowledge advancement is possible using single cell transcriptomics for the identification of cell types and their expression profiles/responses (Table 1). Such work can reveal which cell types and their sub-populations are conserved with a reference species (e.g., model organism), generate evidence for novel cell types, and resolve which cells are the likely drivers for a shift in phenotypic status. Single cell transcriptomics creates an abundance of novel marker genes for cell types and their sub-populations, which can be used to add cell-specific resolution to bulk gene expression studies (i.e., using existing or new datasets), while enhancing the molecular toolbox for aquaculture research. This can be as straightforward as identifying candidate gene markers for a cell population of interest, and quantifying the expression of such genes in bulk samples using targeted assays (e.g., quantitative PCR), providing a routine and cost-effective read-out on a cell population's phenotypic status. Cell-specific marker genes can be taken forward as targets for in situ expression analyses, to validate and provide spatial resolution to their expression (or co-expression with other marker genes) at the tissue level of organization (e.g., Refs. 49,50). A subset of genes will represent targets for which to develop monoclonal antibodies targeting candidate cell surface markers, which can subsequently be used to sort or quantify cells using fluorescence-activated cell sorting (FACS); particularly useful in immunology (Section 3.2) (reviewed in Ref. 51). Finally, data generated by single cell transcriptomics can be used to deconvolute cell-specific expression signals in bulk studies (e.g., Refs., 52–54). This approach may provide the benefits of single cell resolution in larger functional genomics studies, adding significant value without adding additional sequencing costs. For example, a bulk RNA-seq experiment with a complex design could be performed, which is cost-effective per sample, and followed by deconvolution methods that leverage existing single cell transcriptomic data from the same tissue type, with the only additional costs being for the data analysis.

Considering the current status for model species alongside similar aspirations for livestock (e.g., under the FAANG initiative55), many commercially important aquaculture species will likely soon (i.e., in the next 5 years) benefit from cell atlases spanning different tissues relevant to production and health, mirroring what has been done, albeit at a smaller scale, in mice,56 zebrafish,57 humans,58 and Caenorhabditis elegans.59 Such efforts have been merged and integrated into curated resources available for wider exploitation (e.g., Ref. 60), providing a template for aquaculture research communities in the future.

### 3.2 Immunology and vaccinology

The outcome of infectious disease challenge, and the success of vaccination programmes for species with an adaptive immune system, is largely a product of immune cell actions and the responses of these cells to pathogen signals, involving immune cell activation and differentiation, in addition to cell-to-cell interactions. Single cell transcriptomics provides a wealth of information on such processes, enabling a new frontier of investigations into the immune system of aquaculture species.
Unsurprisingly, immunology has been a primary focus in single cell transcriptomics studies published to date in aquaculture species (e.g., Refs. 17,23–26,30,31,63–64; Table 1). Such work has identified novel heterogeneity in the haemocytes (immune cells responsible for phagocytosis in the haemolymph) of kuruma shrimp,31 white shrimp,66 and oysters.64 For example, a recent snRNA-seq study in white shrimp provided evidence for phagocytic haemocytes sharing marker genes with vertebrate macrophages, offering novel future avenues to exploit the basis for cellular immunity in crustaceans.65 The work done to date has also demonstrated that farmed teleosts from different families (Salanidae, Cichlidae and Serranidae) possess diverse immune cell types with identifiable subsets, including T and B lymphocytes, granulocytes, macrophages and dendritic cells (Table 1). A scRNA-seq study focussing on circulating B cells in rainbow trout provided evidence for extensive B cell heterogeneity, likely representing distinct maturation and differentiation states, while also noting substantial differences in B cell marker genes with mammals.25 Two recent multi-organ scRNA-seq studies in turbot provided a major step forward in immunology for this species, demonstrating extensive diversity in multiple immune cell subtypes, along with associated marker genes.64,66 This work evidenced the complex role that T cell heterogeneity plays in the response of turbot to bacterial infection,64 alongside evidence that neutrophils play a central role in turbot trained immunity,66 a process where the innate immune system is more effective in responding to a pathogen due to previous exposure to immunological stimuli. This latter finding could support the design of approaches that stimulate the innate immune system to increase disease resistance independent of vaccination.

Our own study of Atlantic salmon liver used snRNA-seq to uncover the crucial role played by hepatocyte state in the early immune response to bacterial infection, supported by cell-specific responses of hepatic immune cell sub-populations.24 We identified a dominant population of hepatocytes that dramatically remodelled its transcriptome following infection - repressing metabolic and anabolic pathways, while activating the host defence response and up-regulating key genes controlling protein synthesis and secretion, presumed to support the translation and secretion of high concentrations of acute phase proteins into circulation.24 An snRNA-seq study of orange spotted grouper brain following challenge with nervous necrosis virus (causative agent of viral nervous necrosis in many marine teleosts), revealed heterogeneity in brain macrophages, and described putative macrophage differentiation pathways supporting antiviral responses.27 This study employed a bioinformatic tool called Monocle, which aims to identify how far a cell-type has transitioned along a developmental or differentiation state.57,68 Monocle is one of several so-called trajectory inference methods,69 which may have applicability to identify pathways of immune cell activation and differentiation in aquaculture species. For instance, a recent scRNA-seq study of Atlantic cod spleen used an alternative trajectory inference method to reveal a potential B cell differentiation pathway leading to antibody-producing plasma cells.70

Single cell transcriptomics has great potential to improve our understanding of vaccine responses in finfish, and for identifying novel correlates of protection that may expedite tests of vaccine efficacy. This represents an important opportunity, considering: (i) that the cellular basis for immunological memory in fishes remains poorly defined,70,71 (ii) that reliable correlates of protective immunity are yet to be established for many aquaculture vaccines72 and (iii) the pressing need to reduce the number of fish used in vaccine testing.73 Single cell work in mammals (reviewed in Ref. 62) offers a useful direction of travel for farmed finfish. A recent study focussed on vaccination responses to dengue virus (DENV), known, like many viruses, to depend on T cell reactions.74 The authors identified a novel population of CD8+ T cells activated in response to vaccination with high memory effector potential that endured for 4 months post-vaccination.74 Another single cell transcriptomic study provided evidence for individual variation in vaccination response to hepatitis B virus within a human cohort, which correlated with the proportion of two rare dendritic cell populations showing distinct and highly specific marker genes (NDRG2 and CDKN1). The authors showed it was possible to identify these dendritic cell subtypes by quantitative PCR of NDRG2 and CDKN1, providing avenues to predict vaccine responsiveness prior to vaccination.75 Such studies demonstrate promise not simply to identify cellular mechanisms leading to vaccine protection, but also to identify marker genes for cell types that correlate with variation in vaccine protection outcomes across individuals, which may be present either before, or early post-vaccination, and can potentially be measured cheaply.

### 3.3 | Host-pathogen interactions

Another emerging single cell approach with potential applications in aquaculture research involves the profiling of host-pathogen interactions. Such methods apply sc/snRNA-seq to samples including both a host species and an infecting pathogen or parasite.76,77 The bulk equivalent, often called dual-RNA-seq, has been used to investigate problematic host-pathogen interactions in aquaculture. This includes the joint profiling of transcriptomic responses of Atlantic salmon tissues with parasites and pathogens during infection scenarios, including salmon louse (Lepeophtheirus salmonis),78 Neoparamoeba perurans (causative of amoebic gill disease)79 and the intracellular bacterium Piscirickettsia salmonis (causative of piscirickettsiosis).80 These bulk studies have revealed genes potentially involved in host resistance, and candidate mechanisms by which parasites and pathogens circumvent host defence or otherwise interact with the host during infection. However, dual-RNA-seq methods cannot directly inform on cell types involved in host-parasite interactions.

Single cell Dual-seq (scDual-seq) aims to directly measure cell-specific responses and cell-to-cell interactions in samples containing both host and pathogens. For intracellular pathogens, such insights extend to distinguishing infected from uninfected host cells, which has proven fruitful in studies of human pathogens.76,81 A recent study...
developed a bioinformatics approach to capture pathogenic viruses in infected host cells using scRNA-seq, allowing the immune responses of infected and uninfected bystander cells to be distinguished. A similar strategy was used in an Atlantic salmon head kidney cell line to study the transcriptome of cells infected with infectious salmon anemia virus, and compare their responses to bystander cells, providing novel insights into the interaction between the virus and host cells. As intracellular bacterial and viral infections pose a ubiquitous challenge in aquaculture (e.g., Refs. 72,84), an improved understanding of which cell types are infected, along with individual variation in cell-specific responses to infection, will have value when designing vaccines that aim to elicit cell-mediated immunity, but also for elucidating cellular mechanisms underlying the genetic basis for disease resistance, for example, targets for viral entry into host cells. However, it must be noted that in some settings, co-profiling of host and pathogen RNA brings challenges using bulk samples, and achieving effective scDual-seq pipelines will be even more difficult. 77

3.4 Genome editing

The discovery of CRISPR-Cas systems has greatly facilitated the field of genome editing, revolutionising practically all fields of biology. The ability to modify the genome of aquaculture species has attracted great interest from both researchers and industry, and CRISPR/Cas9 genome editing has already been applied to target traits in farmed finfish species including Atlantic salmon, Nile tilapia, Channel catfish, and various carps (reviewed in Refs. 85–88), in addition to farmed shrimp 89 and oyster 90 species. Legislation is rapidly evolving and genome editing seems to be gaining traction as a promising method to improve aquaculture sustainability and animal welfare.

The main biological challenge limiting applications of genome editing to improve aquaculture stocks is the identification of appropriate targets. Recently, the combination of genome editing and single-cell transcriptomics has enabled the study of candidate gene function, even at genome-wide resolution. In CRISPR screens, numerous genes are knocked-out simultaneously in vitro, with most cells being edited for a single gene. If the appropriate construct has been used for editing, single-cell transcriptomics can be used to simultaneously identify the guide RNA and determine the impact of the knock-out of that gene on the cell transcriptome. This is commonly known as a perturbation screen, and has, for example, been used recently to knock-out all expressed human genes (in cell lines) simultaneously to uncover the function of many uncharacterized genes on the basis of expression phenotypes in the edited cells. 91 Several approaches including Perturb-seq 92, CROP-seq 93, and CRISP-seq 94 rely on the same principle, that is, the identifying the RNA guide that edited each cell in parallel to measuring that cell’s transcriptome. In addition to loss of function screens, CRISPR activation, allowing the selective up-regulation of genes, was recently coupled to single cell transcriptomics in mouse embryonic stem cells, revealing key genes involved in transcriptional regulation. 95 Perturbation screens have huge potential to improve our understanding of gene functions in aquaculture species, which currently rely heavily on extrapolations from model species. Nonetheless, these approaches are limited by the relevance of the in vitro model of choice for the specific trait of interest, and in this sense there is an acute need for the development of novel cell lines in aquaculture species.

As a reverse strategy, it is also possible to interrogate the impact of targeted genome edits for candidate genes by applying single cell transcriptomics. The combination of in vitro perturbation screens and in vivo characterisation of gene function using the above highlighted approaches offers a powerful new toolbox to identify target genes for the genetic improvement of aquaculture animals, prioritise causative genetic variants in regions of the genome explaining trait variation, and also to validate the potential impact of off-target edits. Finally, single-cell technologies may also help us better understand and tackle mosaicism, a frequent phenomenon where edited animals are a mixture of edited and non-edited cells, representing a well-known issue in aquaculture species (reviewed in Ref. 85). It remains unclear whether mosaicism is stochastic in nature or is biased towards certain cell lineages. Single-cell sequencing may help us answer this question, and possibly lead to a better understanding of the molecular pathways underlying these potential biases, a necessary step towards improving the efficiency of in vivo editing in aquaculture species.

3.5 Sex and reproductive biology

Aquaculture is a relatively young industry and encompasses hundreds of unique species at different stages of domestication. One of the main challenges during the domestication of aquaculture species is achieving reproduction in captivity, an issue that affects even long-farmed species such as Senegalese sole, forcing the industry to rely on wild broodstock and curtailing scope for selective breeding. 96 The lack of reproduction in captivity can have different underlying causes, ranging from impaired gonad maturation to ineffective or inexistenct courtship. In finfishes, these processes depend on complex signalling along the hypothalamic–pituitary–gonadal axis, with specific cell-types secreting different sex hormones that control sex differentiation and reproduction. 99 The hormonal cascades involved in regulating reproductive processes in the many shellfish lineages used in aquaculture are equally complex and highly diverse (e.g., Refs. 100,101).

Bulk transcriptomics lacks the resolution to detect subtle, cell-type specific changes that may cause disruptions to reproduction in captivity. 102,103 In this sense, single-cell technologies can help dissect the complex hormonal systems controlling reproduction at higher resolution, as done recently in the model teleost medaka, enabling the characterization of reproductive disruptions by comparing wild versus F1 individuals. Such information offers a strong base to tackle reproductive problems faced by aquaculture species. Simultaneously, understanding how reproductive cells integrate and process hormonal signals during sex differentiation will also shed light into sexual dimorphism, commonly affecting traits of commercial relevance in aquaculture species, such as growth. 105 In this area, single cell transcriptomics has already provided insights into mechanisms...
underlying the expression of sexually dimorphic traits in mouse 44,106 and zig-zag eel. 107

Single-cell technologies further have great potential to refine our understanding of sex determination. Aquatic species have extremely labile and consequently diverse sex determination systems. For instance, in fish, the current model of sex determination suggests a network of different interacting genetic and environmental factors, where small changes can tip the scale towards males or females, providing multiple opportunities for novel sex determination mechanisms to evolve (e.g., Ref. 108). The development of germ cells is intimately linked to sex determination in many species, 109 and single-cell transcriptomics can improve our understanding of this process during gonad development, identifying factors underpinning germ cell proliferation and the underlying genetic networks, as done recently in mammals (e.g., Refs. 110,111), and avians. 112,113 For example, a recent study in zebra finch (Taeniopygia guttata) discovered three primordial germ cell sub-types, representing the first evidence of heterogeneity in this cell type. 113

With respect to work performed to date in aquaculture species, two single cell transcriptomics studies have provided insights into both germ and somatic cells in gonads. The first reported a comprehensive scRNA-seq atlas of testis cells in orange-spotted grouper, a protogynous hermaphrodite, revealing a candidate developmental trajectory of germ cells during spermatogenesis, providing novel markers genes at different stages of the transition from spermatogonial stem cells to mature spermatozoa. 28 The second offered evidence for five distinct cell types in the ovary of Asian seabass (Lates calcarifer), a protandrous hermaphrodite, including germ cells; revealing novel oocyte marker genes, including shared marker genes with human oocytes. 44 Such a comparative approach across different species may reveal shared factors underpinning sex determination and early sex differentiation, contributing to our understanding of the rapid evolutionary turnover of sex determining mechanisms and helping towards sex control efforts. Finally, an improved understanding of the transcriptome and development of germ cells, including the associated marker genes, may also be useful to expedite progress in surrogate broodstock technologies, which have major future applications in aquaculture research and stock genetic improvement. 114

3.6 | Selective breeding

Selective breeding is the main route for the genetic improvement of aquaculture stocks. At the centre of these efforts are genome-wide association studies (GWAS), which have identified quantitative trait loci (QTL, i.e., regions in the genome correlated with variation in a target phenotype, usually captured by SNP markers) for traits of interest in many aquaculture species, including growth and resistance to diverse diseases. 39 Yet the causative genes and mutations underlying these QTL remain elusive, and selection efforts rely on neutral markers in linkage disequilibrium with causative genetic markers, which has limitations for cross-generation and cross-population selection using genome-wide information (e.g., Ref. 115). Single-cell studies can contribute towards dissecting QTL through more precise assessment of the genes co-localizing with QTL. This could involve inferring cell-specific expression of genes within QTL regions, or investigating cell-type-specific differences between individuals carrying distinct genotypes for the QTL. The power of scRNA-seq to understand the cell-specific nature of GWAS hits has been demonstrated in humans, 116,117 paving the way for similar studies in aquaculture species.

Single-cell technologies also allow for more precise definition of connections between molecular phenotypes (e.g., gene expression) and genetic variation. 118 It is now known that most causative variants fall within regulatory regions of genomes (e.g., Ref. 119), making expression QTL (eQTL, i.e., genomic regions explaining individual variation in gene expression levels) increasingly to determine the genetic basis for trait variation in aquaculture populations. While initiatives such as FAANG aim to improve our understanding of non-coding regions and eQTL in farmed animals, including aquaculture species, 55 such work has been based on bulk methods. However, many eQTL are cell-specific (e.g., Refs., 120,121), highlighting an increasing need for cell-resolved eQTL maps (e.g., Ref. 13). Furthermore, in addition to discovering the causes underlying QTL for prioritized traits in aquaculture, cell-resolved eQTL can be fed directly into selective breeding models to prioritise functional variation (e.g., Ref. 122). It should also be noted that cell-resolved eQTL analysis is possible following deconvolution of bulk RNA-seq datasets. 52–54 Therefore, as the generation of population-scale bulk RNA-seq is becoming increasingly affordable, it should be readily possible to design studies with aquaculture species that combine bulk data with a smaller set of single cell data for deconvolution, enabling eQTL analysis.

To summarise, up-taking single cell transcriptomics into future research on the genetic basis for commercial trait variation will help increase the accuracy of selective breeding, leading to more efficient and resilient aquaculture stocks through genetic improvement.

4 | EXPERIMENTAL CONSIDERATIONS

Single cell transcriptomics is performed using two fundamental strategies, scRNA-seq or snRNA-seq, which require cells and nuclei as the input, respectively. Consequently, a central consideration is which strategy to select (Figure 2). This decision is influenced by practical issues including whether fresh tissue is readily available, or whether it’s essential to freeze samples, which may be the case when sampling aquaculture species. The quality of scRNA-seq and snRNA-seq data is highly correlated with the quality of cells or nuclei input to library preparation, demanding optimization efforts to ensure high-quality outcomes are achieved. In this section, we focus on considerations when designing single cell experiments, accounting for issues faced by researchers working with aquaculture species. We first outline fundamental considerations around the choice of performing scRNA-seq versus snRNA-seq (Section 3.1), before reviewing methods for isolating cells and nuclei from fresh and frozen tissues as the input to scRNA-seq (Section 4.2) and snRNA-seq (Section 4.3), respectively.
4.1 Sequencing cell or nuclei transcriptomes?

A large number of studies have considered the relative merits of scRNA-seq and snRNA-seq across diverse biological contexts and tissues, including mammalian brain, kidney, liver, cardiomyocytes, adipose tissue, peripheral blood mononuclear cells, and cell lines. While the general consensus is that both strategies typically perform well using the same platform, there are important distinctions and potential strengths and weaknesses to consider when designing an experiment. RNA captured from whole cells is derived from all compartments of the cell, which means scRNA-seq datasets are most compatible with knowledge gained from transcriptomic studies using bulk datasets, which likewise capture all cellular fractions. Furthermore, as extensive post-transcriptional regulation occurs outside the nucleus, scRNA-seq and snRNA-seq capture distinct information on gene expression dynamics, with snRNA-seq lacking scope to capture transcriptional regulation after RNA nuclear export, but on the other hand providing more direct readouts on transcriptional regulation after RNA nuclear export, but on the other hand providing more direct readouts on transcriptional regulation. Past work has shown that scRNA-seq datasets are enriched for mitochondrial and ribosomal genes, while snRNA-seq datasets are enriched for nuclear RNAs, including long non-coding RNAs. snRNA-seq data also shows a markedly higher abundance of unspliced mRNA containing intronic sequences, which has implications for downstream bioinformatics (Section 5) and may lead to biases in the expression of particular genes for some sample types (e.g., Ref. 124). Despite these clear differences, a range of studies have shown that data derived from nuclear and cell transcriptomes is highly correlated in many sample types (e.g., Refs. 128,129,131,132).

Furthermore, at least some of the apparent differences in gene expression between the two strategies may be due to sampling (e.g., impact of freezing vs. not freezing), rather than inherent differences between nuclei and cell transcriptomes. A known consideration is that scRNA-seq and snRNA-seq datasets often capture a very distinct representation of cell type diversity. Numerous studies have shown that cell diversity is biased in scRNA-seq datasets across different tissue types, through there also exist cases where specific cell types were underrepresented in snRNA-seq datasets (e.g., human microglia). For scRNA-seq, this issue relates mainly to the distinct sensitivity of different cell types during cell dissociation from tissues (Section 4.2), which is not straightforward for many sample types, and in some cases, it may be difficult to recover some cell types for sequencing (e.g., Ref. 135). While less well recognized, different cell types likely have distinct sensitivities to being processed through platforms used for scRNA-seq, particularly droplet based methods employing microfluidics (Section 2). The need to dissociate cells before scRNA-seq also activates stress associated genes (Section 4.2), which may impact downstream data interpretation (e.g., Ref. 133). This issue is thought to be largely avoided in snRNA-seq experiments, and it is also true that nuclear isolation is more straightforward for many samples poorly amenable to cellular isolation (Section 4.3).

Building on these general issues, it is important to consider practicalities surrounding sampling using aquaculture species, which often necessitates trips to field situations (e.g., fish or shellfish farms) and/or sampling sites physically located far from the facilities where single...
cell transcriptomics is performed. In such cases, another challenge of using scRNA-seq is the need to move from cell isolation to library preparation quickly to limit negative impacts on cell viability and/or cell type representation. This issue is especially important considering that we will often start with limited expectations about cell diversity/representation in samples for many aquaculture species and their tissues, making it impossible to detect biases owing to a lack of baseline understanding. On the other hand, snRNA-seq is commonly performed using flash frozen tissue samples, which is compatible with sampling set-ups used widely in aquaculture research. Based on the literature reviewed above, snRNA-Seq is likely to give a less biased representation of cell types for many sample types, which is desirable when working with poorly characterized species and cells. Consequently, the recognized benefits of snRNA-Seq are particularly relevant to studies of aquaculture species cell transcriptomes. When working with a new species or sample, it would be wise to initially compare the results of both snRNA-Seq and scRNA-Seq, to determine the comparative representation of cell type diversity as a trade off against the relative quality of data captured. This approach is becoming increasingly common in the mammalian literature.

### 4.2 Obtaining high-quality cells for scRNA-seq

High quality scRNA-seq data are dependent on achieving cell suspensions with a high proportion (i.e., >90%) of viable individual cells. High cell viability also helps ensure that a sample’s cell diversity is present at the start of library preparation. Using cell suspensions containing dead or damaged cells increases the detection of RNA located outside cells, increasing background noise in the dataset (Section 5). Achieving a high quality single cell suspension is strongly dependent on the protocol used. Generally speaking, cell dissociation involves converting fresh tissue into a heterogeneous soup of its constituent cells. It is common to dissociate tissues using mechanical means like douncing, cutting or pipetting up and down, with care required to avoid negative impacts on fragile cell types. Enzymatic digestion is also widely used, requiring considerations around which enzyme or enzyme cocktail to employ (e.g., trypsin, collagenase, etc.), in addition to the length and temperature of digestion (typically 30–60 min at 30–37°C). It is also common to combine mechanical and enzymatic digestion. Ideally, every study will achieve a balance between releasing ‘difficult to dissociate’ cell types, while avoiding damage to more fragile cells, which may not be easily achieved.

Most cell dissociation protocols were developed and optimized using mammalian samples. This poses issues when working with aquaculture species, as the architecture of tissues and sensitivity of cells to mechanical or enzymatic digestion may differ greatly from mammals. It is well-established that mammalian cells experience transcriptome-wide changes in response to common dissociation protocols, with incubations at 37°C inducing stress response genes. The same responses will be strongly amplified for cells from ectothermic species used in aquaculture. Providing an overview on this issue, Machado et al. concluded that virtually all cell types will express stress signatures given sufficient dissociation time. The choice of enzyme can also affect gene expression in cells. One possible solution to reduce negative impacts of cell dissociation is to use cold active proteases (active at <6°C), limiting cells heat stress. At this temperature, transcription is largely inactive in mammals, limiting artifacts linked to heat stress. This approach has been used with success in mammalian kidney, brain and solid tumour samples, and will presumably greatly reduce dissociation induced artefacts in ectothermic species.

Another way to avoid dissociation issues in aquaculture species is via the use of cells that do not require dissociation, including immune cells in the blood of rainbow trout, or oyster haemolymph. A gentle approach for cell dissociation may further be possible for soft tissues lacking extensive structure, for example, the spleen and head kidney of teleosts. In these studies, tissues were subjected to filtering and centrifugation to achieve cellular dissociation, which can be carried out at 4°C to limit stress responses. In summary, diverse options are available for cell dissociation, but care must be taken to limit the negative impacts of enzymatic digestion and associated heat stress. Ideally, every new tissue type will be subjected to trials to achieve optimal viability of cells before scRNA-seq.

The need to rapidly process fresh samples for library preparation in scRNA-seq can also be circumvented using fixation protocols. These include the use of fixatives including methanol and formaldehyde, in addition to cryopreservation in DMSO. Such options provide more flexibility when sampling and storing samples for scRNA-seq, but will still result in the same dissociation biases associated with using fresh cells. A promising method called ACME (ACetic-MEthanol) dissociation was recently established that overcomes this issue by simultaneously dissociating and fixing cells for later sequencing. Here the authors demonstrated that scRNA-seq after ACME dissociation avoided biases in cell diversity, emphasizing benefits of maintaining the complete cell transcriptome—in other words, avoiding limitations of scRNA-seq, while allowing cells to be stored and sequenced at a later date, that is, a major benefit of snRNA-seq.

### 4.3 Obtaining high quality nuclei for snRNA-Seq

For the reasons outlined in Section 4.1, it may not be possible or desirable to dissociate cells from fresh tissue samples, especially when working with aquaculture species. In such cases, flash freezing freshly-sampled tissues on liquid nitrogen or dry ice is compatible with the recovery of high quality nuclei for snRNA-seq at a later date, following a period of storage at an ultra-low temperature. An important consideration, as with any bulk experiment, is that the integrity of RNA will degrade through time even at very low temperatures. Therefore, it is advantageous to minimize the time samples are frozen prior to nuclear isolation and snRNA-seq library preparation, though it is not possible to give precise guidance here, given that there will be considerable variation in RNA degradation rates...
across tissues and species. Additional advantages of snRNA-seq include that standardising nuclear isolation across different tissue types is less onerous than attempting the equivalent cell dissociations, and that nuclear isolation protocols are done on ice at 4 °C, limiting the transcriptional activity of nuclei and associated impacts on gene expression.35

Many of the first snRNA-seq studies employed a nuclear isolation protocol using a standard commercial nuclear isolation buffer (‘EZprep’) with a combination of douncing and centrifugation.132 This protocol has been modified to incorporate a sucrose gradient to accommodate more delicate tissue types.150 However, a drawback when using a sucrose gradient is that this additional step increases the time between dissociation and library construction, potentially damaging or losing fragile nuclei. A similar protocol has been used in other sequencing assays that require nuclei from frozen tissue, including ATAC-seq.151 With minor adjustments such as swapping the protease inhibitor cocktail for RNase inhibitor, these protocols can be readily adapted for snRNA-seq.

A growing volume of literature has compared methods of nuclear isolation across different tissues.13,36,152 This work highlights disadvantages in the original EZprep method, including nuclei loss and high levels of ambient RNA. The faster and cheaper chopping extraction approach152 was shown to represent the most effective method for nuclear isolation with frozen tissue in terms of capturing diverse cell types and reducing background RNA.13,36 Chopping extraction is when nuclei are dissociated from cells using a custom nuclear extraction buffer, while chopping with precision scissors. In Eraslan et al.13 a toolbox is presented to optimise detergent use for chopping extraction in different tissue types that is very applicable to different species. With minor modifications, these protocols have been used in a diverse range of tissue panels in Atlantic salmon for successful snRNA-seq (e.g., Ref. 24). The addition of RNase to this protocol is desirable to reduce RNA degradation and background ambient RNA, especially for tissue types that show high endogenous RNase activity.153

5 | BIOINFORMATIC AND ANALYSIS CONSIDERATIONS

The analysis of single cell transcriptomics data is more complicated than bulk RNA-seq, owing to its higher complexity and dimensionality, which often captures the expression of tens of thousands of genes in thousands of cells or nuclei. The data tends to be much sparser, consisting largely of zeros, with sequencing depth varying extensively between different cell types. These features require dimensionality reduction approaches to make the analysis computationally tractable, alongside statistical methods that compensate for sparseness and noise in the data, in addition to inventive visualisations that make the outputs interpretable. Reviews exist elsewhere that outline general approaches to single cell transcriptomic data analysis, for example, Ref. 154, and this section mainly outlines considerations relevant to studies with aquaculture species, which transfer well to other non-model organisms (also see review by Ref. 155). The analysis pipelines discussed were designed for droplet-based technologies being widely applied in aquaculture species (Section 2, Table 1), but can generally be used with data derived from microplate based methods like SPLIT-seq.

5.1 | Limitations of genome annotations and cross-species cell markers

A key outcome of single cell transcriptomic data analysis is the generation of a count matrix, representing the number of sequencing reads or UMIs captured for each gene per cell/nuclei (the basis for all downstream analyses and visualizations). In the first step, the sequence data are usually mapped against an annotated reference genome to determine read counts for genes (Section 5.2). While analysis frameworks exist that do not require this a reference genome,156,157 they have not been widely benchmarked.155 There are many considerations surrounding genome annotation that impact on data generation and interpretation. The first is that if a genome assembly is of low quality and fragmented, this will impact gene prediction, meaning key marker genes may be missing, split or only partially represented in the predicted gene models. Annotated reference genomes are available for many aquaculture species,39 with most being of high quality owing to modern sequencing technologies. However, even in complete and accurate reference genomes, many correctly predicted gene models will be assigned names that are challenging to interpret. This results from the fundamental nature of functional annotation (assigning names or features to genes based on sequence characteristics), which is primarily derived from similarity to gene products from characterised species in public databases. The consequence is that genes may be named incorrectly, have low-confidence annotations, or lack functional annotations (e.g., ‘uncharacterised protein’).

An associated problem is that gene names assigned by automated genome annotation may often fail to represent the true orthologue (i.e., same gene inherited in different species from their common ancestor) to the genes from which their names were derived. Many gene families have complex evolutionary histories, characterized by losses and expansions, in addition to divergent evolutionary rates across lineages, which challenges accurate gene name assignment based solely on sequence similarity. For such gene families, sophisticated phylogenetic approaches may be required for accurate homology assignment (e.g., Ref. 158). As a simple example, if a gene has been lost by pseudogenization during the evolutionary history of a target species, it may be assigned a name for the next most closely related gene from a larger gene family. In other cases, gene families have been biasedly expanded in a particular species of interest, such that multiple co-orthologues exist to single genes in taxa from which gene names have been derived. This represents the rule for species with a recent history of whole genome duplication (WGD), including some of the most important farmed finfishes globally; that is, salmonids159,160 and cyprinids,161,162 which occurred on top of a WGD event in the common teleost ancestor. In such species, it is common for there to exist three or four genes sharing equal orthology to...
mammalian species, and these duplicated copies often show distinct expression patterns, which presumably extends to different cell types, for example, Ref. 24. However, duplicated genes retained from recent WGD events are poorly annotated in public sequence or genome databases, so care is required to ensure they have been properly captured in genomics studies to avoid spurious conclusions about functional differences between a target species and well-characterised taxa with more compact gene families.

Such issues are common to comparative genomics studies involving non-model organisms, but are particularly important to consider in single cell transcriptomics, owing to the standard practise of determining cell identity on the basis of cell marker genes. This process inescapably requires transfer of knowledge about cell marker genes from well-characterised species. Lying at the heart of this strategy is an assumption of genetic orthology, which as detailed above may often not be met, in addition to the assumption that gene cell-type expression is typically conserved across species. On this latter point, we already know from single cell studies of species with well-annotated genomes where gene orthology assignment is straightforward (i.e., humans vs. mice), that while some orthologous genes are reliable cross-species markers for the same cell types, others are not. Moreover, for species that possess multiple co-orthologues of marker genes from well-characterised species, it is clearly important to consider them holistically, rather than in isolation, to avoid spurious conclusions. For example, Taylor et al. identified that salmonid co-orthologues of established mammalian marker genes for specific hepatic cell types showed highly distinct cell-specific expression, pointing to the need for more work to define such patterns globally.

Despite the above issues, single cell studies of aquaculture species cited elsewhere in this review indicate that a sufficient number of conserved marker genes exist to confidently identify major cell types, for instance, the main classes of immune cells shared by all jawed vertebrates (i.e., B cells, T cells, macrophages, dendritic cells, etc.) in recent teleost work. Where things get more challenging is in assigning identity to distinct subsets within conserved cell types, which may have evolved recently, reducing the effectiveness of marker gene information from distantly related species. A good example is the extensive heterogeneity observed in rainbow trout B cells, with distinct subsets identified lacking shared marker genes for B cell subsets in mammals. As a separate related point affecting our ability to transfer knowledge on cell markers across species, technical differences between studies, such as the use of snRNA-seq versus scRNA-seq, as well as the platform used for analysis, can also change the repertoire of captured marker genes, even for the same species (e.g., Ref. 127).

The above points are intended to highlight the need for a critical approach to cell identification in non-model aquaculture species, including the need to be aware of the possibility of species-specific cell biology, marker genes that have yet to be characterized and limitations in the use of marker genes from distantly related taxa. Nonetheless, several strategies exist to address such challenges and support a more reliable analysis. Firstly, if there are prioritized genes of interest, it is possible to manually annotate them before adding these to the reference genome annotation. For example, in a recent single cell study in turbot, manual annotations of novel immune-type receptors (NITR) were added to the annotation prior to mapping using BLAST searches against the better annotated zebrafish genome. It can also be useful to bolster the quality of gene annotation using databases containing phylogeny-derived information on homology relationships. For example, the Ensembl database, via the Biomart function, allows researchers to extract information on the predicted orthologue to their full set of genes from any species in Ensembl. We perform this approach routinely to compare annotations from salmonid genes to their predicted human, mouse and zebrafish orthologues. For taxa that are not included in Ensembl, global orthology predictions can be derived using methods such as Orthofinder, and carried into single cell data analysis and interpretation (e.g., Ref. 166). On a smaller scale, constructing manual phylogenetic trees for key gene families is a useful strategy, as done recently in Atlantic cod, revealing several questionable annotations in the reference genome.

Overall, we advise that all single cell transcriptomic studies done in aquaculture species are based on a foundation that attempts to capture and interpret the correct evolutionary relationships between a target species and the taxa from which knowledge of cell biology is being inferred or transferred.

### 5.2 Data mapping and initial filtering

The first step towards generating a count matrix is to determine the genomic, cellular, and transcript origin of each sequenced read. This is typically performed using a single pipeline, which determines the genomic origin of reads via alignment to a reference genome, and assigns the cell or nuclei of origin using the cellular barcode (CB) and (when applicable) the transcript of origin using the UMI associated with each read. A popular pipeline to perform these steps is the 10x Genomics Cell Ranger software, which aligns reads to a reference genome, and associates each read to an error-corrected CB and UMI. Alternative packages such as STARsolo or Alevin perform the same function as Cell Ranger but also allow for the adjustment of sequence alignment parameters, which has particular value when working with non-model species, and the use of non-10x cellular/UMI barcode configurations.

Multi-mapping occurs when a read can be assigned to more than one location in the genome with similar statistical probability. This is a major issue for lineages with genomes characterized by the extensive presence of duplicated genes retained from recent WGD events, including salmonids and cyprinids. In some cases, duplicated genes are so similar that a significant proportion of reads map equally well to both locations in the genome. If we retain only the uniquely mapping portion of reads in species where duplicated features are common, extensive data loss (including marker genes) may occur, which will reduce the power of downstream analyses. Cell Ranger automatically discards multi-mapping reads that map to more than one gene, whereas Alevin and STARsolo offer models allowing probabilistic assignment of multi-mapping reads across duplicated genes. Thus,
more data can be retained, and an effort is made to accurately estimate expression levels of duplicate genes using information about the number of uniquely mapping reads to the different gene copies. Ale- vin and STARso also offer more flexible approaches for error cor- recting CBs and UMIs, as well as allowing user-specified read structures to allow the processing of data from non-10x Genomics platforms. There are several other differences, notably faster running time than Cell Ranger, as well as differences in final UMI and gene counts, which are described elsewhere.169

The final step is to determine which CBs are associated with real cells or nuclei, based on the raw UMI count associated with each CB.170 This step is non-trivial and even the best algorithms can result in the erroneous filtering of genuine low-UMI count cells. For exam- ple, the incorrect filtering of neutrophils in mammals by Cell Ranger, due to low RNA levels and high RNAse content, is a known issue that needs manual intervention to address.171 Suggested solutions are to bypass the automated cell filtering step and specify a set number of cells to be returned, or alternatively count intronic reads in addition to exonic reads to increase UMI count (both of these approaches are possible in all tools). This issue is likely to be particularly important for aquaculture species due to the current sparsity of studies profiling cells in these species, and the heterogeneous nature of datasets generated from whole tissues. For instance, in Atlantic salmon, there can be an order of magnitude difference in the RNA content of cells (e.g., Ref. 23) and erythrocytes in particular have extremely low RNA content and can be erroneously filtered by automated tools.24

5.3 | Additional filtering and quality control steps

After the count matrix has been generated, there are a number of additional steps that can be performed to enhance quality of downstream data. Most commonly, this involves a manual inspection of data to remove empty droplets and poor quality cells, the use of bioin- formatic tools to remove ambient RNA, the removal of ‘doublets’ (i.e., chimeric transcriptomes derived from more than one cell or nuclei), and imputation of missing expression values. The removal of empty droplets or poor quality cells is conducted through filtering thresholds on UMI count, gene count and mitochondrial content, and is usually conducted in downstream packages such as Seurat172 and ScanPy,173 both of which include detailed vignettes on the process. These packages also provide a diversity of clustering, differential expression and visualisation options and are recommended for many downstream data analysis purposes such as dimensionality reduction, cluster- ing and visualisation (Section 5.4).

In principle, each droplet in a droplet-based single cell dataset should contain either RNA originating from a single cell or nucleus, or no RNA at all. In reality, even in high quality datasets every droplet contains non-negligible amounts of contaminating RNA,174 and this can vary significantly between datasets. This effect can be readily observed in a gill snRNA-seq dataset from Atlantic salmon19 where all cell types ‘express’ haemoglobin, that has likely originated from erythrocytes. Similarly, abundant hepatocyte genes encoding acute phase proteins showed leakage to all cell types in a liver snRNA-seq dataset in Atlan- tic salmon.24 Several effective tools including SoupX174 CellBender175 and DecontX176 have been designed to remove ambient RNA, using the presence of known empty droplets containing only ambient RNA to estimate ambient RNA content in non-empty droplets. Cellbender in particular is highly suitable for the analysis of data gener- ated in aquaculture species, as it requires little prior knowledge of the data and few user set parameters, while SoupX and DecontX each require input on meaningful cell identifications, which may be chal- lenging to establish in the early stage of analysis with a novel species. Cellbender has the advantage of also performing cell filtering based on the de-contaminated dataset, when the removal of ambient RNA can clarify the distinction between empty and non-empty droplets.

5.3.1 | Doublet removal

Three approaches exist to deal with doublets in single cell datasets. The simplest is to apply an upper threshold on the UMI or gene count in the quality control step, under the assumption that doublets will contain more RNA and therefore more UMIs/genes. While this may work well in homogeneous data sets where all cells/nuclei express similar number of transcripts, in a tissue level dataset there will often exist enormous variation in transcriptional activity between cell types, resulting in an upper threshold erroneously removing the most transcriptionally active cells, while missing doublets containing less active cells. Therefore, this strategy is not generally recommended.

The second approach is to cluster the cells or nuclei (e.g., with Seurat or ScanPy) (Section 5.4), perform a differential gene expression test between each cluster and all other cells, and manually identify clusters that differentially express specific markers of two other cell types, but no unique markers of their own. This approach can be time consuming, allows for human error, and only identifies heterotypic doublets (i.e., two different cell types) but has the advantage of a bio- logical justification for the removal of cells.

The third approach is to employ a dedicated bioinformatic pack- age, many of which have been usefully benchmarked against each other.177 These usually operate by generating a set of artificial dou- blets based on an initial clustering of the data, which are used as the basis to identify actual doublets in a sample. In our experience, care should be taken when using a doublet removal package, as there can be significant disparities between packages in which cells are identi- fied as doublets. A manual inspection of the cells identified as dou- blets is recommended prior to removal. The scDbiFinder package178 uses a function that identifies clusters of likely heterotypic doublets based on an algorithmic version of the manual strategy of performing differential gene expression tests and identifying cell clusters that express markers of two other clusters, but no unique markers.

5.3.2 | Imputation

Imputation aims to fill missing data values by comparison to a ‘true’ reference. In the context of single cell data, this means using
transcriptionally similar cells to impute missing expression of genes that have been lost due to incomplete capture or sequencing of the RNA in the cell. The rationale for using imputation is that the imputed dataset will be superior for downstream analysis, but this is by no means clear. For example, while the recovery of gene expression profiles observed in bulk RNA-seq can be enhanced through imputation, this may result in little downstream enhancement to clustering and trajectory analysis, and there is the potential to introduce false positive results in differential gene expression tests. Despite this, imputation has been used to recover biologically meaningful expression of very lowly expressed genes that was not present in the non-imputed dataset (e.g., Ref. 181). Imputation should be conducted with care and the benefits weighed against the risks of introducing unwanted artefacts.

5.4 Clustering and cell type identification

To characterize the heterogeneity of cells in a single cell dataset, it is necessary to group cells sharing similar molecular profiles in a process known as clustering, which is fundamental to most downstream applications. Most clustering algorithms use machine learning approaches to cluster cells in an unsupervised fashion (i.e., without user input) and generally perform well. Prior to clustering, data dimensionality is reduced to make the analysis computationally tractable and to remove uninformative variation. This typically involves using a subset of genes that show the highest level of expression variation (perhaps 5%–10% of genes), before performing further dimensionality reduction, such as principal component analysis (PCA), and selecting the PCs explaining the most variation to eliminate uninformative noise. Cells are then grouped by similarity along these axes, with a graph-based clustering algorithm used to determine cluster number and assign cells.

Two critical parameters to consider during clustering are ‘resolution’, defining how fine-grained the definition of cell types will be, and the degree of variation used to inform the clustering (i.e., the number of PCs used as input). A study aiming to describe only the broad cell lineages present in the data should opt for low resolution clustering, requiring few PCs, while a study attempting to describe all variation, including heterogeneity within particular cell lineages, should opt for high resolution clustering, requiring more of the variation in the dataset (hence more PCs). For most datasets, particularly in under described aquaculture species, the clustering process will inevitably require multiple attempts at clustering, with better understanding of the data from early clustering attempts leading to better informed choices for parameters. With very heterogeneous datasets, such as those derived from primary tissue, the best approach is often to perform an initial global clustering to identify the major cell types in the samples, then sub-setting these cell lineages and performing separate clustering with parameters tuned to each lineage, for example, Ref. 24. This approach is commonly used and avoids under clustering very diverse cell lineages (e.g., haematopoietic, including immune cells) or over clustering homogeneous populations, but is time consuming. Alternatively, other systematic approaches have been developed to deal with this issue.

At this stage on an analysis, it is important to visualise the data in order to assess the performance and biological relevance of the clustering. Currently the most widely used visualisation of cell clustering is uniform manifold approximation and projection (UMAP), a non-linear embedding method that aims to project all variation in data onto two axes, resulting in a “map” where the proximity of individual cells reflects similarity in their transcriptome composition.

<table>
<thead>
<tr>
<th>Application</th>
<th>Purpose</th>
<th>Challenges</th>
<th>Example tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential expression tests between conditions</td>
<td>Identify cell type specific changes in gene expression in response to challenge or changes in environment.</td>
<td>Low numbers of biological replicates may result in false positives. Low RNA content in single cell datasets may lead to bias against lowly expressed genes being accurately identified as differentially expressed.</td>
<td>Many bulk RNA-seq differential gene expression tests work well with single cell data, in addition to single cell specific tests. A variety of DGE tests are available in scRNA-seq workflows, Seurat, SCANPy and Scater.</td>
</tr>
<tr>
<td>Trajectory inference</td>
<td>Infer dynamic changes in transcriptomic profiles, for example, along a developmental trajectory.</td>
<td>Sampling must take place at the time points where dynamic changes are taking places. Uncertainty in number of cells required to reliable infer trajectories.</td>
<td>Monocle, Velocyto, Slingshot</td>
</tr>
<tr>
<td>Deconvolution of bulk RNA-seq datasets</td>
<td>Estimation of cellular composition of bulk RNA-seq data.</td>
<td>No aquaculture specific challenges.</td>
<td>scBio, Bisque</td>
</tr>
<tr>
<td>Gene network analysis</td>
<td>Identify signalling pathways and molecular interactions between cell populations.</td>
<td>Reliance on derived knowledge of molecular interactions, which is not available for most species Noisy nature of single cell data can confound inference of networks.</td>
<td>CellPhoneDb, BTR</td>
</tr>
</tbody>
</table>
Identification and annotation of cell types involves the use of ‘a priori’ knowledge of marker genes to assign cellular identity. This can be performed by either visualising the expression of these markers in each cluster, for example with “violin” plots or heatmaps, or by performing differential gene expression tests between each cluster and all other cells, then referencing the most differentially expressed genes in each cluster against the ‘a priori’ markers. The design of differential expression tests employed to define cell marker genes is important to the outcome of this approach. For example, when investigating cell subtypes within a lineage, the resulting marker genes from the differential expression tests strongly depend on the background used for comparison, that is, whether the test is performed against all other cells in the experiment, or against only other cells in that lineage. For example, in Atlantic salmon liver, T cell subsets were better identified by comparison of expression within the T cell lineage, rather than against all other liver cells. For this reason is it often useful to analyse subsets of the data separately when attempting to annotate cell sub-types. Once clustering and annotation of cell types has been completed, many options exist for downstream analyses, and the choice will be informed by the goal of the study. Table 2 summarises several of the common bioinformatic analyses that can be performed.

6  |  FUTURE PERSPECTIVES AND CONCLUSIONS

Looking ahead, spatial ‘omics’ is a related set of technologies we expect to make a big impact on the characterisation of cell biology in aquaculture species. A limitation of the single cell methods reviewed here is the lack of in situ data to contextualize cell-specific gene expression (including in response to stimuli) in the background of where the cells are physically located or co-located within a tissue. The spatial organization of cells within tissues is vital to cell and tissue function, and may radically change under different physiological conditions, for example, the migration and interaction of cells of the immune system following disease challenge. Several methods are already widely used to explore gene or protein expression within cells and tissue organizations (e.g., in situ RNA hybridization and immunohistochemistry), which can be combined with novel cell marker genes gained from single cell transcriptomics. However, such approaches are limited in throughput. Spatial transcriptomics encompasses a group of recently developed methods that bridge the gap between low-throughput in situ expression methods, and high-throughput single cell transcriptomics (reviewed in Ref. 187). In essence, these methods capture transcriptomic read-outs in minute regions (tens of micrometre scale) sampled from tissue sections, maintaining the spatial location of each region to build up a bigger picture of gene expression across the sampled tissue. This approach is complementary to single cell transcriptomics, helping to interpret the function of cell types according to their spatial location in relation to known features of a tissue, particularly insightful when used in an experimental framework comparing different conditions.

To wrap-up, single cell genomics is being rapidly uptake in aquaculture species, and when applied alongside other emerging technologies will revolutionise our understanding of the cell-specific basis for traits of significance to sustainability and production goals. We have outlined some of the key envisaged applications and potential barriers to successfully adopting single cell technologies in aquaculture research, highlighting considerations for experimental design and execution both in the lab and during data analysis, with major implications for sampling decisions, data quality and interpretation, and even cost. All single cell studies require careful consideration and planning, and there exist no blanket options to guarantee standardized high quality data and interpretations in all species and systems. Clearly this field is moving rapidly, and will build rapidly upon the emerging knowledge gained from pioneering studies published in recent years, providing increasing assurance concerning methods best suited to different aquaculture species. In the future, we envisage a data-derived approach will increasingly drive forward advances in species-specific cell biology, which is needed to move beyond the current constraints of knowledge transfer from a few well-characterised species.

AUTHOR CONTRIBUTIONS

Rose Ruiz Daniels: Conceptualization; writing – original draft; writing – review and editing; visualization; funding acquisition. Richard S. Taylor: Conceptualization; writing – original draft; funding acquisition; writing – review and editing. Diego Robledo: Conceptualization; funding acquisition; writing – original draft; writing – review and editing; supervision; visualization; project administration; resources. Daniel J. Macqueen: Supervision; conceptualization; funding acquisition; writing – original draft; writing – review and editing; visualization; project administration; resources.

ACKNOWLEDGEMENTS

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

FUNDING STATEMENT

We acknowledge funding for our single cell research, including grants from the Scottish Universities Life Sciences Alliance (Technology Seed Funding Call) (all authors), the University of Edinburgh’s Data Driven Innovation Initiative (Scottish Funding Council Beacon ‘Building Back Better’ Call) (Rose Ruiz Daniels, Richard S. Taylor, Daniel J. Macqueen), the Biotechnology and Biological Sciences Research Council, including BBS/E/D/10002071 and BBS/E/D/20002174 (all authors), BB/W005859/1 (Richard S. Taylor and Daniel J. Macqueen) and BB/V009818/1 (Diego Robledo), and the Norwegian Seafood Research Fund 901631 (Diego Robledo).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.
REFERENCES


48. The State of World Fisheries. FAO. Food and Agriculture Organization of the United Nations. 2022. 10.1038/s41586-019-0969-x


62. The State of World Fisheries. FAO. Food and Agriculture Organization of the United Nations. 2022. 10.1038/s41586-019-0969-x


Estermann MA, Williams S, Hirst CE, et al. Insights into gonadal sex differentiation provided by single-cell transcriptomics in the chicken


How to cite this article: Daniels RR, Taylor RS, Robledo D, Macqueen DJ. Single cell genomics as a transformative approach for aquaculture research and innovation. Rev Aquac. 2023;15(4):1618-1637. doi:10.1111/raq.12806