Employing core regulatory circuits to define cell identity

Nathalia Almeida†, Matthew W H Chung†, Elena M Drudi†, Elise N Engquist†, Eva Hamrud†, Abigail Isaacson†, Victoria S K Tsang†, Fiona M Watt* & Francesca M Spagnoli**

Abstract

The interplay between extrinsic signaling and downstream gene networks controls the establishment of cell identity during development and its maintenance in adult life. Advances in next-generation sequencing and single-cell technologies have revealed additional layers of complexity in cell identity. Here, we review our current understanding of transcription factor (TF) networks as key determinants of cell identity. We discuss the concept of the core regulatory circuit as a set of TFs and interacting factors that together define the gene expression profile of the cell. We propose the core regulatory circuit as a comprehensive conceptual framework for defining cellular identity and discuss its connections to cell function in different contexts.

Keywords cell identity; core regulatory circuit; GRN; regenerative medicine; transcription factor

Subject Categories Chromatin; Transcription & Genomics; Development; Stem Cells & Regenerative Medicine

DOI 10.15252/embj.2020106785 | Received 14 September 2020 | Revised 3 February 2021 | Accepted 4 February 2021 | Published online 5 May 2021

The EMBO Journal (2021) 40: e106785

Introduction

The nature of cell identity is a central problem in biology. Accurate identification of cell types deserves significant attention due to its impact on many areas of research and clinical applications, including regenerative medicine. Cell identities are influenced by external stimuli, such as signaling molecules, growth factors, and intercellular communication, which in turn affect downstream gene expression and jointly dictate cell phenotype and function(s) (Holmberg & Perlmann, 2012; Wagner et al, 2016). Even though these distinct facets of a cell’s identity are interdependent, they are often considered separately. Nevertheless, the cell’s phenotype and functional characteristics ultimately represent the readout of a specific gene-expression program. Typically, a small number of transcription factors (TF), which show a lineage-restricted expression pattern, are considered sufficient to establish gene expression programs that define the identity of a cell (Holmberg & Perlmann, 2012; Zaret & Mango, 2016). Often, these TFs have the ability to bind to inaccessible nucleosomal DNA, acting as “pioneer” TFs (Zaret & Carroll, 2011; Zaret & Mango, 2016).

The concept that differentiated cell identity is established and continuously maintained by a set of TFs was proposed several decades ago (Blau & Baltimore, 1991). This was supported by pioneering studies with cell hybrids and heterokaryons, in which terminally differentiated cells could be successfully reprogrammed into muscle cells by cell fusion (Weiss & Green, 1967; Blau et al, 1983; Pomerantz et al, 2009), and later by gain-of-function approaches based on key TFs (Davis et al, 1987). While these experiments established that cell identity is actively maintained by TFs, it was only in 2008 that Hobert proposed the term of terminal selector gene (TSG) (Hobert, 2008). A TSG was defined as a gene that specifies individual identities by directly controlling the expression of a set of downstream differentiation genes (a.k.a. effector genes) via common cis-regulatory motifs (a.k.a. terminal selector motifs) (Hobert, 2008). Though initially described within the context of neuron-specific lineage determination and maintenance in C. elegans (Etchberger et al, 2007), the existence of TSGs has been confirmed in a plethora of other cell types and also in vertebrate model systems (Hobert, 2008) (Box 1). Features of neuronal cell TSG expression that may well apply to other cell types are as follows: (i) the initiation and maintenance of TSG expression are independent events; (ii) the initiation may be the result of transient expression of distinct regulatory factors, either extrinsic signals or TFs; (iii) after initiation, TSGs autoregulate their expression, ensuring continuous expression and regulation of downstream targets (Hobert, 2008, 2011).

In higher vertebrate species, acquisition of a differentiated cell identity seems to require more complex circuitries, whereby a larger panel of TFs act in a combinatorial manner (Fig 1A) (Holmberg & Perlmann, 2012). Target/effector genes are not all controlled by a similar cis-regulatory logic, but instead different combinations of lineage-specific TFs co-regulate different subsets of target genes in distinct ways. Thus, only when the complete set of TFs is co-expressed in a cell, the full repertoire of differentiation genes is induced and maintained (Holmberg & Perlmann, 2012). Davidson pioneered the concept of gene regulatory networks (GRN) governing the development of body plan and organ formation in the embryo...
Efforts to classify neuronal identity have greatly contributed to our understanding of CRCs, with studies in C. elegans being the first to conceptualize various components of CRCs such as TSGs (Hobert, 2008). For example, PBX/CEH-20, part of the PBX TALE (three-amino-acid loop extension) homeodomain proteins (Selleri et al, 2019), was first identified to initiate and maintain the terminally differentiated state of dopaminergic (DA) neurons, thereby acting as a TSG (Doitsidou et al, 2013). It was later found that PBX factors, in particular Pbx1, have a conserved role in mouse midbrain DA neurons (Villaescusa et al, 2016). More recently, a genetic approach was used to specifically ablate Pbx1 expression in mouse DA neurons to achieve temporal control over its expression, confirming the involvement of Pbx1 in an evolutionarily conserved CRC (Remesal et al, 2020). This study not only confirmed the involvement of Pbx1 in the production of dopamine, but also showed that this TF is required for the expression of a broad range of olfactory bulb DA effector genes (Remesal et al, 2020). Such a genetic approach enabled the distinction between the late roles of Pbx1 in terminal differentiation and preservation of neuronal identity (Remesal et al, 2020) and its early activities in neuroblasts as well as in midbrain DA neuron specification (Grebbin et al, 2016; Villaescusa et al, 2016) (Box 1 Figure).

Box 1 Figure. The role of Pbx1 in the CRC of olfactory bulb DA neurons.
Pbx1 is a TF that is continuously expressed from progenitor to mature neurons. Conditional knockout approaches were key for elucidating the role of Pbx1 not only in the specification of midbrain DA neurons (Villaescusa et al, 2016), but also specifically in the CRC of olfactory bulb DA neurons (Remesal et al, 2020). DA: dopaminergic.

The transcriptional characterization of cell populations can be facilitated by the prior knowledge of TFs that promote cell identity and unfold a CRC network. For instance, the use of known DA lineage marker genes enabled Fernandes and colleagues to describe a previously unknown heterogeneity of DA neurons derived from human induced pluripotent stem cells (Fernandes et al, 2020). Using scRNA-seq to obtain an unsupervised clustering of the population of cells, the group identified six distinct cell types, two being neuron progenitor populations and four being subpopulations of DA neurons. Although these populations differed in expression of certain genes, all expressed typical DA lineage markers, including Pbx1. Additionally, the in vitro transcriptional data overlapped well with single-cell transcriptomic datasets of post-mortem substantia nigra, which validated the transcriptional heterogeneity found in subpopulations of human DA neurons.

Given that DA neurons are degenerated in individuals with Parkinson’s disease, building the CRC network in DA neurons will not only enrich our understanding of this cell type, but also, potentially, contribute to the development of disease therapies (see section “Assigning functional relevance to CRCs”).
Terminally circuitry” (CRC) (Boyer connected auto-regulatory loops have been termed “core regulatory pluripotency (Boyer thereby forming cross-regulated feed-forward loops that maintain bind to their own loci, but also mutually regulate one another, TFs, such as NANOG, SOX2, and OCT4, called core TFs, not only from embryonic stem cells (ESCs) have revealed that a small set of Davidson, 2010; Xia & Yanai, 2019). Furthermore, seminal studies for example, in the hematopoietic lineage (Graf & Enver, 2009; transcriptional control of binary fate choices in stem cell differentiation, processes, GRN circuit design has been applied to describe the tran- network (Peter & Davidson, 2016). Beyond early embryonic development. GRNs adapt in response to external signals and other influences during development, resulting in a series of different developing cell states. Once cells become terminally differentiated, the TF network becomes more stable and can be defined as a CRC, which is autoregulating and activates the expression of the terminal effector gene battery. While GRNs and CRCs can be identified using similar methods, studies of GRNs additionally benefit from lineage tracing and pseudotime analysis to account for their temporal aspect.

Figure 1. Cell identity is regulated by CRCs.

(A) Conceptualizing different types of information (e.g., transcriptomics, epigenomics) in the flow of biological information from DNA to function in order to shape our knowledge of CRCs. Downstream processes (purple), such as gene and protein expression, are routinely measured using transcriptomics and proteomics. Further downstream of this is cellular phenotype, a more complex readout which is measured using various assays and microscopy techniques. Factors that influence the CRC of a cell (green) include intrinsic factors and extrinsic factors, such as epigenetic memory and the external environment, respectively. While an overall flow of information is unidirectional (from top to bottom), many factors influence each other in more complex ways. (B) Model of cell identity being regulated by GRNs through development and CRCs in differentiated cells. We propose that CRCs define differentiated cell types and GRNs are temporally transient networks which drive cellular differentiation during development. GRNs adapt in response to external signals and other influences during development, resulting in a series of different developing cell states. Once cells become terminally differentiated, the TF network becomes more stable and can be defined as a CRC, which is autoregulating and activates the expression of the terminal effector gene battery. While GRNs and CRCs can be identified using similar methods, studies of GRNs additionally benefit from lineage tracing and pseudotime analysis to account for their temporal aspect.

[Davidson & Erwin, 2006]. TFs and transcriptional regulators are GRN components, and their target sites are the cis-regulatory DNA modules. Because each module is regulated by multiple TFs and each TF interacts with multiple modules, it is possible to represent developmental patterns of gene expression as an interlocking network (Peter & Davidson, 2016). Beyond early embryonic processes, GRN circuit design has been applied to describe the transcriptional control of binary fate choices in stem cell differentiation, for example, in the hematopoietic lineage (Graf & Enver, 2009; Davidson, 2010; Xia & Yanai, 2019). Furthermore, seminal studies from embryonic stem cells (ESCs) have revealed that a small set of TFs, such as NANOG, SOX2, and OCT4, called core TFs, not only bind to their own loci, but also mutually regulate one another, thereby forming cross-regulated feed-forward loops that maintain pluripotency (Boyer et al., 2005). The core TFs and their intercon- nected auto-regulatory loops have been termed “core regulatory circuitry” (CRC) (Boyer et al., 2005; Young, 2011).

Arendt et al (2016) further extended these models and introduced the concept of core regulatory complex (CoRC), whereby cell type-specific gene expression not only requires the activity of a specific combination of terminal selectors but also depends on their physical cooperativity. Based on such a model, the origin of a new cell type in evolution coincides with the occurrence of a unique CoRC, distinct from its evolutionary sister cell type (Arendt, 2008; Arendt et al, 2016). While the primary function of CRC or CoRC factors is to keep cells in a stable differentiated state, the notion of GRN describes a temporally hierarchical framework of gene expression that controls a differentiation process and adapts in response to external signals and other influences during development (David- son, 2010; Marioni & Arendt, 2017) (Fig 1B).

Since CRC factors provide the ultimate instructive “code” underlying the expression of the effector genes in differentiated cells, cellular identifiers, such as the functional output, cannot be consid- ered separately. Thus, the CRC concept might provide a standard- ized and comprehensive definition of a cell type, as the TF regulatory network, which is necessary for the induction and mainten- ance of cell type-specific gene expression program in differenti- ated cells.

In this review, we discuss how CRC TFs can be employed to define cell identity in the context of differentiation strategies, which can benefit regenerative medicine.

Identifying a core regulatory circuit

Several efforts have been made to identify individual components of cell type-specific CRCs (Graf & Enver, 2009; Xia & Yanai, 2019). To date, most of our knowledge is based on the use of expression profiles of core TFs as a proxy for CRCs. However, to build the network, transcriptomic data need to be integrated with chromatin analyses in computational models for protein–protein, gene–protein, and regulatory element interactions (Fig 1A).
Transcriptomics: from population to single-cell analyses

Cell type-enriched sets of TFs, the main components of CRCs, are still primarily discovered by transcriptome analyses (Xia & Yanai, 2019) (Fig 2A). Over the last decades, the shift from bulk transcriptomics to single-cell or single-nucleus RNA-sequencing (scRNA-seq and snRNA-seq, respectively) has started to provide new insights into gene modules underlying individual cell types (Menon, 2018). Moreover, these approaches in genomics and transcriptomics at a single-cell resolution have led to depositories such as the Human Cell Atlas (HCA). The HCA is a global initiative, which aims to create a comprehensive reference map of all human cell types based on their molecular profiles and their classical cellular descriptions (Regev et al, 2017). The purpose is to provide a unique identification of each cell type and a common framework for understanding biological processes in health and disease. Single-cell atlases are already available for adult human tissues, including the lung.

Figure 2. Methods employed to identify CRCs.

(A) Single-cell transcriptomics allows the identification of cell populations or states (top). Putative CRC components for these cell identities can be identified by defining the TFs and downstream genes enriched in these cells (bottom). (B) Epigenetic methods allow the identification of cis-regulatory elements that make up the CRC. Chromosome conformation capture (3C/HiC) identifies regions of DNA, which are in close contact with each other, potentially including enhancer–promoter interactions (left). ATAC/DNase- and ChIP-seq for histone modifications identify regions of open chromatin, which can be used to identify enhancers as well as promoters and actively transcribed genes (right). (C) Computational methods are used in multiple aspects of CRC identification. Clustering of single-cell transcriptomics data allows discovery of previously unknown cell types, while pseudotime analysis help identify transcriptional states when cell fate decisions along developmental trajectories are made (top). Several algorithms can make data-driven predictions of CRCs by analyzing TF co-expression and performing GRN inference (middle). Other relevant data supplied by users or deposited in databases can inform on CRC mechanisms (e.g., chromatin accessibility, promoter and enhancer states, TF-binding and protein–protein interactions) and be integrated to refine CRC predictions (bottom).
kidney, pancreas and liver, and sequencing of fetal tissues is also ongoing (https://data.humancellatlas.org/).

To date, transcriptome analyses have enabled the classification of specific mammalian brain cells in spatiotemporal and cell-type databases (Keil et al, 2018; Arlotta & Pașca, 2019). The spatial transcriptome atlas of the adult human brain from the Allen Human Brain Atlas (AHBA), for example, comprises histological analysis and comprehensive microarray profiling of nearly 900 neuroanatomically precise microdissected sites of the brain in two individuals (Hawrylycz et al, 2012). More recently, in 2014, the U.S. National Institutes of Health funded the BRAIN Initiative Cell Census Consortium (BICCC) (Ecker et al, 2017). The initiative combines ten pilot projects spanning multiple approaches, including single-cell omics and species (mice, rats, zebrafish, and humans) with the final goal to classify brain cell types based on integrated analysis of their molecular, anatomical, and physiological properties. The BICCC network works closely with the HCA to develop a comprehensive atlas of all cell types in the human body within a common coordinate framework (Regev et al, 2017). BICCC groups developed new technologies for profiling single neurons that identified new cell types or cell states in the nervous system (Tasic et al, 2016).

The availability of single-cell data has allowed the characterization of heterogeneous transcriptional profiles, context-dependent regulatory relationships, and functional interactomes with higher granularity (Aibar et al, 2017; Mohammadi et al, 2019). Kelley et al (2018) used scRNA-seq data to examine cell-type variations across brain regions in intact human tissue. This resulted in a robust strategy to define gene modules enriched in major neuronal subtypes, which they termed “core transcriptional identities” (Kelley et al, 2018; Menon, 2018).

Despite its many advantages, scRNA-seq techniques are susceptible to several influences, which can bias the results (Chen & Zhou, 2017; Keil et al, 2018). Several technical factors can introduce variations in the sequencing data; cell dissociation and suspension preparation may introduce technical noise; and stress to the cell-type viability could lead to alterations in the gene expression profiles (Ecker et al, 2017; Menon, 2018; Kelley et al, 2018). As some classes of cells are more fragile and prone to rupture than others, this will introduce bias in the populations captured. Other challenges include transcripts of short length or of low abundance in a single cell. The low amount of material may result in uneven RNA loss leading to gene drop-out events which can be difficult to measure accurately (Chen & Zhou, 2017; Keil et al, 2018). In particular, Mawla and Huising have illustrated the limitations of pancreatic islets transcriptomics where the impact of endocrine cells, other than the insulin-producing β-cells, or auxiliary cells in the disruption of blood glucose homeostasis is often overlooked due to their lower abundance (Mawla & Huising, 2019). Although whole islet analysis is limited by the mixture of cells, which differ in abundance, Bramswig and Kaestner discussed the reliability of adding a sorting strategy to determine cell type-specific changes (Bramswig & Kaestner, 2014).

Another challenge of developing a comprehensive human cell atlas is that scRNA-seq requires fresh tissue and therefore relies on limited tissue donations collected either surgically or post-mortem (Ecker et al, 2017; Kelley et al, 2018). A valuable alternative is snRNA-seq which can be applied to archived frozen samples and provides less biased cellular coverage (Bakken et al, 2018). In fact, Lake et al (2016) revealed 16 neuronal subtypes using nuclear RNA from single nuclei harvested from post-mortem tissue, demonstrating snRNA-seq as a promising method to analyze the human brain. Similarly, snRNA-seq overcame the technical problems due to rapid enzymatic RNA degradation upon resection of pancreatic tissue, which historically have led to underrepresentation of exocrine cells and hampered comprehensive sequencing of human exocrine pancreatic cells (Tosti et al, 2021).

An additional challenge in single-cell transcriptomics is to classify cell variability, to define cell “types” and to distinguish them from transient cell “states”. A consensus on whether a cell going through different states should still be considered the same cell type has not yet been achieved. Xia and Yanai proposed a “periodic table” approach to distinguish cell types from cell states. Typically, scRNA-seq analysis relies on unsupervised clustering algorithms based on the differential expression of genes to identify the cell types (Xia & Yanai, 2019). This uncovers modules of genes and provides an initial map of the relative proportions of different cell types (Regev et al, 2017; Menon, 2018). However, clustering based on differential gene expression might overlook the fact that cell states, such as the cell cycle or stress, are also captured (Kiselev et al, 2019). By contrast, by defining cell identity using the concept of CRCs, a given cell is expected to show a unique set of TFs regardless of its state, which would help to distinguish between cell types and cell states. Hence, Xia & Yanai propose a cell clustering approach that combines both differentially expressed genes and the expression profile of TFs (Xia & Yanai, 2019). This represents a practical approach for distinguishing cell states within the cluster of a given identity.

**Epigenetic modifications and chromatin landscapes**

Defining CRC factors and building a network requires elucidation of the relationships between the regulators of gene expression (TFs) and the target genes (effector genes). TFs activate or inhibit the expression of genes by binding specific regulatory sequences, including promoters and enhancers (Spitz & Furlong, 2012). Identifying the enhancers that regulate genes of interest or are bound by key TFs is therefore crucial to understand the connections between the players in the CRC. As enhancers cannot be uniquely characterized by a particular sequence or feature (Coppola et al, 2016), they are identified using multiple approaches combined (Fig 2B).

Coordinated experiments interrogating transcriptional responses and chromatin binding via chromatin immuno-precipitation with next-generation sequencing (ChIP-seq) can offer insights into different levels of gene regulation, TF-binding motifs, DNA and chromatin modifications, and how each component is coupled to a functional output (Holmberg & Perlmann, 2012; Wilson & Filipp, 2018). Examples of CRCs in specific lineages are included in Box 1 and Box 2.

The majority of enhancers, in order to influence gene expression, are located in proximity to their target gene’s promoter. Pairs of genomic loci which are nearby in 3D space can be identified using chromosome conformational capture (3C) (Dekker et al, 2002) or Hi-C (Belton et al, 2012). More conveniently, the genome can be scanned for accessible chromatin regions. Accessibility can be assayed by DNase-seq (Boyle et al, 2008) or ATAC-seq ( Buenrostro et al, 2015), which work by partial DNA digestion or transposases,
Mistour knowledge of CRCs. The function of Mist1 and Ptf1a in acinar tissue has been established thanks to mouse genetic studies (Krapp et al, 1996; Lemercier et al, 1997; Pin et al, 2001). Together, Mist1 and Ptf1a bind and drive the transcription of over a hundred downstream acinar genes through reiterated feed-forward regulatory loops (Jiang et al, 2016). However, the depth and nature of these TFs’ involvement in acinar cell identity was not understood until more recently when a combination of epigenetic and transcriptomic analyses revealed that they are part of a CRC (Jiang et al, 2016). ChIP-seq analysis revealed that Mist1 and Ptf1a share many target genes with highly juxtaposed binding sites. Ptf1a drives expression of Mist1 through binding to its enhancer, thus generating a self-sustaining regulatory loop between the two factors capable of maintaining not only itself, but also expression of effector genes essential for acinar cell identity.

Within the endocrine compartment of the pancreas, loss-of-function experiments also uncovered the roles of potential CRC constituents [comprehensively reviewed in (Romer & Sussel, 2015)]. Specifically, the development of insulin-producing β-cells depends on several TFs such as Pdx1, Ngn3, and Nkx6.1 (Murtaugh, 2007; Best et al, 2008; Arda et al, 2013; Romer & Sussel, 2015; Jennings et al, 2015). While some of these developmentally crucial TFs are also members of the CRC governing terminal β-cell identity, additional TFs such as MafA and MafB are required to maintain the mature β-cell phenotype through regulation of downstream effector genes involved in β-cell function (Kataoka et al, 2002; Matsuoka et al, 2004; Nishimura et al, 2015; Zhu et al, 2017; Russell et al, 2020).

scRNA-seq studies have unveiled a remarkable heterogeneity within mouse and human β-cells (Baron et al, 2016; Muraro et al, 2016; Segerstolpe et al, 2016; Xin et al, 2016; Lawlor et al, 2017; Mawla & Huising, 2019), which has further contributed to our understanding of these cell types. Wang and colleagues have taken advantage of single-cell transcriptomic data to model the relationship between eight master TFs (Pdx1, Ngn3, Arx, Ngn1, Hes1, Sox9, Hes1, Arx, Ngn3, and Pax4) in the pancreatic cell lineage (Wang et al, 2020). An adaptive landscape was constructed in which states were annotated either as mature or progenitor cell types based on prior knowledge of the relationships between these factors (Wang et al, 2020). The model infers additional transition states along different pancreatic lineage trajectories as well as previously unrecognized progenitors characterized by distinct CRC systems (Wang et al, 2020).
Chromatin accessibility can also be detected indirectly by searching for associated histone or DNA modifications using ChIP-seq (Creyghton et al., 2010; Rada-Iglesias et al., 2011). High levels of histone H3 lysine-27 acetylation (H3K27ac) typically mark active proximal and distal (e.g., enhancers) regulatory elements, while monomethylated H3 lysine-4 (H3K4me1) marks primed or active enhancers in the absence or presence of H3K27ac, respectively (Shlyueva et al., 2014). This is a general phenomenon reported in various cell types, including human and mouse ESCs undergoing differentiation (Creyghton et al., 2010; Rada-Iglesias et al., 2011).

Tiwari et al. (2018) integrated transcriptomic and epigenomic analyses to delineate gene regulatory programs that drive the developmental trajectory from mouse ESCs to astrocytes. By examining H3K4me1 enrichment patterns at stage-specific H3K27ac sites during astrogliogenesis, they were able to define regulatory elements unique to each stage. Next, by inferring the most highly associated TF-binding motifs at these elements, they unveiled drivers of the underlying differentiation trajectory. In this way, NFIA and ATF3 were identified as drivers of astrocyte differentiation from neural precursor cells, while RUNX2 promotes astrocyte maturation (Tiwari et al., 2018). Another histone modification, trimethylated H3 lysine-4 (H3K4me3), is commonly used to identify promoters (Guenther et al., 2007). The number and range of histone modifications that can be assessed is limited by the availability of appropriate antibodies (Satterlee et al., 2010), which can be of variable quality (Park, 2009). Another limitation of ChIP-seq is the requirement for cross-linking the DNA, which can cause epitope masking and technical artifacts (Satterlee et al., 2010; Baranello et al., 2016) meaning a large amount of starting material is required for accurate results, typically 10 million cells (Park, 2009).

High levels of H3K27ac, together with high abundance of TFs, transcriptional coactivators, and chromatin remodelers binding characterize a class of regulatory elements that have been termed super-enhancers (Hnisz et al., 2013; Whyte et al., 2013; Mooothy et al., 2017). These are major regulatory components of the gene expression program that shapes cell identity (Wang et al., 2019). Core TFs typically bind super-enhancers of their own genes, positively regulating their own expression, as well as the super-enhancers of many other cell type-specific genes, thereby establishing an interconnected regulatory network (i.e., CRC) (Hnisz et al., 2013; Whyte et al., 2013). Saint-André and colleagues have mined super-enhancers as an unbiased approach to identify core TFs in human ESCs, creating a map of the transcriptional regulatory circuitry involved in pluripotency and other cell lineages (Saint-André et al., 2016). Integrated and interactive databases of super-enhancers for human and mouse genomes have been made available as resources (Khan & Zhang, 2016; Qian et al., 2019; Jiang et al., 2019). Super-enhancers have been recently identified as having a unique role in transcriptional regulation in cancer (Bradner et al., 2017). As oncogenic events can go hand in hand with a loss of cell identity, further investigation of transcriptional dysregulation in cancer may shed light on transcriptional programs in normal cells and unveil cell type-specific CRCs (Bradner et al., 2017).

Similar to transcriptomes, epigenetic signatures can be detected at single-cell level resolution. For example, single-cell ATAC-seq (scATAC-seq) can reveal cell-specific regulatory signatures characteristic of CRCs (Fullard et al., 2018). Single-nucleus ATAC-seq of the mouse forebrain has identified cell type-specific genomic elements, many of which are distal enhancer elements (Preissl et al., 2018). TFs that bind these elements are candidate master regulators of different neuronal identities (Preissl et al., 2018). Recent studies based on single-nucleus methylomes have also expanded the atlas of brain cell types and identified regulatory elements that drive conserved brain cell diversity (Luo et al., 2017). Taken together, these studies highlight how studying chromatin modifications across different cell types can help identify candidate CRCs.

Computational approaches to predict CRCs

Advances in high-throughput sequencing technologies have led to the development of multiple computational algorithms designed to predict candidate core TFs and map CRCs (Fig 2C). Some approaches allow the integration of data using gene–gene, protein–protein, gene–protein, and regulatory element interactions and provide resources and insights into basic principles governing transcriptional regulatory networks (Neph et al., 2012; Rolland et al., 2014; Saint-André et al., 2016; Khan & Zhang, 2016; Qian et al., 2019; Jiang et al., 2019; Moore et al., 2020).

Computational methods have been established to predict the minimum combination of TFs required for inducing changes in cell identity as well as improving the efficiency of reprogramming to pluripotency (Cahan et al., 2014; D’Alessio et al., 2015; Rackham et al., 2016; Biddy et al., 2018; Nicetto & Zaret, 2019; Schiebinger et al., 2019). Based on just transcriptomic data, a simple method to select TFs is to calculate expression specificity for each cell type against multiple cell types. This measurement provides information about transcriptional control of cell identity and candidates can then be verified by reprogramming experiments (D’Alessio et al., 2015). However, this prediction may include false positives due to oversimplification. Other computational algorithms have tackled this problem by inferring GRNs. These algorithms explore and model the relationships between TFs and target genes based on the expression patterns across samples. For example, ARACNe is an algorithm which identifies potential TFs and putative target genes by “mutual
information”, a measure of mutual dependence in information theory (Basso et al, 2005). GENIE3 is a machine learning method that infers GRNs by learning the complex co-expression relationship of TFs and candidate target genes (Huynh-Thu et al, 2010). It weighs TFs by their ability to predict expression of target genes and construct a TF network with the highest weights. A common limitation to GRNs inference is the requirement for large amounts of gene-expression data. The data complexity, partially in the form of expression variability, is key to constructing gene-gene relationships, such as co-expression. This variability is assumed to be representative of perturbations in cells of identical types. Experimental conditions need therefore to be carefully designed to examine such biologically relevant variability without compromising cell identity regulation. Moreover, the mechanisms of inferred networks need to be extensively validated in experiments. The recent accumulation of reliable data of multiple omics-types provides readily available information to direct mechanistic studies.

Multi-omics data can also be integrated during CRC inference to make supervised predictions. Ideally, this type of data is generated from a single purified population of cells. Transcriptomic data inform TF-target gene relationships and co-factor expression, which may indicate phenotypic specificity of the regulatory complex. TF activity on target genes can be inferred from ChIP-seq peak profiles, especially when they overlap with active promoters (H3K4me3 peaks) and enhancers (H3K27ac peaks). Chromosome conformation capture technologies can provide a basis for cell type-specific predictions of enhancer-promoter interactions. Also, TF activity can be studied by looking at overlapping peaks mapped by DNase-seq or ATAC-seq. Some of these epigenetic data are available in databases, such as ENCODE (Dunham et al, 2012), Roadmap Epigenetics (Roadmap Epigenomics Consortium et al, 2015), Blueprint (Martens & Stunnenberg, 2013), and GeneHancer (Fishilevich et al, 2017), especially for stable cell lines. Furthermore, experimental protein–protein interaction information curated in databases, such as STRINGDB (Szklarczyk et al, 2018), IntAct (Orchard et al, 2014), and BioGrid (Oughtred et al, 2018), enable consideration of co-factors and TF complexes in building a CRC. While acquiring deep multi-omics data is not yet technologically feasible, computational algorithms have been designed to integrate data from various databases with minimum user input for CRC inference. SCENIC infers TF GRNs using GENIE3 and applies RcisTarget to further reduce false-positive TF-target gene relationships in the network by performing cis-regulatory motif analysis (Aibar et al, 2017). While intended for single-cell data, it is applicable to any transcriptomic datasets with sufficient size and complexity. The recently developed computational method Multi-omics network inference (Moni) integrates TF ChIP-seq data, protein–protein interactions, enhancer–promoter interactions, and reference RNA-seq data from databases, along with individual user input of cell type-specific datasets to identify TFs and co-factors, and eventually reconstructs enhancer-promoter GRNs (Jung & del Sol, 2020). Although this approach is limited by the multi-omics data available for each cell type, the construction of precise CRCs will become more feasible as “omics” technologies mature and become more affordable. Advances in computational methods will need to be pursued in parallel to address challenges associated with data integration—to link data from heterogeneous sources and different measurement types with increased complexity and perform correction of batch effects.

The availability of single-cell data has inspired novel approaches to integrate data. Schiebinger et al (2019) applied the mathematical concept of optimal transport, which efficiently computes a distance between distributions, to scRNA-seq profiles of mouse embryonic fibroblasts undergoing reprogramming to pluripotent stem cells. TFs predictive of various fates were inferred and then experimentally tested. Among others, the homeobox Obox6 was found to correlate strongly with the pluripotent cell state and when combined with OCT4, KLF4, SOX2, and MYC factors, it enhanced reprogramming efficiency (Schiebinger et al, 2019). However, as previously mentioned, while single-cell transcriptomics allow identification of new cell types and states, an immediate challenge is to establish and validate consensus assignment of cell types and states, and to standardize experimental procedures to generate comparable results. Future techniques that enable collection of multiple data types from single cells or a highly homogenous population will (1) enhance the granularity and precision of CRC inference methods, (2) help to gain deeper insights into the complexity of CRCs, such as the hierarchy and molecular mechanisms of regulation, and (3) enable characterization of variability of CRCs across time, cell states and types.

In developmental systems, the concept of cell state refers to cell fate transition along a particular developmental trajectory (Kester & van Oudenaarden, 2018). With time included as a variable, more advanced computational methods are needed that can integrate data across a time scale. An example is dynGENIE3, which infers dynamical GRN models from time series expression data (Huynh-Thu & Geurts, 2018). These methods are particularly useful for studying changing GRN along the developmental trajectory. Furthermore, cross-sectional single-cell data that captures continuous developmental stages can be reconstructed into pseudo time-series. More than 50 bioinformatics tools have been developed for pseudotime analyses in scRNA-seq data (Saelens et al, 2019; Tritschler et al, 2019). Beyond embryonic development, single-cell pseudotime approaches provide powerful means for identification of differentiation trajectories in adult stem cell compartments as well as in disease (Tritschler et al, 2019). Since pseudotime analysis orders cells according to their overall transcriptomic similarity (Trapnell et al, 2014), it should be interpreted as a relative measure of cellular differentiation state, or maturity, rather than one on an absolute time-scale. Once such a trajectory is inferred, the transcriptional cell states at which fate decisions are made and the TFs driving these decisions can potentially be identified. For example, pseudotime series can be used to calibrate models such as non-linear ordinary differential equation models (Ocone et al, 2015) that capture CRC dynamics. It should be noted that pseudotime approaches cannot replace traditional cellular lineage tracing techniques (Kester & van Oudenaarden, 2018) but the two are complementary and should be used to validate each other.

Assigning functional relevance to CRCs

Transcriptomics, epigenetics, and computational approaches can be used to predict CRC components and architecture. However, functional experiments, such as genetic perturbation and reprogramming studies, provide validation of these predictions. An important challenge is to match transcriptional and functional profiles of a cell
population. In some cases, cells with apparently identical CRCs exhibit disparate functions. A compelling example of this is the existence of dynamic, interchangeable states between pancreatic β-cells, which are likely to be controlled at multiple levels and influenced by the pancreatic islet microenvironment (Dominguez-Gutierrez et al, 2019); here, the existence of multiple combinations of states may correlate with varying levels of insulin secretion. Diverse calcium responses to glucose stimulation are also found among β-cells, and this has recently been shown to further fluctuate when the cells are detached from their host islets (Scarl et al, 2019). Functional differences of this kind have been reconciled in the context of neuronal subtypes possessing the same CRC, where the activity of terminal selectors was demonstrated to vary in the presence of repressor proteins confined to a specific cell subtype, thereby curbing the expression of the terminal gene battery (Hobert, 2016). Consequently, there exist limitations when considering cell identity exclusively from a molecular or functional perspective and the CRC should not be considered in isolation.

The concept of CRC not only provides a more comprehensive way of defining cell identities but might also have direct implications in regenerative medicine. For example, the knowledge of core TFs and CRCs underlying a desired cell identity may have a direct impact in lineage reprogramming and advance its clinical translation. Indeed, direct lineage reprogramming represents a strategy for generating desired functional cells that can be used in cell therapies (Heinrich et al, 2015), as the idea underlying successful lineage reprogramming is based on the knowledge of transcriptional networks governing cellular identity (Graf & Enver, 2009; Heinrich et al, 2015). A lot of progress has been made in this field, since it is now possible to obtain an array of different cellular types from distinct mature populations (Zhou et al, 2008; Vierbuchen et al, 2010; Heinrich et al, 2015; Xu et al, 2015). Direct lineage reprogramming unfolds developmental programs and argues for the engagement of hierarchical developmental CRCs, providing another strategy for discovering hierarchical developmental CRCs, providing another strategy for discovering hierarchical developmental CRCs.

Concluding remarks

We propose that the CRCs provide a comprehensive and uniform framework for defining the identity of a cell. The significant increase in our understanding of gene expression, particularly from single-cell datasets, underscores the need to unify and integrate new information with prior knowledge. At the same time, these new insights have reopened the definitions of a cell’s identity. While resources such as the HCA has given us access to information about each cell of the human body, additional functional studies are required to build a complete map of CRCs. In particular, the knowledge inferred from all the available datasets has to be tested in human models using high-throughput approaches, such as CRISPR-based screening platforms in human pluripotent stem cells. Large-scale observational studies offer another way to assess the relevance of suspected CRC components in a human setting. For example, the Human Knockout Project (https://www.broadinstitute.org/cardiovascular/human-knockout-project) studies loss-of-function phenotypic consequences in naturally occurring human genetic variants. Besides assessing the functional relevance of TFs, a fundamental challenge is to combine the transcriptome and epigenomic characterization of individual cell types with concurrent CRISPR-based genome and enhancer-targeting editing approaches. Only such a systems approach will elucidate if a CRC is self-sustaining and drives the expression of genes, which maintain unique cellular traits. Finally, future efforts will be directed to integrate core TFs with cell behavior and function into a more comprehensive concept of cell regulatory networks.

Acknowledgements

Our apologies go to all authors whose important work could not be mentioned due to space limitations. We gratefully acknowledge the financial support of the Wellcome Trust PhD program “Cell therapies and regenerative medicine” (108874/2/15/2). We thank Dr. Fay Minty for her support of the program.

Conflict of interest

F.M.W. is currently on secondment as executive chair of the UK Medical Research Council. The other authors declare that they have no conflict of interest.

References

Abar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J et al (2017) SCENIC: Single-cell regulatory network inference and clustering. Nat Methods 14: 1083 – 1086

Ardha HE, Benitez CM, Kim SK (2013) Gene regulatory networks governing pancreas development. Dev Cell 25: 5 – 13

Arendt D (2008) The evolution of cell types in animals: Emerging principles from molecular studies. Nat Rev Genet 9: 886 – 882

Arendt D, Musser JIM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlösser G, Widder S, Laubichler MD et al (2016) The origin and evolution of cell types. Nat Rev Genet 17: 744 – 757

Arliotta P, Pašca SP (2019) Cell diversity in the human cerebral cortex: from the embryo to brain organoids. Curr Opin Neurobiol 56: 194 – 198

Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Avermann B, Barkan E, Bertagnolli D, Casper T, Dee N et al (2018) Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One 13: e0209648

Baranello L, Kouzine F, Sanford S, Levens D (2016) Chip bias as a function of cross-linking time. Chromosoma Res 24: 175 – 181

Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM et al (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 3: 346 – 360.e4

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37: 382 – 390

Belon JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 58: 268 – 276

Best M, Carroll M, Hanley NA, Piper Hanley K (2008) Embryonic stem cells to beta-cells by understanding pancreas development. Mol Cell Endocrinol 288: 86 – 94

Biddy BA, Kong W, Kamimoto K, Guo C, WAYE SE, Sun T, Morris SA (2018) Single-cell mapping of lineage and identity in direct reprogramming. Nature 564: 219 – 224

Blau HM, Baltimore D (1991) Differentiation requires continuous regulation. J Cell Biol 112: 781 – 783
Blau HM, Chiu CP, Webster C (1983) Cytoplasmic activation of human nuclear genes in stable heterocaryons. Cell 52: 1171–1180
Boyer LA, Tong IJ, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947 – 956
Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311 – 322
Bradner JE, Hnisz D, Young RA (2017) Transcriptional addiction in cancer. Cell 168: 629 – 643
Bramswig NC, Kaestner KH (2014) Transcriptional and epigenetic regulation in human islets. Diabetologia 57: 451 – 454
Buenrostro JD, Wu B, Chang HY, Greenleaf WJ, Dierckx RA, Sheffield DN, Lee TI, Johnson AK, Simon IS,-sectional enhancers and target genes in GeneCards. Database (Oxford) 2017: 1 – 17
Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, Motj J, Ehrlich ME, Hurl YL, Roussos P (2018) An atlas of chromatin accessibility in the adult human brain. Genome Res 28: 1243 – 1252
Graf T, Enver T (2009) Forcing cells to change lineages. Nature 462: 587 – 594
Grebbin BM, Hau AC, Grob A, Anders-Maurer M, Schramm J, Koss M, Wille C, Mittelborm M, Selleri L, Schultz D (2016) PBX1 is required for adult subventricular zone neurogenesis. Dev. 143: 1281 – 1291
Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77 – 88
Hawrylycz MJ, Lein ES, Guillemot-Bongaarts AL, Shen EH, Ng L, Miller JA, Van De Lagemaat LN, Smith KA, Ebbert A, Riley ZL et al. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489: 391 – 399
Heinrich C, Spagnoli FM, Berninger B (2015) In vivo reprogramming for tissue repair. Nat Cell Biol 17: 204 – 211
Hnisz D, Abraham BJ, Lee TI, Lau A, Muley V, Sigova AA, Hoke HA, Young RA (2013) Transcriptional super-enhancers connected to cell identity and disease. Cell 155: 934 – 947
Hobert O (2008) Regulatory logic of neuronal diversity: Terminal selector genes and selector motifs. Proc Natl Acad Sci USA 105: 20067 – 20071
Hobert O (2011) Regulation of terminal differentiation programs in the nervous system. Annu Rev Cell Dev Biol 27: 681 – 696
Hobert O (2016) Terminal selectors of neuronal identity. Curr Top Dev Biol 116: 455 – 475
Holmberg J, Perlmann T (2012) Maintaining differentiated cellular identity. Nat Rev Genet 13: 429 – 439
Huynh-Thu VA, Geurts P (2018) dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. Sci Rep 8: 3384
Huynh-Thu VA, Irmrath A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS One 5: e12776
Jung S, del Sol A (2020) Multiomics data integration unveils core transcriptional regulatory networks governing cell-type identity. npj Syst Biol Appl 6: 1 – 4
Kataoka K, Han S, Shioda S, Hirai M, Nishizawa M, Handa H (2002) MaFA is a glucose-regulated and pancreatic beta-cell-specific transcriptional activator for the insulin gene. J Biol Chem 277: 49903 – 49910
Keil JM, Qaliev A, Kwan KY (2018) Brain transcriptome databases: a user’s guide. J Neurosci 38: 2399 – 2412
Fernandes HJr, Patikas N, Foskolou S, Field SF, Park JE, Byrne ML, Bassett AR, Metzkoop E (2020) Single-cell transcriptomics of Parkinson’s disease human in vitro models reveals dopamine neuron-specific stress responses. Cell Rep 33: 108263
Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iyng Stein T, Rosen N, Kohn A, Twik M, Safra M et al. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017: 1 – 17
Kelley KW, Nakao-Inoue H, Molofsky AV, Oldham MC (2018) Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nat Neurosci 21: 1171–1184

Kester L, van Oudenaarden A (2018) Single-cell transcriptomics meets lineage tracing. Cell Stem Cell 23: 166–179

Khan A, Zhang X (2016) DbsUPER: a database of Super-enhancers in mouse and human genome. Nucleic Acids Res 44: D164–D171

Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20: 273–282

Krapp A, Knoffler M, Frutiger S, Hughes CJ, Hagenblüche O, Mandanian J, Hopp U, Johnson C, Konieczny SF (2017) The p48 DNA-binding subunit of transcription factor PTF1 is a new exocrine pancreas-specific basic helix-loop-helix protein. EMBO J 15: 4317–4329

Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung HL, Shen S et al (2016) Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 352: 1586–1590

Lawlor N, George J, Bolisetty M, Kursawre R, Sun L, Sivakamasundari V, Kycia I, Robson P, Stitzel ML (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res 27: 208–222

Lemercier C, To RQ, Swanson BJ, Lyons GE, Konieczny SF (1997) Mist1: a novel basic helix-loop-helix transcription factor exhibits a developmentally regulated expression pattern. Dev Biol 182: 101–113

Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, Castanon R, Lucero J, Nery JR, Sandoval JP et al (2017) Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 604: 600–604

Marioni JC, Arendt D (2017) How single-cell genomics is changing evolutionary and developmental biology. Annu Rev Cell Dev Biol 33: 537–553

Martens JHA, Stunnenberg HG (2013) BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98: 1487–1489

Matsuoka TA, Artner I, Henderson E, Means A, Sander M, Stein R (2004) The Mafa transcription factor appears to be responsible for tissue-specific expression of insulin. Proc Natl Acad Sci USA 101: 2989–2993

Mawla AM, Huising MO, Mawla AM, Rawles C, Jafri M, Gorkin DU, Zhang Y, Sos BC, Preissl S, Michael Rukstalis J, Johnson C, Konieczny SF (2001) The bHLH transcription factor Mist1 is required to maintain exocrine pancreas cell organization and acinar cell identity. J Cell Biol 155: 519–530

Pomerantz JH, Mukherjee S, Palermo AT, Blau HM (2009) Reprogramming to a muscle fate by fusion recapitulates differentiation. J Cell Sci 122: 1045–1053

Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE et al (2018) Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat Neurosci 21: 432–439

Qiao FC, Li XC, Guo JG, Zhao JM, Li YY, Tang ZD, Zhou LW, Zhang J, Bai XF, Jiang Y et al (2019) SEanalysis: a web tool for super-enhancer associated regulatory analysis. Nucleic Acids Res 47: W248–W255

Rackham OJ, Jaffes J, Fang H, Oates ME, Holmes ML, Knaupp AS, Suzuki H, Neff Z, Dau B, Shin JW et al (2018) A predictive computational framework for direct reprogramming between human cell types. Nat Genet 48: 331–335

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysoczka J (2012) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470: 279–285

Regev A, Teichmann SA, Lander ES, Amit I, Bondyra S, Birney E, Bodenmiller B, Campbell P, Carninci P, Catzeflis M et al (2017) The human cell atlas. eLife 6: e27041

Remesal L, Roger-Baynat I, Chirivella L, Maicas M, Brocal-Ruiz R, Perez-Villalba A, Cucarella C, Casado M, Flames N (2020) PBX1 acts as terminal selector for olfactory bulb dopaminergic neurons. Development 147: dev186841

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilinkny M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317–329

Rolland T, Tsaun M, Charloteaux B, Pevzner S, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R et al (2014) A proteome-scale map of the human interactome network. Cell 159: 1212–1226

Nicetto D, Zaret KS (2019) Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. Curr Opin Genet Dev 55: 1–10

Nishimura W, Takahashi S, Yasuda K (2015) Mafa is critical for maintenance of the mature beta cell phenotype in mice. Diabetologia 58: 566–574

Ocone A, Haghverdi L, Mueller NS, Theis FJ (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics 31: i89–i96

Orchard S, Ammari M, Aranda B, Breuza L, Bragianti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N et al (2014) The MintAct project-IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42: D358–D363

Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O’Donnell L, Leung G, McAdam R et al (2018) The BioGRID interaction database: 2019 update. Nucleic Acids Res 47: S29–S54

Park P (2009) Applications of next-generation sequencing: ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669

Peter IS, Davidson EH (2016) Implications of developmental gene regulatory networks inside and outside developmental biology. Curr Top Dev Biol 117: 237–251

Philpott M, Cribs AP, Brown T, Oppermann U (2020) Advances and challenges in epigenomic single-cell sequencing applications. Curr Opin Chem Biol 57: 17–26

Pin CL, Michael Rukstalis J, Johnson C, Konieczny SF (2001) The bHLH transcription factor Mist1 is required to maintain exocrine pancreas cell organization and acinar cell identity. J Cell Biol 155: 519–530

Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE et al (2018) Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat Neurosci 21: 432–439

Qian FC, Li XC, Guo JG, Zhao JM, Li YY, Tang ZD, Zhou LW, Zhang J, Bai XF, Jiang Y et al (2019) SEanalysis: a web tool for super-enhancer associated regulatory analysis. Nucleic Acids Res 47: W248–W255

Rackham OJ, Jaffes J, Fang H, Oates ME, Holmes ML, Knaupp AS, Suzuki H, Neff Z, Dau B, Shin JW et al (2018) A predictive computational framework for direct reprogramming between human cell types. Nat Genet 48: 331–335

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysoczka J (2012) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470: 279–285

Reg viv A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Catzeflis M et al (2017) The human cell atlas. eLife 6: e27041

Remesal R, Roger-Baynat I, Chirivella L, Maicas M, Brocal-Ruiz R, Pérez-Villalba A, Cucarella C, Casado M, Flames N (2020) PBX1 acts as terminal selector for olfactory bulb dopaminergic neurons. Development 147: dev186841

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilinkny M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317–329

Rolland T, Tsaun M, Charloteaux B, Pevzner S, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R et al (2014) A proteome-scale map of the human interactome network. Cell 159: 1212–1226

© 2021 The Authors

The EMBO Journal 40: e106785 | 2021

11 of 12
Romer Al, Susel L (2015) Pancreatic islet cell development and regeneration. *Curr Opin Endocrinol Diabetes Obes* 22: 255 – 264

Russell R, Carnese PP, Hennings TG, Walker EM, Russ HA, Liu JS, Giacometti S, Stein R, Hebrok M (2020) Loss of the transcription factor MAFB limits β-cell derivation from human PSCs. *Nat Commun* 11: 1 – 15

Saelens W, Cannoodt R, Todorov H, Saes Y (2019) A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 37: 547 – 554

Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young RA (2016) Models of human core transcriptional regulatory circuitries. *Genome Res* 26: 385 – 396

Satterlee JS, Schübel D, Ng HH (2010) Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol* 28: 1039 – 1044

Scarl RT, Corbin KL, Vann NW, Smith HM, Satin LS, Sherman A, Nunemaker CS (2019) Intact pancreatic islets and dispersed beta-cells both generate intracellular calcium oscillations but differ in their responsiveness to glucose. *Cell Calcium* 83: 102081

Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Gould J, Liu S, Lin S, Berube P et al (2019) Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 176: 928 – 943.e22

Segerstolpe A, Palasantza A, Eilasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK et al (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 24: 593 – 607

Sellers L, Zappavigna V, Ferretti E (2019) ‘Building a perfect body’: control of vertebrate organogenesis by PRX-dependent regulatory networks. *Genes Dev* 33: 258 – 275

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from genome-wide predictions. *Nat Rev Genet* 15: 272 – 286

Spitz F, Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 13: 613 – 626

Sun Y, Miao N, Sun T (2019) Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* 156: 29

Szkarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Jensen LJ et al (2018) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47: 607 – 613

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T et al (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 19: 335 – 346

Thibodeau A, Uyar A, Khetan S, Stitzel ML, Ucar D (2018) A neural network based model effectively predicts enhancers from clinical ATAC-seq samples. *Sci Rep* B: 1 – 15

Tiwari N, Pataskar A, Péron S, Thakurela S, Sahu SK, Figuères-Oñate M, Marichal N, López-Mascarache L, Tiwari VK, Berninger B (2018) Stage-specific transcription factors drive astrogliogenesis by remodeling gene regulatory landscapes. *Cell Stem Cell* 23: 557 – 571.e8

Tosti L, Hang Y, Trefzer T, Steiger K, Ten FW, Lukassen S, Baille S, Kuehl A, Spieckermann S, Bottino R et al (2021) Single-nucleus and in situ RNA-sequencing reveal cell topographies in the human pancreas. *Gastroenterology* 160: 1330 – 1344.e11

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32: 381 – 386

Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, Lickert H, Theis FJ (2019) Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* 146: dev170506

Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* 463: 1035 – 1041

Villaescusa JC, Li B, Toledo EM, di Val R, Cervo P, Yang S, Stott SR, Kaiser K, Islam S, Gyllborg D et al (2016) A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson’s disease. *EMBO J* 35: 1963 – 1978

Wagner A, Regev A, Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* 34: 1145 – 1160

Wang J, Yuan R, Zhu X, Ao P (2020) Adaptive landscape shaped by core endogenous network coordinates complex early progenitor fate commitments in embryonic pancreas. *Sci Rep* 10: 1 – 17

Wang L, Tan TK, Durbin AD, Zimmerman MW, Abraham BJ, Tan SH, Ngoc PCT, Weichert-Leahey N, Akahane K, Lawton LN et al (2019) ASC1L1 is a MYCN- and LM01-dependent member of the adrenergic neuroblastoma core regulatory circuitry. *Nat Commun* 10: 1 – 15

Weiss MC, Green H (1967) Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes. *Proc Natl Acad Sci USA* 58: 1104 – 1111

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kayge MH, Rahl PB, Lee TI, Young RA (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 158: 307 – 319

Wilson S, Filipp FV (2018) An network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer. *npj Syst Biol Appl* 4: 24

Xia B, Yanai I (2019) A periodic table of cell types. *Development* 146: 1 – 9

Xin Y, Kim J, Oikamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, Grámona D (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab* 24: 608 – 615

Xu J, Du Y, Deng H (2015) Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell* 16: 119 – 134

Young RA (2011) Control of the embryonic stem cell state. *Cell* 144: 940 – 954

Zaret KS, Carroll JS (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* 25: 2227 – 2241

Zaret KS, Mango SE (2016) Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr Opin Genet Dev* 37: 76 – 81

Zhou Q, Brown J, Kanarek J, Melton DA (2008) In vivo reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* 455: 627 – 632

Zhu Y, Liu Q, Zhou Z, Ikeda Y (2017) PDX1, Neurogenin-3, and MAFA: critical transcription regulators for beta cell development and regeneration. *Stem Cell Res Ther* 8: 240

License: This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.