Chapter 5

Concluding remarks

Developments in single-cell genomics are still shaping the way we define cellular identity. With the increasing number of cell types, organs and species profiled, we are bound to obtain an exhaustive overview of eukaryotic cell diversity, together with their genomic determinants. This work illustrates the importance of studying cell types across different tissues, and discussed computational challenges as well as solutions for the integrative atlasing of cellular diversity.

5.1 Cells and genes trade-offs in single-cell profiling

The number of cells profiled per study is still increasing exponentially (Svensson et al., 2018). This has been accompanied by a marked expansion in the number of studies using single-cell technologies (Svensson and Beltrame, 2019), much of it due to the spread in use of a more standardised cell isolation and sequencing pipeline, 10x Genomics' Chromium technology. This democratisation of single-cell omics is resulting in more cell types, tissues and species being profiled. Nevertheless, single-cell studies should be designed with a clear goal, and the choice of protocol should be adequate to the question at hand.

With regards to the type of sequencing, scRNA-seq protocols can be broadly split between full length transcript profiling and 5'/3' RNA tagging. Full length protocols - the most widely used being Smart-seq2 (Picelli et al., 2014) - follow in the footsteps of the majority of bulk RNA-seq studies. Smart-seq2 is still dependent on mRNA isolation by the poly-A tail, and thus does not reveal changes in non-polyadenylated transcript as other protocols might (Hayashi et al., 2018; Verboom et al., 2019). Despite this, Smart-seq2's full length characteristics have been important to study immune cells.

The development of TraCeR (Stubbington et al., 2016) and BraCeR (Lindeman et al., 2018) have allowed the detection of TCR and BCR transcripts in single-cell data, which in turn have been used to lineage trace T cells with similar developmental origin (Lönnberg et al., 2017) and Treg cell migration between tissues (Miragaia et al., 2019) (Chapter 2). Smart-seq2 has also allowed uncovering the diversity of KIR receptors in NK cells at the maternal-fetal interface (Vento-Tormo et al., 2018). Splicing-oriented studies with this protocol, on the other hand, have been scarce (Arzalluz-Luque and Conesa, 2018), yet splicing can be important in revealing important features of cell identity. Combination of Chromium and PacBio long read sequence has revealed cell type specific isoforms in mouse cerebellum (Gupta et al., 2018). Other changes in isoform usage also exist that can influence cell identity, yet this has been underappreciated.

The dominance of 3' and 5' sequencing protocols stems from the fact that a large number of cells is more important to revealing cell diversity in a given tissue or condition than increased sequencing depth or number of genes per cell (Svensson et al., 2019), which has been showed early on when profiling bipolar retinal cells in mouse (Shekhar et al., 2016). Droplet-based protocols allow the user to more easily isolate a large number of cells, which are then only sequenced at lower levels. This increase in cell numbers was necessary in the Treg cell work presented in Chapter 2 to detect the subpopulations composing the lymph node-peripheral tissue trajectory (Figures 2.2 and 2.3). Thus, at the transcriptomic level, different protocols can serve complementary functions - either increasing the resolution of the cellular census, or providing a more detailed representation of the molecular makeup of cell populations.

5.2 Building a transcriptomic atlas of cell types

The study presented in Chapter 2 shows that, to unravel the full extent of cell identity, it is not enough to unbiasedly profile a tissue, since even low-frequency cell populations may reveal functional heterogeneity. Furthermore, the relevance of this is sometimes only apparent once more tissue-specific context is added. In isolation, the census of colonic Treg cells would only reveal different levels of activation, but once this was combined with the draining mesenteric lymph node (mLN) populations, and compared with Treg cells in the brachial lymph nodes, it became clear that these subpopulations formed a continuum across organs, and the subpopulations present in the mLN expressed genes that coded for homing chemokine receptors specific for the colon.

The development of single-cell sequencing methods has unlocked the ability to perform unbiased cellular phenotyping. Yet there are several layers, from DNA, RNA and protein, to probe this phenotype. From these, at the single-cell level, RNA is by far the most widely available. While it is not as close to cellular function as proteins, it is a good approximation, and can be unbiasedly amplified. The spread of cellular transcriptomic profiling was not initially accompanied by a development of dedicated databases for this type of data, although more recent efforts have been made towards this end (Alavi et al., 2018; Franzén et al., 2019), and it is the goal of the Human Cell Atlas to gather and standardise single-cell expression data. Moreover, most of the data produced is not accompanied by cell type annotations in a machine-readable format, nor does it follow a standardised nomenclature. This is likely because the existing ontologies (Bard et al., 2005) were not prepared for this explosion in cell profiling and the diversity of cell types and states it brought. Thus, the existence of these scRNA-seq datasets creates an opportunity to develop and update an informed cell type reference (Aevermann et al., 2018). Chapter 3 introduced *CellTypist*, a method to integrate scRNA-seq data from multiple sources and tissues. This pipeline does not require a uniform annotation *a priori*, and produces an interpretable model for annotation of new data.

While most cell population profiling focuses on RNA, other aspects are also relevant. Open chromatin regions, which can be identified through scATAC-seq, are often involved in regulation of gene expression, and have been shown to be sufficient to distinguish cell types similarly to expression profiling (Cusanovich et al., 2018). An open chromatin cell type atlas can then provide a more regulatory perspective on cell identity, perhaps more clearly illustrating what effective alterations at the DNA level result in acquisition or loss of cellular phenotypes.

Cell type references like *CellTypist* have a multitude of applications, ranging from basic science to applied biomedicine. The predictive capabilities of this sort of models can be used to test cellular responses in organoids (Brazovskaja et al., 2019). This assessment can range from evaluating the differentiation potential of cultured cells, to measuring deviations and responses caused by external factors, like varying differentiation molecules or infectious agents. Ultimately, this can further improve the efforts in the field of tissue engineering, by guiding the development of *in vitro* differentiation protocols (Camp et al., 2018). Likewise, these references can also be used in a clinical setting to probe changes in cell diversity in disease. In cancer, infiltration and phenotypic changes of immune cells can be assessed, and single-cell phenotyping of tumour cells can monitor its progression. Monitoring of

cell abundances and diversity in the clinic can provide a new view of disease from a "cell ecology" perspective.

More fundamentally, *CellTypist*, as an integrated, cross-tissue cell type compendium can inform us on the key genes that define the core cellular phenotype.

5.3 Defining cellular identity

The advent of single-cell genomics has re-ignited the debate on the definition of cell type identity. Historically, cell types have been defined based on their morphology, location, function, or developmental origin. Development of cellular staining, and especially flow cytometry, have added molecular phenotyping to this list. While flow cytometry already offered a large cell throughput capable of detecting even the smallest populations, the revolutionary aspect of single-cell transcriptomics has been the unbiased probing of RNA molecules, revealing a new high-resolution cellular map of gene expression programmes.

It is only through integration that we can achieve a organism-scale picture of cell identity. This is achieved by *CellTypist*, which is capable of resolving cell type correspondences across tissues (Figure 3.1, Figure 4.3), and provides the list of genes at the core of each cell grouping (Figure 4.4). Despite the discussed limitations, owed in part to the still limited diversity of data available, *CellTypist* lays the groundwork and reveals the first systematic picture of human cell types (with an expansion to other species in sight).

The transcriptomic composition of cells is vastly informative for their taxonomy, yet only makes up a small portion of the information we can obtain. Other omics modalities (open chromatin, chromatin modifications, methylation, proteomics, ...) can provide equally informative yet complementary perspectives on cellular identity. Furthermore, these can be integrated computationally (Stuart et al., 2019) or obtained simultaneously using appropriate protocols (Angermueller et al., 2016; Clark et al., 2018). Nonetheless, an ideal compendium of cell types should strive to go beyond this low level and invasive characterisation, and merge back into the knowledge obtained from other modalities. Cellular interactions are of great importance to cell function, and thus spatial information adds a relevant layer to this. Mapping the developmental trajectories of all cell types can inform us on their origin and generative processes. Morphology is the most easily observed characteristic, and heavily related to cell function, thus controlled by the genome. Only through integration can this systems view of cell biology come to fruition.

When possessing information on these many layers of cell phenotypes, we will be able to more accurately define the boundary between cell types and cell states. Often these can be observed in each individual modality - transient versus definitive cell shapes, immune lineages and their response to pathogens, or intermediate versus leaf stages in cellular differentiation. Yet these perspectives need each other, as cellular form and function should be understood in the context of its origin and genomic programming. Reconciling these different perspectives through a multi-window approach will provide us with a complete blueprint of the basic unit of life. It is expected that the unified view provided by the Human Cell Atlas - and indeed all cell atlases - results in a Modern Synthesis of Cell Theory.