Chapter 1

Cellular identity in the genomics era

Cell biologists have attempted, from the inception of the discipline, to categorize the extensive variability of cells that are found in Nature. This endeavour is hampered by the intrinsic complexity of cells, which associated to their small size and sensitivity to the surrounding environment, makes cellular phenotypes hard to probe in an integrated and comprehensive way. The last decade however has seen extraordinary improvements in the detail to which molecules can be assayed in individual cells. Single-cell RNA-sequencing (scRNA-seq) has for the first time provided an unbiased, transcriptome-wide census of RNA molecules for one cell at a time. By acquiring the transcriptome of large numbers of cells, we can group them by their gene expression programmes - a proxy for their function - and thus define their cell identity. The definition of this cell type identity from the massive amounts of transcriptome data produced in recent years has required the continuous adoption of new computational and analytical methodologies.

This chapter provides an introduction to the definition of cell types. It will show how more recently developed experimental and computational approaches are shaping our understanding of how cells are categorized.

1.1 Cell type discovery and definition

The term "cell" was coined by Robert Hooke in the 17th century to describe the empty cell walls he observed in cork samples through his microscope (Hooke, 1667). This observation was complemented some years later, when Antonie van Leeuwenhoek first observed live unicellular organisms and other cells with a microscope composed of more powerful lenses (Mazzarello, 1999). Research and observations in the following 200 years led to the formulation of cell theory. Its first tenet was introduced by Schleiden and Schwann, and states that all living structures are composed of cells or their byproducts (Schwann, 1847). The theory was later complemented by Robert Remak, Rudolf Virchow, and Albert Kölliker to include the postulate that all cells are derived from other cells (in the latin formulation popularized by Virchow, *omnis cellula e cellula*).

These early studies looked at a variety of sources to unveil different types of cells. Leeuwenhoek reported observations from blood, brain, muscle and semen (Leeuwenhoeck M, 1674; Leeuwenhoek Antoni Van, 1677). Subsequent developments of microscopy techniques led to improved imaging of a variety of tissues and the cells that compose them. For the first centuries of cell biology, microscopy was the method of choice to identify cell types. While this was mostly due to the relatively reduced knowledge of cellular biochemistry, it was immediately apparent that morphology was intrinsically tied to cellular function. The most illustrative example of this is the neuron, whose unique structure was only unravelled after subsequent improvements in tissue preparation and staining, as well as increases in resolution and development of electron microscopy (Mazzarello, 1999). Microscopy was also important in understanding where cell types come from by mapping their developmental origin. The three germ layers - endoderm, mesoderm, ectoderm - were identified in the 19th century, and was postulated that each of them would give rise to different sets of tissues (Collins and Billett, 1995). Developmental studies have since had a central role in defining cell lineages, and thus how cell types are related. Advances in microscopy were also crucial to the identification of organelles. While larger structures, like nuclei, are still identifiable with simpler microscopes (Brown, 1866), others required improved resolution and staining or preparation to be identified (Golgi and Lipsky, 1989). Other advancements in microscopy like live-cell imaging or super resolution microscopy are constantly perfected to expand the boundaries of cellular functional characterization.

Advances in biochemistry and molecular biology revealed that most organic molecules that compose cells are directly responsible for their function. Proteins are responsible for most cellular functions, being involved in enzymatic reactions, signalling and regulatory pathways or structural components. They became a prime target for cellular phenotyping with the development of immunostaining (Coons et al., 1941), whereby an antibody that specifically targets a certain protein is usually tagged with a fluorophore. Immunostaining can identify protein expression in tissue slices, and the use of different fluorophores allows for the imaging of cells expressing multiple proteins. The usefulness of immunostaining became especially apparent when it was combined with high-throughput microfluidics methods and used for fluorescence-activated cell sorting (FACS) (Bonner et al., 1972). This introduced the first high-throughput studies on molecular phenotyping of cells, and sorting allowed cell function to be probed in parallel (Julius et al., 1972). More recently, mass cytometry has allowed for a further expansion of the repertoire of proteins assayed (Bandura et al., 2009; Di Palma and Bodenmiller, 2015). This technique, while destructive, has also been combined with tissue imaging, adding a spatial component to the cell populations examined (Chang et al., 2017).

The identification and classification of cell types is dependent on their function. Function is deeply related to cellular morphology (Prasad and Alizadeh, 2019), and both are ultimately a consequence of the molecular pathways shaping them. Additionally, even though recent advances permit high throughput cell sorting through imaging (Nitta et al., 2018), the limited resolution hinders the identification of finer details of cell and organelle shape, which are frequently more informative of cellular activity. Cell sorting with fluorescent antibodies and mass cytometry can reveal more details on the molecules underlying cellular behaviour, but they are targeted approaches that depend on prior knowledge of the effector molecules. The more recent attempts at defining cell identity have therefore relied on the unbiased, high-throughput character of single-cell RNA-sequencing methods.

1.2 Defining cell types using scRNA-seq

Methods to sequence the transcriptome of individual cells started to be developed shortly after the advent of RNA-seq (Mortazavi et al., 2008; Tang et al., 2009). This early development was pushed not by a need to define the molecular makeup of the unit of life, but rather to allow transcriptomic studies to be performed in low-input samples. Nonetheless, this seminal work still sparked the improvements that occurred in the decade that followed (Svensson et al., 2018) (Figure 1.1).

Initial developments focused on increasing sensitivity, since the original scRNA-seq protocol was performed on cells from very early developmental stages, which are larger and contain more RNA than most differentiated cell types. Different methodologies quantified gene expression by sequencing distinct transcript segments (either the 5' or the 3' end, or the full transctipt) (Hashimshony et al., 2012; Islam et al., 2011; Picelli et al., 2014; Ramsköld et al., 2012). The idea of multiplexed scRNA-seq also started gaining traction with the use of multi-well plates or molecular

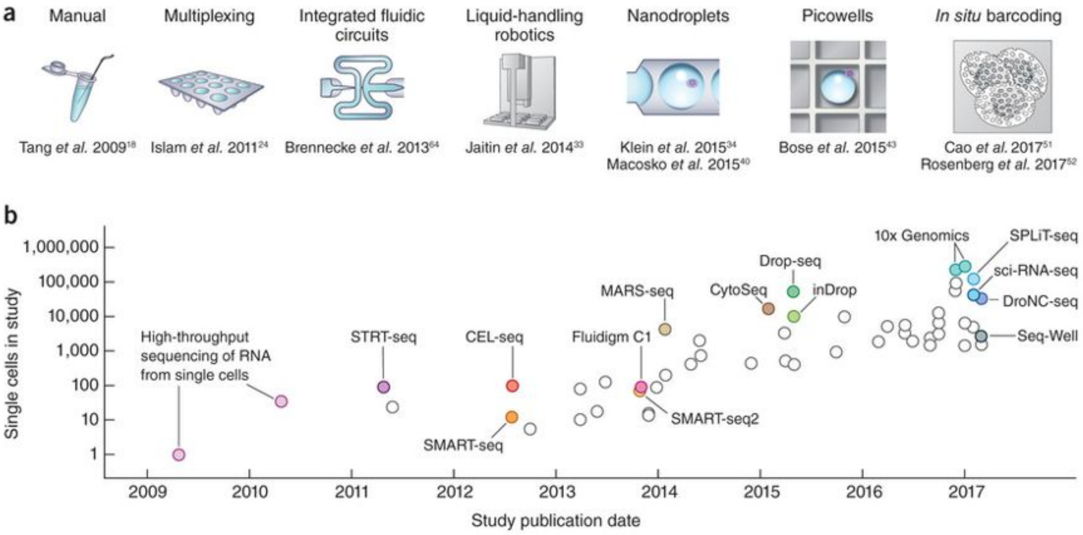

Fig. 1.1: **Timeline of scRNA-seq technology development**
(A) Key technologies that have allowed jumps in experimental scale. A jump to ∼100 cells was enabled by sample multiplexing, and then a jump to ∼1,000 cells was achieved by large-scale studies using integrated fluidic circuits, followed by a jump to several thousands of cells with liquid-handling robotics. Further orders-of-magnitude increases bringing the number of cells assayed into the tens of thousands were enabled by random capture technologies using nanodroplets and picowell technologies. Recent studies have used in situ barcoding to inexpensively reach the next order of magnitude of hundreds of thousands of cells. **(B)** Cell numbers reported in representative publications by publication date. Key technologies are indicated. *Original figure published in (Svensson et al., 2018).*

barcodes for cells. The company Fluidigm eventually introduced the first commercially available microfluidics chips (called the "Fluidigm C1 system") for miniaturized cell isolation, RNA extraction and reverse transcription (Brennecke et al., 2013). It is from this point that increased cell capture becomes the major technological driver (and has gained even great importance as discussed in Section 1.3). The major contributors to this have been nanodroplet-based technologies, that have put the number of profiled cells per dataset in the range of 10.000 to 100.000 (Klein et al., 2015; Macosko et al., 2015). The importance of this increase in throughput has been demonstrated by Shekar and colleagues (Shekhar et al., 2016), where they demonstrate that a Drop-seq dataset of approximately 25.000 cells sequenced at low depth could identify more *bona fide* cell types and subtypes than a smaller, more deeply sequenced Smart-seq2 dataset. Currently, most single-cell RNA-seq datasets use droplet-based technologies, chiefly the protocols designed for the Chromium instrument by 10x Genomics (Zheng et al., 2017), which have a higher sensitivity to detect different transcripts. Other more recent methods have followed the trend of increase in cell throughput by using multiplexed barcoding, which allows for different samples to be combined and reducing sample processing costs, reaching 10^5-10^6 cells for less than $0.01 per cell (Cao et al., 2019a; Rosenberg et al., 2018). A list of the most up-to-date scRNA-seq methods can be found in Table 1.1.

Table 1.1: Current methods for single-cell RNA-sequencing

Method Name	**Reference**
Fluidigm C1	(Brennecke et al., 2013)
Smart-seq2	(Picelli et al., 2014)
Drop-seq	(Macosko et al., 2015)
inDrop	(Klein et al., 2015)
CEL-seq2	(Hashimshony et al., 2016)
Chromium	(Zheng et al., 2017)
ICELL8	(Goldstein et al., 2017)
Quartz-seq2	(Sasagawa et al., 2018)
mcSCRB-seq	(Bagnoli et al., 2018)
SPLiT-seq	(Rosenberg et al., 2018)
MARS-seq2	(Keren-Shaul et al., 2019)
sciRNA-seq3	(Cao et al., 2019a)
Seq-Well S^3	(Hughes et al., 2019)

The exponential developments in single-cell sequencing technologies were accompanied by essential computational developments to analyse the resulting data. From a cell type discovery perspective, the key methods are clustering and pseudotime analysis (Rostom et al., 2017), which assign to cells a discrete or a continuous label, respectively. These are of course dependent of the upstream processing steps of normalisation, feature selection and dimensionality reduction, as well as often used batch correction methods (Luecken and Theis, 2019). Most of these analysis steps are available in accessible software toolkits (Butler et al., 2018; McCarthy et al., 2017; Wolf et al., 2018).

With clustering, the goal is to identify discrete cell populations. The most widely used methods for clustering are the louvain and leiden community detection algorithms (Blondel et al., 2008; Traag et al., 2019). These populations are commonly considered an approximation of the cell types present in a sample of dataset, often justified by examining the presence of known markers for known cell types across clusters. Further application of differential expression methods (extensively benchmarked in (Soneson and Robinson, 2018)) between clusters can identify other potentially novel genes that are, within that context, unique to that population. This can be used to characterise newly discovered populations (Montoro et al., 2018; Shekhar et al., 2016; Villani et al., 2017) and to identify new markers that can be used to isolate or understand known cell types (Bjorklund et al., 2016; Shulse et al., 2019; Vento-Tormo et al., 2018).

Pseudotime analysis consists on describing a set of cells from a continuous perspective. The name derives from the original application to obtain a dimensionless temporal trajectory from time course scRNA-seq data (Trapnell et al., 2014). There are several methods to perform this analysis (exhaustively reviewed in (Saelens et al., 2019)), all with the goal of defining a latent variable from the data along which a biological process, reflected in gene expression, is changing. Pseudotime is especially useful to study response to stimuli (Lönnberg et al., 2017; Trapnell et al., 2014) and developmental trajectories (Cao et al., 2019a; Watcham et al., 2019), but has also been used to model changes to cellular spatial distribution (Scialdone et al., 2016). These methods can differ in the way they model biological trajectories, with some explicitly allowing for branched trajectories. This is of special importance in development, where the goal is usually understanding which daughter cell types share progenitors. The direction of differentiation is usually just assumed according to previous knowledge and of the experimental conditions. This is not completely possible in all situations, yet can be inferred from expression data. By considering RNA kinetics, and using the quantification of spliced/unspliced reads, the current and future (i.e. still circumscribed to the nucleus) transcriptomic states can be untangled as a "velocity" vector (Manno et al., 2018). In differentiation trajectories, cell types are therefore usually defined as the endpoints, with the cells in between forming more transient cell states, along which gene expression is dynamically adjusting to the final cellular identity. It should be noted that this "cell type vs cell state" nomenclature is context-dependent, and there is no absolute agreement on how cell types should be formally and empirically defined (Various, 2017).

Globally, the increasing adoption of scRNA-seq is due to its multi-gene and unbiased profile. It allowed for the first time the non-directed profiling of molecules driving heterogeneity in cellular populations. Nonetheless, its use for defining cell identity still has some drawbacks. Even though the cost of high-throughput sequencing keeps dropping, single-cell RNA-seq still requires costly protocols, especially at the scale that it is currently performed for cell type discovery. This however can be mitigated by more targeted approaches, aimed at characterizing specific subsets of already known cell types isolated by their broad markers. scRNA-seq is also prone to batch effects, which can become more pronounced when comparing or integrating data generated by different protocols. This has been a very active topic of research, and several batch alignment and correction methods can now account for these integration of different protocols (Butler et al., 2018; Haghverdi et al., 2018; Park et al., 2018; Stuart et al., 2019). From the protocol side, sample barcoding for multiplexed processing also greatly reduces batch issues (Shin et al., 2019; Stoeckius et al., 2018).

One last concern, although perhaps the largest, is the fact that profiling a tissue or a cell type with scRNA-seq does not inherently give any functional information about the cells. Cellular function has been from the beginning the major point to categorize cells. RNA, despite being easily correlated with protein presence, is not in most cases the effector molecule in a biological process. Additionally, most single-cell methodologies destroy the cell without imaging it, making the link between molecular makeup and morphology harder to obtain. While this is an ongoing research topic, profiling cells through the use of multi-omics technologies can help obtain a deeper mechanistic characterization. Information on open chromatin regions (Buenrostro et al., 2015), histone modifications (Kaya-Okur et al., 2019) or surface proteins (Stoeckius et al., 2017) have the potential to be combined, directly or indirectly, with single-cell RNA-seq (Clark et al., 2018). This can provide information on how these molecular layers interplay and learn about the intrinsic regulatory processes of gene expression (Gorin et al., 2019; Qiu et al., 2019). CRISPR screens with single-cell expression readout can also reveal more about cellular function (Datlinger et al., 2017; Dixit et al., 2016). Lastly, developments in spatial transcriptomics hold the promise of providing spatial context to cellular transcriptomes profiled individually, providing information on the tissue context for cell identity determination (Rodriques et al., 2019; Vickovic et al., 2019). Overall, while the discussion about where to draw the line between cell types still lasts, technological developments provide us with ever increasing information to approach a decisive and informative definition.

1.3 Methods for cell type classification

Single-cell RNA-seq was initially developed to obtain the whole transcriptome from samples with very low starting material (Tang et al., 2009). Nonetheless, the notion of using it to define cell types through their transcriptome was very early on envisioned. In 2011, Islam and colleagues end the discussion on their newly developed scRNA-seq method (STRT-seq) by stating "We envisage the future use of very large-scale single-cell transcriptional profiling to build a detailed map of naturally occurring cell types, which would give unprecedented access to the genetic machinery active in each type of cell at each stage of development." (Islam et al., 2011). The exponential increase in the number of cells profiled per experiment eventually made this prediction come true. A large amount of single-cell projects have used the technology to profile cells captured from various tissues, in steady-state or disease conditions. Yet the most direct example of how this quote reflects the evolution of the field is the Human Cell Atlas (HCA) (Regev et al., 2017). This consortium has been established as a forum for scientists around the world to share their expertise on genomics, bioinformatics, and tissue biology, and coordinate the high-throughput profiling of cellular heterogeneity in the human body. The HCA has groups focusing not just on individual organs, but also on development (Behjati et al., 2018; Taylor et al., 2019) and disease.

In parallel, there have been increased efforts to obtain similar references for other species, in particular animal models (Cao et al., 2017; Fincher et al., 2018). The data collected for these species tends to have a greater cell coverage since the tissue samples can be more readily available. Furthermore, these atlases are by no means less important or useful than the human reference. The cell atlases produced for mouse (Han et al., 2018; Various, 2018) were of especial relevance, since they constitute the first broad, multi-organ cellular census of a mammalian organism, and one for which a large portion of biomedical science has relied on. The accessibility of human tissues for profiling and *in vitro* testing will be crucial in the near future. Nonetheless, having a mouse reference that can be related to human can not only teach us about the evolutionary principles that shape cell type evolution through gene expression, but also serve as a bridge to transpose mouse-based biomedical discoveries into a human context.

For a cell atlas to be used as a reference, it needs not only the expression data to be annotated, but also a computational framework that can use it to classify new datasets of interest. Over the last two years, several methods have been developed to handle scRNA-seq data (a comprehensive list can be found in Table 1.2), which can be added to other general purpose classification methods. These methods vary in complexity, but in general they rely on machine learning approaches to map the reference cell labels to the target dataset. While the most accurate method for this classification is still up for debate (see (Abdelaal et al., 2019; Köhler et al., 2019) in addition to benchmarks in individual method papers), there is agreement about the major challenges for this task. Classification methods should be aware of batch differences, be they caused by use of different scRNA-seq protocols or other technical differences in tissue processing. Different cell isolation and library preparation protocols can have a large impact on the number and type of genes detected (Mereu et al., 2019).

Table 1.2: Comprehensive list of papers detailing methods for automated cell state matching

Method Name	Short Description	Reference
scmap	k-nearest-neighbor search with cosine distance	(Kiselev et al., 2018)
matchSCore	Jaccard Index for cluster markers	(Mereu et al., 2018)
ClusterMap	Hierarchical clustering with marker gene binary expression	(Gao et al., 2018)
CaSTLe	XGBoost classification	(Lieberman et al., 2018)
Moana	Linear SVM on (sub)clusters	(Wagner and Yanai, 2018)
SAVER-X	Autoencoder	(Wang et al., 2018)
scQuery	Neural network classifier	(Alavi et al., 2018)
PopAlign	oNMF, Gaussian Mixture model and Jeffrey's divergence	(Chen et al., 2018)
scGen	VAE and linear classifier	(Lotfollahi et al., 2018)
scVI	VAE and hierarchical Bayesian model	(Lopez et al., 2018)
scPred	SVM in principal component space	(Alquicira-Hernández et al., 2018)
SingleCellNet	Random Forest on binary marker expression	(Tan and Cahan, 2018)
CellAssign	Multi-variable model with marker genes and hierarchical Bayesian framework	(Zhang et al., 2019a)
ACTINN	Neural network	(Ma and Pellegrini, 2019)
scID	Linear Discriminant Analysis with marker genes	(Boufea et al., 2019)
SingleR	Spearman correlation with training data	(Aran et al., 2019)
Garnett	Elastic net multinomial classifier using markers from hierarchical cell types	(Pliner et al., 2019)
SCINA	bimodal distribution of signature genes,	(Zhang et al., 2019b)
Cell BLAST	Adversarial Autoencoder and nearest neighbour search	(Cao et al., 2019b)
scMatch	Correlation with individual sample or average of references	(Hou et al., 2019)
SuperCT	Neural network with binary expression	(Xie et al., 2019)
CellO	Hierarchical binary classifiers	(Bernstein and Dewey, 2019)
scCoGAPS & projectR	NMF and projection in that latent space	(Stein-O'Brien et al., 2019)
SciBet	Entropy test and Bayesian comparison of multinomial distributions	(Li et al., 2019a)
Seurat "Anchors"	CCA, L2-normalisation and mutual nearest neighbours	(Stuart et al., 2019)
LIGER	integrative NMF and joint clustering	(Welch et al., 2019)
cellHarmony	Correlation with cluster centroids of mean marker gene expression	(DePasquale et al., 2019)
CHETA	Correlation with marker genes of hierarchical reference	(de Kanter et al., 2019)
scPopCorn	Co-membership Propensity Graph and (joint) k-partition	(Wang et al., 2019)
p-DCS	Voting based on known marker genes	(Domanskyi et al., 2019)
EnClaSC	Ensemble neural network classifier	(Chen et al., 2019)
scClassify	Ensemble classifier from inferred cell type tree	(Lin et al., 2019)

Many methods also mention the need to build a comprehensive reference, that should be integrated taking the into account the technical variability mentioned above. Training, and especially the prediction phases of the method should also be scalable. Models can take a very long time to train on larger references, and prediction steps that involve extensive manipulation or transformation of the target data can become time consuming with the ever growing size of expression matrices.

Lastly, some methods try to approach this classification problem from a hierarchical point of view (Lin et al., 2019; Pliner et al., 2019; Wagner and Yanai, 2018). This is based on the notion that cell types can be organised trees depicting phenotypic relationships. These trees represent not just developmentally-related lineages, but also the increasing specification of cellular function (still mostly correlating with terminal differentiation). This can be of great value in instances like describing cells from the immune system or the brain, where functional diversification leads to more intricate phenotypes (see Section 1.4). Notwithstanding, a hierarchical classification can also be seen as a method that reflects the uncertainty in the prediction. Each individual cell ideally conforms to a determined phenotype, which would correspond to a leaf node in an ideal cell hierarchy. Assigning a cell to a parent node rather than a terminal one (or not doing it with a high confidence) can be caused by data sparsity or low coverage, and thus not necessarily reflecting a naturally occurring hierarchy of gene expression-driven cellular phenotypes. Yet this structure is intuitive and informative, and projects like the Cell Ontology have considerable value in creating a controlled vocabulary to name and relate cell types (Bard et al., 2005), with some of the methods listed here explicitly conforming to it. The use of a curated and specific nomenclature should thus be incentivized when doing *de novo* annotation of scRNA-seq data, and supplying these labels can greatly accelerate the data interpretation and its application in the development of new algorithms.

Large collections of data and development of informative references can be of use in multiple ways. A steady-state cell identity reference can serve as a baseline to which a disease sample can be compared. Having a sufficiently comprehensive cell registry can do away with the need to generate a reference dataset if the goal is quantifying alterations to the proportions of known cell populations. Evolutionary biology can also benefit from predictive models for cell identity. Models can be adapted to function across species, which can help trace the evolutionary origins of cell types. Producing interpretable models from integrated data can also be informative in itself. Some models return the importance of genes or gene sets in classifying each cell type, and as such can help uncover novel features of a cell's phenotype. Finally, organised references can also speed up new in-depth studies of specific cell types, as well as studies focusing on other aspects of cell identity (e.g. open chromatin, methylation, proteome, or spatial interactions). It should be noted that the methods discussed so far in this section, while being in their majority developed for scRNA-seq, can also for the most part be adapted to other data modalities like scATAC-seq (for open chromatin) or CITE-seq (combining RNA and surface protein detection). Modelling cell identity with multiple layers can revel more details about the molecules shaping it, how they interact, and their relative importance.

1.4 Cell identity in the immune system

The immune system is one of the most complex and diverse biological systems across the animal kingdom. The increased evolutionary pressure caused by the need to continuously adapt to the fast evolving pathogens (Barreiro and Quintana-Murci, 2010) has resulted in a broad variety of molecular pathways and cells. The variability in the types of cells found in the immune system is directly related to their intrinsic plasticity in gene expression. Immune cells are very responsive to their environment, having to constantly fine-tune expression programmes to react in a prompt and targeted manner. It then comes as no surprise that many cell states have been determined and named in immunology, and it is, perhaps on par with neurobiology, the field where the definition of cell type and cell state clash the most.

Due to the fact that immune cells are non-adherent cells, immunology benefited immensely from the development of flow cytometry. Immune cells have been deeply characterised by this technology, with antibodies targeting surface receptors as well as cytoplasmic proteins. It then comes as no surprise that the immune system has been an early and major target of single-cell sequencing methods. scRNA-seq has had a role in the fine-grained mapping of gene expression changes in haematopoiesis (Watcham et al., 2019), discovering and reorganising subpopulations (Villani et al., 2017), mapping their heterogeneity across tissues (Miragaia et al., 2019; Scott et al., 2018), studying immune response to pathogens (Lönnberg et al., 2017; Stubbington et al., 2016), and map communication of immune cells with their tissue of residence (Vento-Tormo et al., 2018).

Immunity can be divided into innate and adaptive. The latter depends on a subset of lymphocytes which are responsible for an immune response that can flexibly adjust to invading pathogens in a non-evolutionary way (i.e. without the need for selection at the level of the individual). The key strength of this system is the use of receptors which recombine and mutate (Krangel, 2009), forming a highly diverse repertoire that can eventually be selected to respond to particular invaders. This variability, central to the adaptive immune response, is further complemented by immune memory, that is, the specific repertoire obtained when combating an infection will remain stored in the organism in the form of inactive immune cells, which can be more quickly reactivated should the same threat reappear. This is far more advantageous than having to undergo selection of the receptor repertoire every time the same pathogen is introduced in the system.

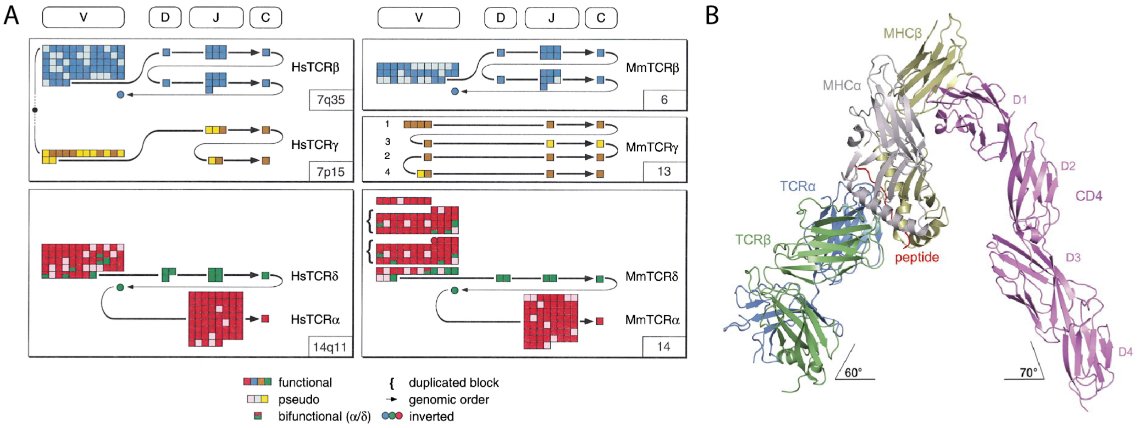

Fig. 1.2: **Gene and protein structure of TCR**
(A) The genomic organization of the human (left) and mouse (right) TCR genes α (red), β (blue), γ (brown), and δ (green), showing clusters of V, D, J, and C gene segments aligned vertically for clarity. Arrows represent the direction of transcription within each of the TCR genes; squares and circles indicate gene elements in the direct and reverse orientations, respectively. The murine TCR $\gamma 2$ gene is inverted relative to the rest of the locus. Dark colors indicate apparently functional gene elements, while lighter shades represent pseudogenes. Curly brackets indicate the duplicated sets of V genes in murine TCR α/δ locus. The TCR β and TCR γ loci are both on human chromosome 7, on opposite sides of the centromere (schematically represented by the black circle). *Original figure published in (Glusman et al., 2001).*
(B) Ribbon diagram of the complex oriented as if the TCR MS2-3C8 and CD4 molecules are attached to the T cell at the bottom and the HLA-DR4 MHC class II molecule is attached to an opposing APC at the top. TCR α chain, blue; TCR β chain, green; CD4, pink; MHC α chain, gray; MHC β chain, yellow; MBP peptide, red. *Original figure published in (Yin et al., 2012).*

Within adaptive immunity lymphocytes, T cells fill various niches, but are broadly considered to be the orchestrators of immune response (Kumar et al., 2018). T cells are characterised by their expression of the T Cell Receptor (TCR), a dimeric surface protein that can recognise an antigen presented by an Antigen Presenting

Cell (APC) (Reinherz, 2014). This receptor's ability to recognize a trove of antigens resides in the original gene's unique recombination capacity. TCR genomic segments are composed of many genes (in addition to a constant region) - grouped into variable (V), diversity (D) and junction (J) genes - that encode the variable section of the final protein, which interacts with the antigen presented by the MHC complex (Figure 1.2A). During T cell development in the thymus, these genes are recombined through the action of RAG enzymes, which target recombination signal sequences to cleave DNA and join them - first D and J (if D is present), then (D)J and V. The insertion of additional non-templated nucleotides at the junctions can result in further variability. There are numerous V and J genes, which gives rise to a large number of possible V-J combinations, thus ensuring the diversity needed for antigen recognition by T cells. This is further augmented by differential combination of TCR chains in the final receptor. The activity of each receptor sub-unit is subject to selective pressures that ensure that it can functionally recognise and respond to foreign antigens, while being unresponsive to self-produced peptides and thus avoid auto-immune responses. In adaptive T cells, these receptors are composed of an α and a β chain. γ and δ chains also exist as a pair, but are less variable which results in a different type of response (Simoes et al., 2018).

The TCR is part of a larger membrane surface complex that assists in the recognition of the antigen being presented, as well as the APC presenting them (Figure 1.1B). T lymphocytes can thus be separated into two subsets with a shared developmental origin, bifurcating depending on the type of antigen-presenting Major Histocompatibility Complex (MHC) they can match. Consequently, each with their own APC matching capabilities and is easily identifiable by the expression of a surface protein that participates in this specific interaction. CD8-expressing T cells recognise antigens presented through MHC class I, which exists on the surface of almost all cells. This recognition elicits the maturation of CD8$^+$ T cells, preparing them for an anti-cellular response. This subset is accordingly also named cytotoxic, and through the use of perforins and granzymes they destroy cancer cells, as well as cells infected by intracellular pathogens (Halle et al., 2017).

CD4$^+$ T cells are the other lineage of T cells. Also known as T-helper (Th) cells, these lymphocytes are credited with the organisation of immune response, producing cytokines that serve as triggers or blockers of particular immune reactions (Luckheeram et al., 2012). Th cells recognise antigens presented by the MHC class II, present only on the membrane of dendritic cells, mononuclear phagocytes, some endothelial cells, thymic epithelial cells (important during T cell selection for functional,

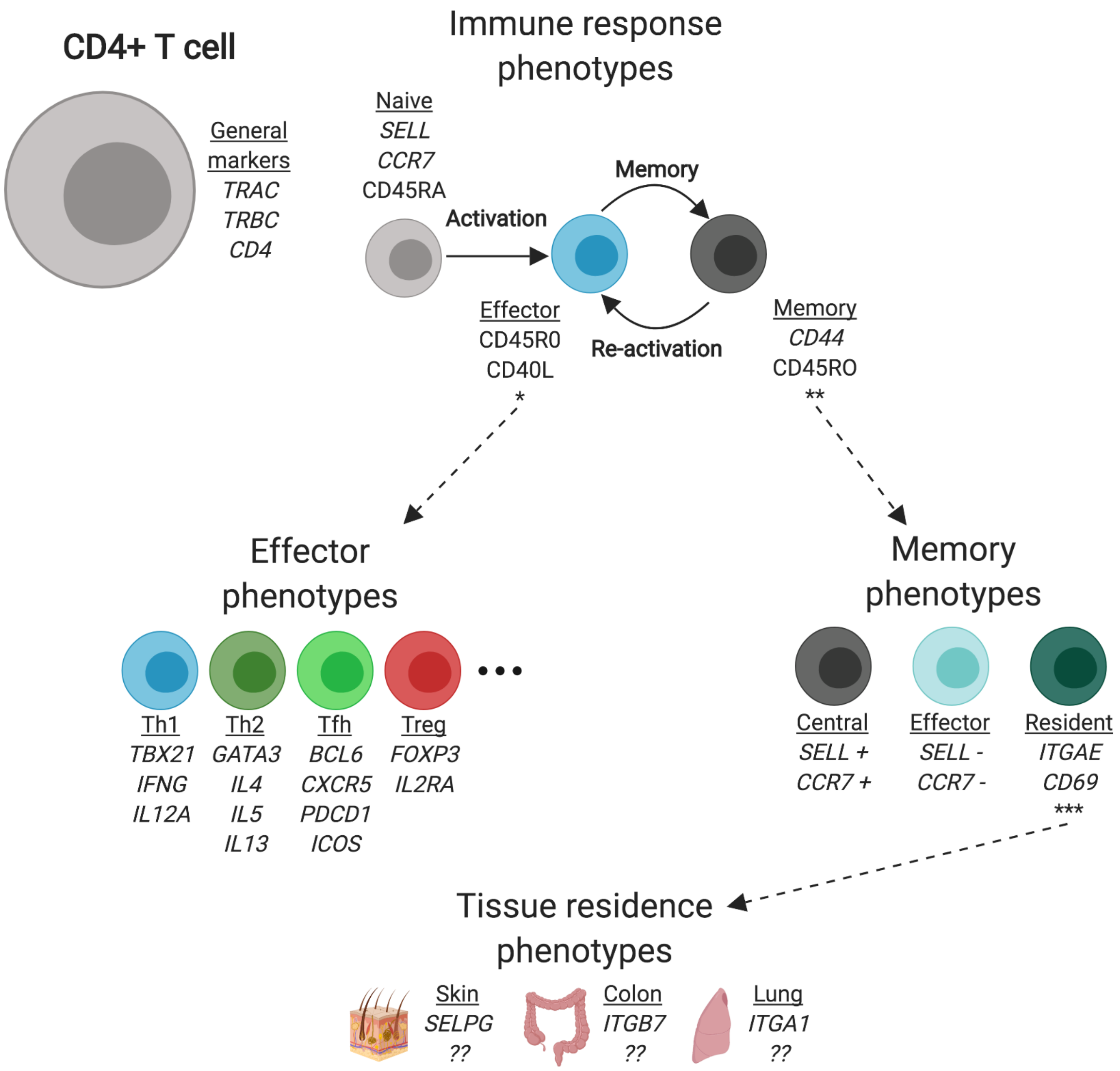

Fig. 1.3: An overview of known T-helper cell heterogeneity and key marker genes. Beyond their core markers, Th cells can be classified into based on different phenotypes that depend on stage of immune response, the type of effector function, the type of memory cell they form and their tissue of residence (a topic understudied comparatively to the rest). Question marks (??) represent unresolved phenotypes.

non-self-responding TCR), and B cells. This interaction, combined with signalling from the media where the cell is acting, induce an activation programme of the cell that is specific to the external threat being handled. CD4$^+$ T cells encompass a large transcriptional plasticity, which results in diverse related phenotypes (Figure 1.3). Th cells have classically been organised into various effector phenotypes based on their cytokine secretion profile (Mosmann et al., 1986; Schmitt and Ueno, 2015), allowing them to orchestrate the professional immune cells in the microenvironment. For instance, IFNγ production by Th1 cells has been identified as a key signalling molecule to combat intracellular parasites through the stimulation of macrophages, as well as class switch recombination of B cells to an IgG isotype. In turn, Th2 cells use IL-4 and IL-13 to stimulate basophils and mast cells to release granules against helminth invaders, and Th17 cells coordinate neutrophil recruitment by epithelial cells through IL-17A and IL-17F (Weaver et al., 2013). While diverse in function, these effector phenotypes are not the sole drivers of variability between Th cells, which also vary according to their activation state (naïve, effector, and memory cells) and with the host environment cues (tissue-specific phenotypes).

Upon finishing responding to an infection, T cells can go into a lowly replicative memory state in which they will save the TCR that drove the specialized response. The various memory states relate to the level of activation of the cell, but also to its tissue of residence. Cells expressing the chemokine receptor CCR7 are in a more naive, non-stimulated state, and also target lymphoid tissues like lymph nodes or the spleen, where most of antigen presenting to CD4$^+$ T cells takes place. In addition, tissue-homing and residency phenotypes exist, all of them characterised by the involvement of one or more chemokine receptors or adhesion molecules like integrins. Nonetheless, tissue-specificity in T-helper cells, and even more broadly in immune cells, is still generally understudied. Recent developments using single-cell high throughout methods have tackled this questions (Scott et al., 2018; Wong et al., 2016a), and it is expected that future efforts will rely on the accumulation of data to extract these patterns from cross-tissue samples.

Among the phenotypic variability of T-helper cells we can find the particular subset termed T-regulatory (Treg) cells. They are different from most Th cell subtypes in that, rather than boosting immune response, they are responsible for dampening it (Sakaguchi et al., 1995). This regulatory role in the immune system is of dire importance. Leaving the immune response unchecked can lead to destructive responses that will adversely affect the organism, as in autoimmune diseases. Treg cells were originally identified by their high expression of CD25, but as a subset they are more clearly defined by the expression of the FOXP3 transcription factor (Hori et al., 2003). Despite the focus on CD4$^+$ Treg cells here presented, CD8$^+$ cells can also have a regulatory phenotype, yet this are understudied compared to its CD4$^+$ counterpart (Yu et al., 2018).

Further subsets of Treg cells have been described, either related to the various parallel programmes that Th cells can adopt or their developmental origin. All T cells derive from a Common Lymphoid Progenitor cell that originates through haematopoiesis in the bone marrow, and travels via the bloodstream to mature in the thymus, where their TCR recombines and is tested for responsiveness to foreign antigens (positive selection) and against self-antigens (negative selection). However, natural Treg cells are derived from a subset of T cells with an intermediate level of response to self-antigens. This subset is further supplemented by induced Treg cells, which originate from other T-helper cells. While both natural and induced T-regulatory cells share a role, their distinct origins extend their TCR repertoire and thus their function (Zhang et al., 2014). Beyond this, Treg cells are also subject to memory and tissue-trafficking phenotypes like the remaining Th cells (Huehn et al., 2004), although these are not as well studied.

Immune cells are also described to have roles beyond defense against pathogens. These roles involve interactions with other non-immune tissues and mostly focus on their maintenance (Gordon and Martinez-Pomares, 2017; Laurent et al., 2017), and the immune system has also been described as relaying signals to the nervous system (Veiga-Fernandes and Mucida, 2016). Treg cells have been increasingly noted to be relevant, not just for their role in the immune system, but also for their functions beyond it. This regulatory subset has been shown to be involved in tissue repair (Li et al., 2018b) (chiefly muscle (Burzyn et al., 2013)), hair growth (Ali et al., 2017), and homeostatic regulation of gut microbiota (Cebula et al., 2013) and adipose tissue (Cipolletta, 2014; Sharma and Rudra, 2018). These functions, being widespread in the organism, consequently rely on an efficient trafficking and tissue localization scheme (Liston and Gray, 2014). Despite the importance of understanding how these migration and adaptation programmes are constituted and regulated (Agace, 2006), this aspect of the immune system is still incompletely understood.

1.5 Tissue-specific gene expression

Histological studies have uncovered many details of organ biology and physiology. Tissue staining is routinely used in pathology, and a better understanding of which molecules are markers of different tissue structures and cells in steady-state has resulted in important medical advancements.

Early studies in transcriptomics using microarrays dissected transcriptional responses to metabolic shifts (DeRisi et al., 1997) and disease (with a particular focus in cancer) (Rhodes et al., 2004), with homeostatic tissue sample comparison only appearing later (Shyamsundar et al., 2005).

RNA-sequencing has, from its inception, been linked to the unraveling of cross-organ and tissue differences (Mortazavi et al., 2008). Compared with preceding technologies, RNA-seq was capable of detecting a broader variety of transcripts in an unbiased way, along with high confidence splice junctions and allele-specific expression, with the added benefit of doing it for a lower cost (Wang et al., 2009). RNA-seq was quickly adopted and improved (see Section 1.2), extending its sensitivity and breadth of applications. Consortia were developed around the use of sequencing technologies for different biomedical purposes, often with RNA-seq taking a pivotal role (Lonsdale et al., 2013; The Cancer Genome Atlas Research Network et al., 2013; The ENCODE Project Consortium, 2012). These large collections of data were instrumental in revealing the functionality of genomic regions and relationships between samples. With data from the Genotype-Tissue Expression (GTEx) consortium, it was revealed how human tissues transcriptionally relate to each other, as well as what genes vary in expression across tissues and individuals (Melé et al., 2015). The Cancer Genome Atlas (TCGA) relied on RNA-seq, as well as other data modalities, from several cancer types to map the similarities between different tumours, and identify potentially important pathways for the treatment of those malignancies (Hoadley et al., 2018). Comparison between disease samples and steady-state can also be particularly informative, for example in understanding how tumours affect their adjacent tissue (Aran et al., 2017), or how tumour growth compares to developmental tissues and which pathways are involved (Young et al., 2018). In short, while large databases of expression data can serve as useful resources for broader applications by the scientific community, they can also be mined for emerging patterns.

Transcriptomic data can also be analysed beyond one species to gain understanding of the evolutionary links of gene expression programmes. Early microarray data analysis showed how human-chimpanzee divergence was especially accentuated when looking at brain RNA (Enard et al., 2002). Collection of samples from more species, combined with the use of RNA-seq, augmented the resolution of what gene expression changes could be observed (Brawand et al., 2011). Varying divergence rates for different tissues, gene groups and genomic regions, could be observed and associated to different selective pressures and tissue functions. Further studies have since compared other species (Li et al., 2014) or aspects of the transcriptome (Barbosa-Morais et al., 2012), revealing the intricate way evolution sculpted molecular programmes in different tissues across the tree of life, and defined the core genes involved in tissue function.

The functional associations observed between tissues are a consequence of the similarities and differences of the cell types that constitute them. These are mostly a result of the developmental processes giving rise to these tissues. For example, most tissues contain epithelial cells, marked by EPCAM, which share certain features such as forming barriers and secretory functions (Trzpis et al., 2007). Epithelial cells have been found to be vastly diverse within and between tissues, adopting different shapes and spatial arrangements (Wang et al., 2012), as well as further cytological changes adapted to the specific tissue biology.

Many aspects of tissue-specific heterogeneity stem from immune cells, perhaps owing to their mobility and plasticity. Various tissue-specific functions of Treg cells have been described above (Section 1.4). Macrophage heterogeneity represents another paradigmatic case of between-tissue phenotypic variability. In adult humans, circulating macrophages derive from bone marrow progenitors; in contrast, tissue-resident macrophages have been demonstrated to be developmentally related to haematopoietic progenitors in the yolk sac (Gomez Perdiguero et al., 2015). These macrophage subsets are important in mediating tissue immunity, while in parallel governing their homeostasis, such as synaptic pruning by microglia, heme recycling by splenic macrophages, or the pro-angiogenic role of Hofbauer cells at the maternal-fetal interface. Importantly, tissue-specific functions are a consequence of signalling in the local environment, which is capable of completely reprogramming macrophage chromatin, gene expression and function (Gosselin et al., 2014; Lavin et al., 2014), and consequently influence their response to tissue-specific injuries (Hoyer et al., 2019). This heterogeneity has also been detected within tissues, and in the gut has been associated with signalling provided by local neurons (Gabanyi et al., 2016). Single-cell RNA-sequencing has also been used to reveal cross-tissue conserved regulators of macrophage identity (Scott et al., 2018), and could in the future be used to further explore potential subpopulation heterogeneity and correlate it with gene expression spatial data to identify associations with specific anatomic locations within organs.

The application of scRNA-seq methods can extend these methods to comparisons between cell types, which results in larger scale comparisons, yet will open a window into how different programmes are specified for cell function in evolution and how they translate across species. It has recently been showed how variability in expression relates to evolution of innate immune response in fibroblasts (Hagai et al., 2018). Data from this study has been further used to test an artificial intelligence method that was capable of accurately predict species-specific responses solely based on the data from the remaining organisms sampled (Lotfollahi et al., 2018). As well as understanding evolutionary biology of cell types or immune responses, these types of studies and applications can have considerable impact in translating results from model organisms into the clinic.

1.6 Insights and scope of this thesis

Single-cell RNA-seq has revolutionized the profiling of cell type heterogeneity over the last decade. This has allowed for a deep, unbiased look into several organs and organisms, profiling hundreds of cell types at higher resolution. At the same time, progress has been made in computationally combining datasets for further analysis. As an increasing number of scRNA-seq datasets is produced, we come ever closer to a first draft of a transcriptional Human Cell Atlas, showcasing the full spectrum of cellular variety in our species.

The expansion in cell throughput is now permitting the study of smaller, rarer subpopulations. While specific cell types can still be sorted prior to sequencing for deeper profiling, unknown and underrepresented cell types will require larger numbers to be detected. This profound transcriptional portrayal of cells also often results in valuable resources that can be examined for functional targets of novel therapies and assays, which is especially true when studying immune cells. Developing directed cell therapies is a long-term goal of many medical fields, but a thorough knowledge of key cell types is still needed.

A transcriptional reference for cell types can be a key resource for those employing scRNA-seq. Having a ready-to-use resource that draws on the combined knowledge of the data generated would provide immediate assistance for automatic annotation of novel projects. Additionally, an exhaustive and integrated collection can be very informative about cell and tissue biology. However, the limits of this integration should also be tested and examined.

After this introductory chapter, Chapter 2 will show a deep dive into T-regulatory cell heterogeneity using single-cell RNA-seq. Treg cells have been shown to have critical roles in steady-state and disease, but it is still not fully understood which subpopulations fulfill which functions in different tissues, and how this heterogeneity relates to cross-tissue diversity. The chapter will describe Treg cell subpopulations detected in mouse in different tissues and how they compare to other resident T-helper cells. These subpopulations reflect different activation states, and form a phenotypic continuum between peripheral tissues (skin and colon) and their respective draining lymph nodes. The first sections will also discuss the limits of heterogeneity detection using scRNA-seq, especially when using two different protocols. Lastly, a mouse-to-human comparison will be presented, comparing conservation and divergence of gene programmes and Treg cell subpopulations.

Chapters 3 and 4 will focus on the use of broad scRNA-seq data collections to create informative references for automatic cell type annotation. Chapter 3 will detail the development of *CellTypist*, a pipeline to integrate diverse scRNA-seq datasets and cluster them into meaningful groups that approximate commonly defined cell identity, and the training of an updatable classifier that can be used to annotate new datasets. All annotation data available from these datasets is also collected, and the classifier train is also in itself informative. Following this, Chapter 4 will be centred on the dissection of a large collection of human scRNA-seq data. After application of *CellTypist*, it will explore how gene expression at the cell type level influences tissue similarity, as well as uncover the groups of genes characterising cell identity.

This thesis ends in Chapter 5, where I will be discussing the broader picture of the results reported in this thesis. This chapter will explore to what detail cell identity can be deconstructed, and what that means for informative automated annotation of new datasets, as well as to our understanding of cell biology and how they are categorized.