Chapter 1

Introduction

1.1 Embryonic development

Prenatal development in placental mammals begins with fertilization of an oocyte by a sperm cell in the in the ampulla of the fallopian tube. The fusion of these two gametes leads to formation of a diploid cell, which is called the zygote. Zygotes have all the genetic material that is necessary for development into the whole organism. The first cell division is special, because the chromosomes from each pronucleus (one from oocyte, one from sperm) are doubled, and syngamy *i.e.* the combination of maternal and paternal chromosomes only occurs during this first mitosis. During the first rounds of division, all embryonic cells remain totipotent, *i.e.* they can give rise to any tissue, either embryonic or extraembryonic (Chason et al., 2011; Saiz and Plusa, 2013).

When the embryo reaches about 100 cells the first cell fate commitments happen (Wennekamp et al., 2013). At this stage, a blastocoel - cavity within the embryo - is formed and cells differentiate into two groups: trophectoderm cells that position on the outside and inner cell mass cells that are inside on the so-called animal pole of the embryo (Figure 1.1). Further in development, during gastrulation, the trophoblast develops into trophectoderm, which gives rise to the placenta. Inner cell mass cells are pluripotent; they develop into three germ layers (ectoderm, endoderm, and mesoderm) of the embryo proper as well as the hypoblast, which later becomes extraembryonic membranes. Embryonic stem cells are derived from cells of the inner cell mass usually at 3.5 days after fertilisation. The blastocyst develops three days after fertilization and is fully formed on the fourth day. At this stage of development the embryo is ready for implantation (Saiz and Plusa, 2013; Tam and Loebel, 2007).

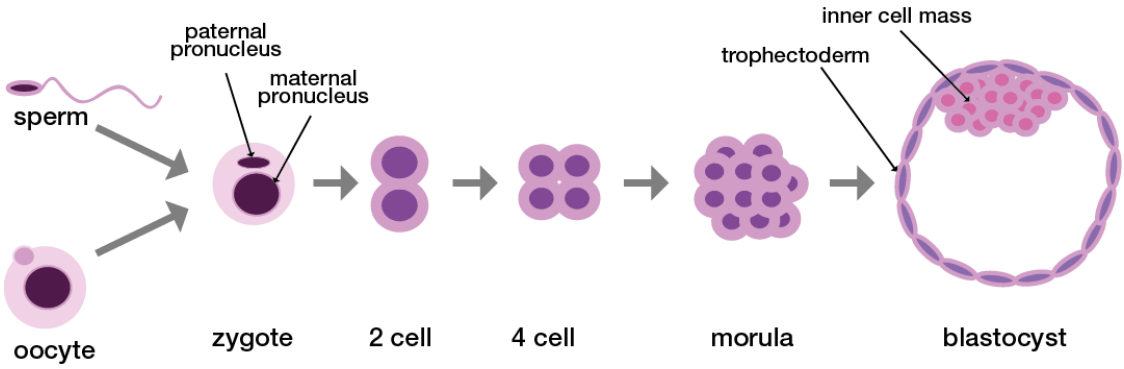

Figure 1.1 Early embryo development
During fertilisation sperm and oocyte combine to form a zygote. It divides giving rise to more totipotent cells. The first two lineages are formed at the blastocyst stage where some cells form a trophectoderm layer which encapsulates the second type of cells -inner cell mass or epiblast and a liquid called blastocoel.

The embryo undergoes gastrulation after implantation, when the body axes are formed and, most importantly, forms the primitive streak with

differentiation of cells into germ layers *via* an epithelial to mesenchymal transition. Later, the endoderm develops into epithelia of the respiratory and digestive tracts, liver and pancreas. The mesoderm becomes muscles, blood, bones, cartilage and other connective tissues, and ectoderm differentiates into skin and neuronal tissues (Tam and Behringer, 1997; Tam and Loebel, 2007).

In contrast to plants, for which totipotency has been known to be a property of each cell for decades (Steward et al., 1958), it was thought that mammalian pluripotent or totipotent cells can only be obtained from embryos until 2006. The discovery and development of induced pluripotent stem cells (iPSCs) revolutionised our understanding of pluripotency in mammals. The expression of four transcription factors, *Pou5f1*, *Sox2*, *cMyc*, and *Klf4* (the 'Yamanaka factors'), causes differentiated cells to be reprogrammed and gain key features of pluripotency: self renewal and the ability to differentiate into different tissues (Takahashi and Yamanaka, 2006).

1.2 Origins of mouse embryonic stem cell cultures

Historically, mouse embryonic stem cell cultures (mESCs) originate from the cultures of teratocarcinomas, tumours of germ cells which occur more commonly in testis, but can also develop within ovaries (Stevens and Little, 1954). Teratocarcinomas are a unique type of tumour, as they contain different types of differentiated tissues, sometimes even teeth or hair (Kleinsmith and Pierce, 1964; Pierce, 1967; Rosenthal et al., 1970). Within teratocarcinomas there are undifferentiated cells called embryonic carcinoma (EC) cells, which can proliferate and differentiate into all cell types of the tumour. Additionally, EC cells are transplantable and self-renewing, and when transplanted to a different animal and they still give rise to all tissues of the tumour. The EC

cells can self-renew and differentiate to all cell types, which are the two main characteristics of pluripotency. This makes them more similar to early embryonic cells than to germ cells (Stevens, 1970). Interestingly, if pluripotent cells from the early embryo are grafted onto a mouse they will develop into a tumour (Stevens, 1970).

These characteristics of EC cells made it possible to establish their cultures *in vitro* already in the 1970s. The cells were cultured in the presence of blood serum on feeder cell layers (usually mitotically inactivated fibroblasts) and they maintained their pluripotency (Martin, 1975, 1980; Martin and Evans, 1974). Importantly, EC cells are inefficient in colonizing embryos when injected into them due to their chromosomal abnormalities, but those without chromosomal abnormalities can indeed colonize embryos (Mintz and Illmensee, 1975).

Successful culturing of EC cells and their similarity to embryonic cells led to the idea that cells from early embryos could be cultured. Indeed, using the same pluripotency-maintaining conditions as for culturing EC cells, mouse embryonic stem cells from the inner cell mass of the 3.5 day blastocyst were cultured (Evans and Kaufman, 1981; Martin, 1981). Soon afterwards, the first mouse embryonic cell lines that efficiently colonized blastocyst stage embryos were established (Bradley et al., 1984).

1.3 Pluripotency signalling in mESC cultures

When culturing embryonic stem cells *in vitro* it is important to ensure that they maintain their pluripotency, meaning they can divide and give rise to more pluripotent cells, and then with appropriate signals, they can differentiate into all other cell types of the organism (Davidson et al., 2015;

Smith, 2001). The first culture conditions were found by an empirical 'trial and error' approach and are very different from the natural environment of the embryo. Culturing cells on feeders in media supplemented with serum has some limitations. Firstly, it efficiently supports pluripotency only for mice of the Sv/129 genetic background or a hybrid of it (Suzuki et al., 1999). It is still unclear which genetic differences make the Sv/129 strain remain pluripotent under these conditions in comparison to C57Bl/6 or other laboratory strains of mice (Nagy et al., 1993). Additionally, the pluripotency of male lines is more successfully maintained for mouse embryonic stem cells derived using this culture condition; female cells tend to lose one of their X chromosomes and grow with a 39,X0 karyotype (Minina et al., 2010; Zvetkova et al., 2005). Finally, these conditions do not support growth of stem cells from other species such as rat and, more importantly, human (Martello and Smith, 2014).

Designing optimal conditions for culturing pluripotent cells requires a thorough understanding of the extracellular signals that lead to pluripotency maintenance and those which lead to differentiation. Cells differentiate in the absence of feeders and serum, suggesting that these additions provide pluripotency-maintaining signals to the mESCs. Media conditioned with feeders or buffalo rat liver cells is able to maintain mESCs in an undifferentiated state for a limited time (Smith and Hooper, 1987). The key factor supplied by the feeder cells was later found to be a secreted protein, leukaemia inhibitory factor (LIF) (Smith et al., 1988; Williams et al., 1988).

The addition of LIF to the culture removes the need for feeder cells, which made culturing and experimenting on mESCs more practical. Supplementation with LIF can also help to achieve good pluripotent cultures in the presence of feeders. LIF binds to the LIFR protein on the surface of

mESCs. This binding causes recruitment of glycoprotein 130 (GP130) and formation of a LIFR-GP130 heterodimer (Gearing et al., 1991). This receptor heterodimer recruits Janus-associated kinases (JAKs) and phosphorylates them. Subsequently, STAT proteins, most importantly STAT3, are phosphorylated, dimerise and translocate into the nucleus. There they in turn regulate expression of many genes including Krüppel Factors, most notably *Klf4*, which function in a gene regulatory network that regulates proliferation and pluripotency maintenance (Figure 1.2) (Hall et al., 2009; Matsuda et al., 1999; Niwa et al., 2009). It was observed that cells cultured in serum supplemented with LIF are more heterogeneous in their morphology than cells cultured in the presence of feeders, suggesting that LIF is not the only signal supplied by the feeder cells (Onishi and Zandstra, 2015).

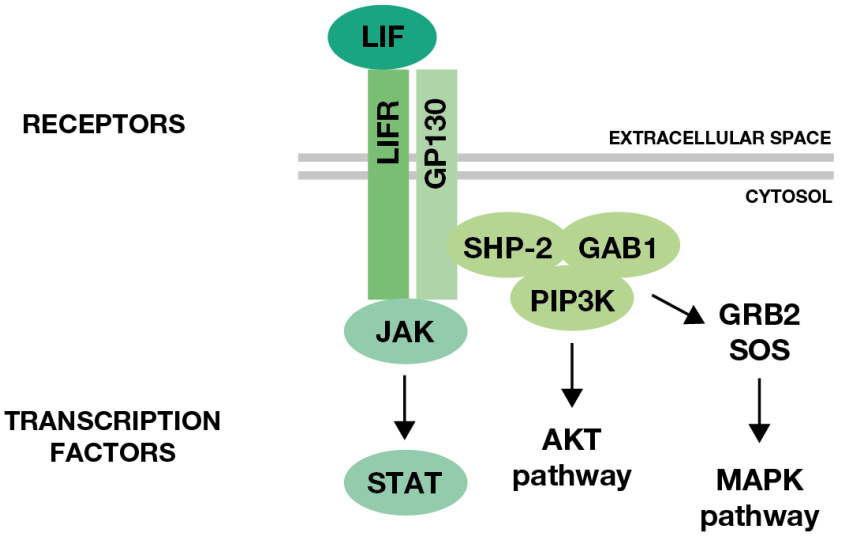

Figure 1.2 LIF signalling

LIF binds its cognate receptor LIFR which dimerises with GP130. They signal to several pathways that alter transcription, most importantly to the JAK/STAT pathway, but also *via* SHP-2, GAB1 and PIP3K to the AKT pathway, and further *via* GRB2 and SOS to the MAPK pathway.

Removal of serum from the culture media causes mESCs to spontaneously differentiate toward the neuronal lineage (Ying et al., 2003b), implying that serum contains factors that inhibit this process. One of the components that play a role was identified to be bone morphogenic protein BMP4. It is an inhibitor of neuronal lineage differentiation *via* induction of inhibitor of DNA binding (*Id*) genes (Ying et al., 2003a).

Another pathway implicated in pluripotency maintenance is the mitogen-activated protein kinases (MAPK) pathway (Burdon et al., 1999). The phosphorylation cascade of MAPK starts by exchange of GDP to GTP bound to the GTPase RAS. This exchange is triggered by extracellular signals binding to receptors such as epithelial growth factor receptor EGFR and subsequent phosphorylation of intracellular SH2 domains of the receptor. The GRB2 protein is phosphorylated during activation of EGFR, and forms a complex with its receptor and the guanine nucleotide exchange factor SOS, which promotes GDP to GTP exchange. GTP-bound RAS activates downstream serine/threonine kinase MAP3K (RAF), which in turn activates serine/threonine kinase MAP2K (MEK1/2) and subsequently tyrosine/threonine kinase MAPK (ERK1/2). Phosphorylated ERK1/2 is an important regulator of the activity of several transcription factors including MYC, CREB, ELK, ETS, SRF and FOS. These regulators modulate transcription of downstream transcription programmes, including the transcription of cell cycle genes (Figure 1.3). Interestingly, ERK1/2 also acts on translation by regulating ribosomal activity *via* phosphorylation of ribosomal s6 kinase (RSK) (Kolch, 2000).

In addition to activating STATs, LIF signalling also activates the MAPK pathway, CREB and PI3K pathway (Burdon et al., 1999; Ernst et al., 1996).

LIFR and the receptor GP130 act indirectly *via* SHP-2, GAB1 and PI3K to cause phosphorylation of GRB2 and trigger the MAPK phosphorylation cascade (Burdon et al., 1999). The MAPK pathway is one of the key signalling pathways in any cell and it regulates several processes, most importantly the cell cycle (Johnson and Lapadat, 2002; Pruitt and Der, 2001; Zhang and Liu, 2002). It may appear contradictory that LIF signalling promotes pluripotency *via* STATs and differentiation *via* ERK1/2. It has been proposed that the balance between these pathways is key for achieving self-renewal and maintenance of potency for differentiation (Niwa et al., 2009).

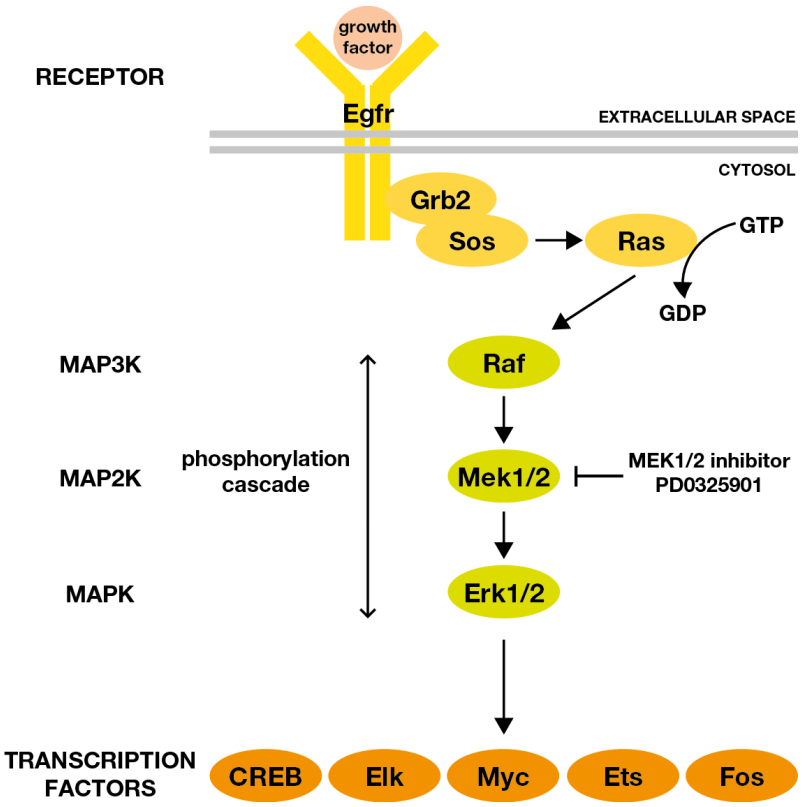

Figure 1.3 MAPK signalling

MAPK signalling starts with a mitogen such as EGF binding to its receptor at the membrane. Subsequently signal is transmitted *via* GRB2 and SOS to RAS, which causes phosphorylation of the first kinase (MAP3K) Raf, which in turn phosphorylates (MAP2K) Mek1/2 and then phosphorylated Mek1/2 phosphorylates (MAPK) Erk1/2, which regulates many transcription factors. Inhibition of this pathway at Mek1/2 helps maintenance of the pluripotent state.

Understanding the importance of MAPK signalling led to attempts to interfere with the pathway with the intention of maintaining a pluripotent state in the absence of BMP4. Serum-free medium with addition of the small molecule inhibitors of MEK1/2 in the presence of LIF was shown to support pluripotency (Kunath et al., 2007). Similarly, inhibition of GSK3β with a small molecule, along with LIF was enough to maintain the self-renewal and differentiation potential of mESCs (Ying et al., 2008). The main effect mediated by GSK3β is accumulation of β-catenin and competition with the DNA binding protein TCF3, which is a repressor of key pluripotency genes (Figure 1.4).

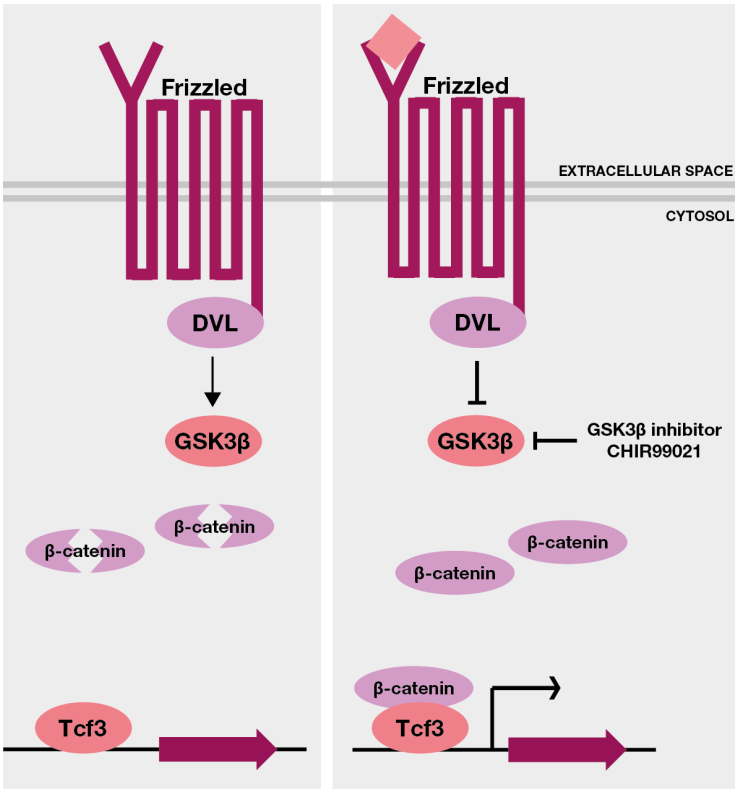

Figure 1.4 Wnt signalling

In the presence of Wnt bound to the Frizzled receptor, Dishevelled activates GSK3β kinase. Phosphorylation by GSK3β and subsequent ubiquitination of β-catenin by the destruction complex leads to degradation of β-catenin by the proteasome. Inhibition

of GSK3β leads to accumulation of β-catenin in the cytoplasm, and its translocation to the nucleus, where it competes with transcription repressors such as TCF3 causing gene expression.

These discoveries led to the formulation of so called "2i medium". This medium owes its name to the fact that it combines two inhibitors: an inhibitor of MEK1/2 and of GSK3β (Ying et al., 2008). 2i medium allows derivation and maintenance of all mESCs regardless of their genetic background, and also supports derivation of embryonic stem cells from other rodents, but not human (Buehr et al., 2008; Nichols et al., 2009). The key advantage of 2i medium is that it is chemically defined and thus standardized, which is not possible to achieve using feeders or serum. Serum contains molecules that act as differentiation factors and, if in a particular batch they are not balanced with factors mediating pluripotency maintenance, the cells respond by differentiating. Moreover, feeders can sometimes be a source of infection with pathogens and it is difficult to control the factors they secrete into the media. Cells in 2i are significantly more morphologically homogeneous than cells cultured in serum supplemented with LIF (Marks et al., 2012). These observations led to a description of the state of mESCs cultured in 2i media as the "ground state" of pluripotency (Ying et al., 2008).

For use in experiments, mESCs are usually cultured on feeder layers or gelatine-coated dishes as the cells usually adhere to the culture surface. Alternatively, they can be cultured as spheroids in suspension in the presence of either serum and LIF (Fok and Zandstra, 2005; zur Nieden et al., 2007) or in a chemically defined medium supplemented with LIF and basic fibroblast growth factor (bFGF) (Andang et al., 2008). Within suspension cultures lacking anti-differentiation factors, mESCs develop into three-dimensional clusters of cells called embryoid bodies. These embryoid bodies recapitulate several

aspects of early embryo development including formation of three germ layers: endoderm, mesoderm and ectoderm (Itskovitz-Eldor et al., 2000; Keller, 1995).

The elasticity of the surface on which mESCs grow plays an important role in maintaining pluripotency, and so dishes on which cells are grown are coated with gelatine. The properties of the surface on which cells grow are important, because mechanical cues of the environment are transformed into biochemical signals by molecules called mechanosensors, such as integrins. Integrins subsequently forward the signal to the cytoskeleton, but also to signalling pathways such as the WNT and MAPK pathways (Ishihara et al., 2013). Inhibition of SRC removes the requirement for an elastic substrate, and replacing MEK1/2 inhibitors with SRC inhibitors also maintains pluripotency. Medium such as this is known as "alternative 2i" (Shimizu et al., 2012).

In addition to mediation of signalling from the focal adhesion kinase (FAK), SRC signals to the MAPK pathway *via* SHC-transforming protein SHCA (Matsui et al., 2012). Hence, inhibition of SRC seems to have a dual role by affecting both MAPK pathway and adhesion signalling (Shimizu et al., 2012). Moreover, inhibition of SRC blocks upstream calcineurin-NFAT signalling, which also plays a role in endothelial to mesenchymal transition (EMT) (Li et al., 2011). Importantly, LIF signalling *via* the JAK-STAT pathway regulates the activity of SRC (Anneren et al., 2004). This suggests that inhibition of either SRC or MEK1/2 achieves a similar effect because both inhibit differentiation (Figure 1.5).

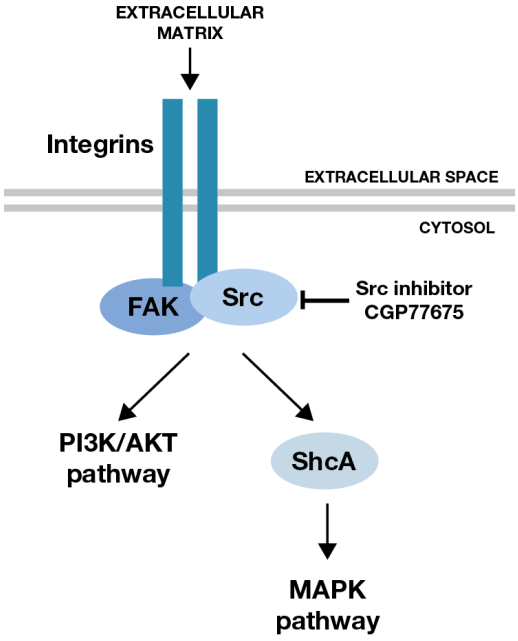

Figure 1.5 Src signalling

Focal adhesion kinase and Src mediate signals arising from the physical properties of the extracellular matrix. They signal further to different pathways including PI3K/AKT pathway, and *via* SHCA, to the MAPK pathway. Inhibition of Src leads to inhibition of downstream pathways, leading to a similar phenotype as inhibition of Mek1/2.

Under appropriate *in vitro* conditions, when pluripotency signals from serum/BMP4 and feeders/LIF are removed, mESCs differentiate into several different cell types. Differentiation is mediated by FGF4, which binds to its receptor, FGFR2, and activates the MAPK pathway (Kunath et al., 2007; Stavridis et al., 2007). There is substantial effort being invested to find signals that cause differentiation towards cell types of interest (Doetschman et al., 1985; Keller, 1995).

The question that arises is whether *in vitro* culture of mESCs is equivalent to the physiological conditions that occur within the embryo. Typically, the prolonged culture of cells from differentiated tissues for long periods of time requires the cells to have abnormal proliferative properties either because they originate from tumours (*e.g.* HeLa cells) or they have been immortalized in

some other way. Under the right conditions, mESCs can self-renew indefinitely without immortalization, which is consistent with their tumorigenic potential (Suda et al., 1987). This property of mESCs seems unexpected because pluripotent cells do not need to multiply indefinitely in the embryo. The fact that mESCs are able to contribute to the embryo even after many rounds of division in culture suggests that they are pluripotent. Even if culturing caused differences between mESCs and cells of the blastocyst inner cell mass these differences must be reversible such that mESCs can take on the fate of inner cell mass cells.

When mice are suckling previous litters and their oestrogen levels are low, embryos do not implant and enter a special quiescent state called diapause, with an almost complete halt of proliferation and metabolism (Renfree and Shaw, 2000). High levels of oestrogen and the presence of LIF are necessary for implantation in mice (Hondo and Stewart, 2004; Mantalenakis and Ketchel, 1966; Renfree and Shaw, 2000). It has been proposed that mESCs in culture may represent diapaused embryos (Nichols et al., 2001). LIF signalling is necessary for survival of diapaused embryos and pluripotency maintenance in mESCs (Nichols et al., 2001). Diapause can be mimicked in mESCs by deleting *Myc* (Scognamiglio et al., 2016) suggesting that this is the factor that mediates proliferation. It is not apparent how this can be explained in light of the fact that STAT3 activates *Myc* (Cartwright et al., 2005), but probably the balance between signalling of JAK/STAT and MAPK pathways plays a crucial role.

1.4 Transcriptional regulators of pluripotency

The master regulator of pluripotency is OCT4, encoded by the *Pou5f1* gene (Pan et al., 2002; Pardo et al., 2010; van den Berg et al., 2010) that is expressed solely in early embryo and germ line cells. Embryos lacking OCT4 develop to the blastocyst stage, but the inner cell mass cells are not pluripotent and can only form extraembryonic tissues (Nichols et al., 1998). Deletion of *Pou5f1* in mESCs leads to loss of self-renewal and causes them to differentiate. Interestingly, overexpression of *Pou5f1* also leads to loss of pluripotency and differentiation to endoderm and mesoderm (Niwa et al., 2000).

In addition to OCT4, the pluripotency network is regulated by homeobox protein NANOG (Saunders et al., 2013). OCT4 and NANOG function in concert and often bind promoters of the same genes (Loh et al., 2006). Deletion of *Nanog* has a similar effect to deletion of *Pou5f1* and causes loss of pluripotency with differentiation toward extraembryonic lineages. *In vivo* loss of *Nanog* causes embryos at the blastocyst stage to form parietal endoderm-like cells and to lack epiblast (Mitsui et al., 2003; Silva et al., 2009). Ectopic expression of *Nanog* from a transgene construct causes cells to remain pluripotent independent of LIF signalling *via* the JAK/STAT pathway (Chambers et al., 2003).

Nanog expression is regulated by the SRY-box transcription factor SOX2 along with OCT4 (Rodda et al., 2005). SOX2 and OCT4 regulate transcription by binding to sox-oct elements in promoter and enhancers of downstream genes, which include many transcription factors and notably also their own promoters of *Sox2* and *Pou5f1* (Chew et al., 2005).

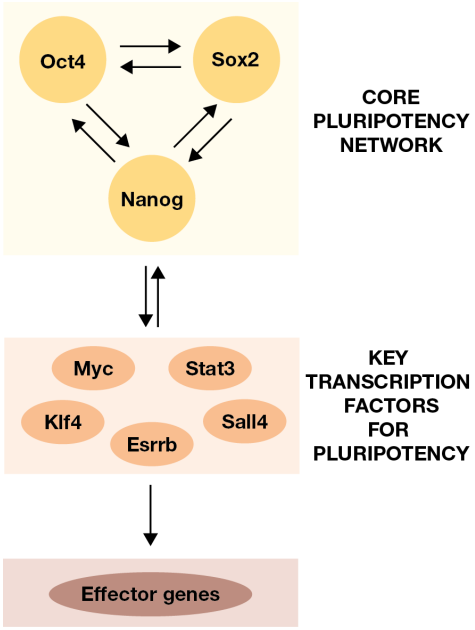

Figure 1.6 Pluripotency network

In the current view of transcription factors regulating pluripotency, key transcription factors OCT4, SOX2 and NANOG are highly interconnected and regulate expression of each other. These genes then signal to other transcription factors important for pluripotency, which propagate signal to effector genes and also regulate extended pluripotency networks.

Our current understanding of the gene regulatory network involving key pluripotency factors describes a highly interconnected network (Figure 1.6) (Boyer et al., 2005; Chickarmane et al., 2006; Kushwaha et al., 2015; Pan and Thomson, 2007). OCT4, NANOG and SOX2 co-occupy promoters of many genes, often transcription factors including themselves, resulting in feed-forward loops (Boyer et al., 2005; Chambers and Tomlinson, 2009). Downregulation by shRNA of *Nanog*, *Pou5f1*, *Sox2*, *Esrrb*, *Tbx3*, *Tcl1* and *Dppa4* also cause impairment in self-renewal (Ivanova et al., 2006). Affinity purification and mass spectrometry demonstrated that NANOG protein interacts with several transcription factors including OCT4, SALL4, SALL1, RIF1 and MYBBP (Wang et al., 2006). It was suggested that the function of the highly interconnected architecture of the network is the robust response to

developmental stimuli whilst dampening random gene expression fluctuations (Sokolik et al., 2015; Torres-Padilla and Chambers, 2014).

1.5 Chromatin state and structure as regulators of pluripotency

DNA in cells is packaged into chromatin to make it possible to fit long DNA molecules into the nucleus, to prevent damage of DNA and to regulate DNA function. The basic unit of chromatin is a nucleosome, which consists of 8 histone molecules (2 copies each of the core histones H2A, H2B, H3, and H4) and 147bp of DNA wrapped around them. Histones tails are posttranslationaly modified to affect their interaction with DNA and other proteins. Methylation, acetylation, phosphorylation and ubiquitination are the most common, but other modifications also occur (Jenuwein and Allis, 2001; Strahl and Allis, 2000). Posttranslational modifications of histones regulate the recruitment of different regulatory proteins. For example, methylation of H3K4 causes gene activation, while methylation of H3K27 and ubiquitination of H2AK119 lead to silencing of gene expression.

An entire organism containing diverse cell types develops from a single zygote. Hugely diverse cellular functions exist despite each cell having the same genome. This is possible due to regulated gene expression. Chromatin state is very important in determining whether a particular gene is active, poised or silenced and is crucial in regulating the transcriptional identity of the cell.

Expression of genes that regulate pluripotency maintenance and development is highly regulated by chromatin structure. mESCs are highly transcriptionally active and express many genes at low levels (Efroni et al., 2008; Efroni et al., 2009). This promiscuous transcription is thought to mediate

pluripotency since low levels of differentiation factors and markers of all lineages are expressed (Efroni et al., 2008; Loh and Lim, 2011). This phenomenon is attributed to largely accessible chromatin throughout the genome during early stages of development (Meshorer and Misteli, 2006). Differentiation leads to the genes that are not needed for the particular cell type becoming silenced by changes in chromatin structure. This causes cells to acquire a particular stable identity that cannot be reversed without intervention, such as reprograming to induced-pluripotent stem cells (iPSCs). Regulation of chromatin structure occurs *via* different mechanisms: DNA methylation, modification of histones and action of ATP-dependent chromatin remodellers (Li et al., 2012).

1.5.1 DNA methylation

The first level of chromatin modification is DNA methylation at cytosines of CpG dinucleotides. There are two types of DNA methylation: (1) maintenance methylation by DNMT1, which methylates hemi-methylated CpGs that arise after DNA replication during S phase of the cell cycle and (2) *de novo* methylation by DNMT3A and DNMT3B (Okano et al., 1999; Pawlak and Jaenisch, 2011; Tsumura et al., 2006). After fertilization, there is a wave of massive demethylation of DNA, which has to be regained in the inner cell mass cells at the blastocyst stage of the embryo (Morgan et al., 2005). As mentioned above, demethylation can happen passively during DNA replication, but can also occur by an active process either *via* the activation-induced cytidine deaminase (AID) pathway or *via* oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by the ten-eleven translocation (TET)

enzymes (Ficz et al., 2011; Ito et al., 2010; Koh et al., 2011; Ooi and Bestor, 2008).

X-chromosome inactivation, needed for female embryos to obtain the same gene dosage as male embryos, happens before implantation. It involves binding of the noncoding RNA Xist, and a subsequent major wave of histone modifications including loss of H3K4me2 and H3K4me3, and the gain of H3K9me2 and H3K27me3, as well as the ubiquitination of H2A (Galupa and Heard, 2015; Pollex and Heard, 2012). Somatic chromosomes are demethylated during preimplantation development, and afterwards methylation is regained through the action of DNMT3B (Watanabe et al., 2002).

Methylation of DNA can be monitored using bisulfite sequencing, in bulk and recently also in single cells (Farlik et al., 2015; Guo et al., 2013; Kantlehner et al., 2011; Smallwood et al., 2014). Cells cultured in 2i have greatly hypomethylated DNA in comparison with cells in serum and, similarly to their transcriptomes, their methylomes exhibit heterogeneous patterns in the serum but not 2i cells (Angermueller et al., 2016; Ficz et al., 2013). This suggests that cells cultured in 2i media are closer to the pluripotent ground state of cells in the inner cell mass, as methylation is lowest in embryos at this stage of development (Smith et al., 2012).

1.5.2 Histone modifications

DNA methylation is a relatively stable modification, and is not easily reversed. Many genes in the inner cell mass are regulated by histone modification rather than methylation due to the generally hypomethylated state of the genome. Key signalling pathways in mouse embryonic stem cells regulate histone modifications. These include the JAK/STAT pathway

(Griffiths et al., 2011), the WNT pathway, the MAPK pathway and FGF signalling (Ficz et al., 2013; Habibi et al., 2013; Leitch et al., 2013).

There are two key complexes implicated in histone regulation in ESCs: the Polycomb repressor complex and the Trithorax complex. Trithorax promotes self-renewal while Polycomb promotes developmental potency to achieve cells with both hallmarks of pluripotency: self-renewal and developmental potency (Ang et al., 2011; O'Carroll et al., 2001).

There are two Polycomb complexes in mouse: PRC1 and PRC2. PRC2 genes *Ezh1* and *Ezh2* are members of a histone methyltransferase complex, and are essential for early mouse development. It is not possible to derive embryonic stem cells from *Ezh2* knockout embryos (O'Carroll et al., 2001). The PRC2 complex deposits histone 3 lysine 27 trimethylation (H3K27me3), a repressive mark, which may lead to chromatin compaction mediated by PRC1 (Boyer et al., 2006; Francis et al., 2004; Ringrose et al., 2004). The other proteins in the PRC complex include zinc finger SUZ12, EED, histone binding protein RBAP48 and other proteins such as JARID2 or PCLs (Margueron and Reinberg, 2011).

PRC1 binds to H3K27me3, deposited by PRC2, and is composed of several different components; PRC1 subunits often have alternative versions. H2K27me3 is bound by members of the chromobox family of proteins (CBX2, CBX4, CBX6, CBX7 and CBX8) and the PRC1 complex may contain any of them. CBX7 is the most common in mESCs and it functions in preventing precocious differentiation (Martin, 2010). Levels of CBX7 decrease during cell differentiation and it is replaced by CBX2, CBX4 and CBX8 (Morey et al., 2012; O'Loghlen et al., 2012). The molecular mechanism involves monoubiquitination of the histone 2A lysine 119 (H2AK119Ub1) by the

ubiquitin ligase Ring1B, and further compaction of the chromatin (Buchwald et al., 2000; de Napoles et al., 2004; Wang et al., 2004). Interestingly, RNA polymerase phosphorylated on S5 but not on S2 of the C-terminal domain can still transcribe genes marked by PRC with H3K27me3 (Brookes et al., 2012).

Hierarchical model for Polycomb repression where PRC2 deposited marks recruit PRC1 is not the only possible pathway. Other studies shown that depending on the composition of the complexes recruitment of PRC1 to the chromatin and histone mark deposition differs (Blackledge et al., 2014; Cooper et al., 2014). This system is highly complex and in addition to changes in function mediated by subunit composition, it also involves interactions between PRC1 and PRC2 complexes (Cao et al., 2014) and different mechanisms of recruitment to the chromatin involving other types of histone modifications, for instance H3K9 methylation, interactions with transcription factors and ncRNAs (Brockdorff, 2013; Mozzetta et al., 2014; Yu et al., 2012).

The Trithorax group protein WDR5 mediates histone 3 lysine 4 trimethylation (H3K4me3). This modification causes recruitment of histone acetylases and remodelling enzymes, and positively regulates transcription (Ang et al., 2011; Pray-Grant et al., 2005; Santos-Rosa et al., 2003; Wysocka et al., 2005). Using ChIP-sequencing it was observed that upstream of some genes, including *Hox* gene clusters, there are both active (H3K4me3) and repressive (H3K27me3) histone marks. These genes are mostly other developmental regulators, and such 'bivalent domains' at their promoters are thought to mediate a poised state of transcription (Bernstein et al., 2006). Cells cultured in 2i have fewer bivalent domains than cells cultured in serum, in accordance with their more naïve state (Figure 1.7) (Marks et al., 2012).

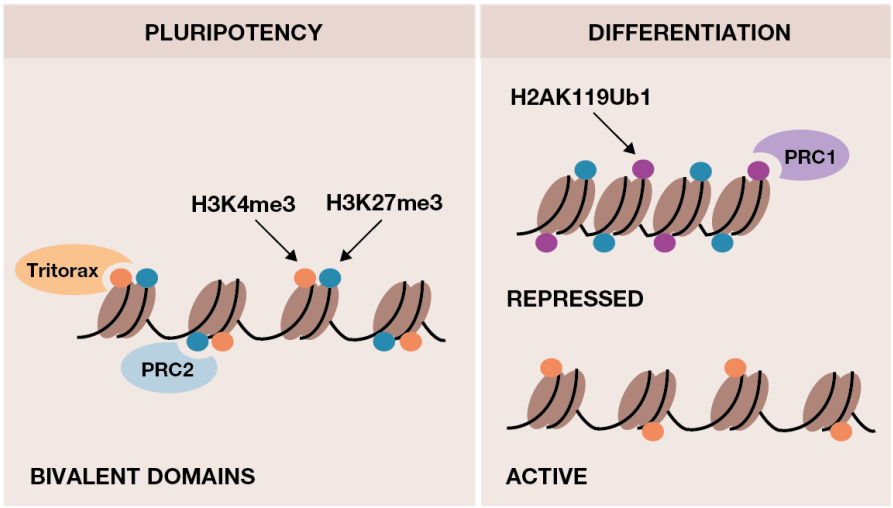

Figure 1.7 Histone modifications in pluripotent and differentiated cells
The promoters of tissue-specific genes and pluripotency genes include both active (H3K4me3) and repressive (H3K27me3) marks deposited by the Trithorax and PRC2 complexes respectively. Upon differentiation, these domains either loose repressive marks and remain active and expressed, or in addition to H3K27me3, gain the compaction chromatin mark (H2AK119Ub1) by PRC1 and become completely silenced.

Enzymes can also remove epigenetic marks. During differentiation, the Lys-specific demethylase 1 (LSD1), which associates with the nucleosome remodelling and deacetylase (NuRD) complex, removes H3K27 and H3K4 methylation marks from enhancers. These enhancers are then no longer occupied by transcriptional activators, and this shuts down the pluripotency expression programme (Adamo et al., 2011; Whyte et al., 2012).

1.5.3 Chromatin remodelling

Chromatin remodellers are typically large, multi-subunit complexes that have diverse functions in cells, including the regulation and maintenance of pluripotency. Depending on the sequence of the ATPase that they contain, chromatin remodellers can be divided into four families: SWI/SNF, CHD, ISWI and INO80 complexes. Their main mode of action is to regulate DNA

accessibility by disrupting the interactions between DNA and nucleosomes in an ATP-dependent manner (Clapier and Cairns, 2009; Narlikar et al., 2013; Saha et al., 2006).

The subunit composition and function of remodelling complexes change during development. The exact composition of the complex tunes its affinity for particular target genes (Ho and Crabtree, 2010; Martin, 2010). During the transition from pluripotency to trophoblast-like cells, the SWI/SNF family complex Brahma Associated Factors (BAF) changes its composition dramatically (Yan et al., 2008). Additionally, the embryonic stem cell-specific BAF complex co-localizes with the pluripotency regulators NANOG, OCT4, SOX2 and STAT3, which suggests that chromatin remodelling is crucial for the action of core pluripotency transcription factors (Ho et al., 2009). BRG1 (also known as SMARCAD4) is a component of BAF whose downregulation results in differentiation and loss of expression of key pluripotency genes (Kidder et al., 2009). *Brg1* knockouts are embryonic lethal in mice due to a failure to form the pluripotent inner cell mass in the blastocyst (Bultman et al., 2009). In comparison, BRM, which is a protein that can replace BRG1 to form a functioning BAF, is dispensable for early development (Bultman et al., 2009).

Nucleosome-remodelling and histone deacetylase (NURD) complexes, which are a subfamily of the CHD family of chromatin remodellers, also play a role in pluripotency maintenance. Their repressor function is mediated by histone deacetylases (HDACs) within the complex. These complexes also include ATPases (CHD3 or CHD4), metastasis-associated proteins (MTA1, MTA2 or MTA3), MBD methyl-CpG-binding domain (MBD2 or MBD3) and retinoblastoma-associated-binding protein (RbBP4 and RbBP7). Deletion of

Mbd3 leads to failure in development of the inner cell mass of the embryo and defects in differentiation of mESCs (Kaji et al., 2006; Kaji et al., 2007).

Another complex, TIP60-P400 was also identified to function in stem cells by integrating NANOG binding and histone H3 lysine 4 trimethylation (H3K4me3) (Fazzio et al., 2008). When ISWI family NURF complex member bromodomain PHD-finger transcription factor (BPTF) is deleted, embryos also die at the early stages of embryo development and ESCs from such embryos are unable to form mesoderm and endoderm (Landry et al., 2008).

Furthermore, higher order chromatin organizers, such as the insulator protein CCCTC-binding factor (CTCF), which organizes chromatin into domains, are also regulated by pluripotency factors. These are thought to play a role in looping chromatin in such a way that pluripotency genes are expressed (Kim et al., 2011).

1.6 Applications of ESCs

The main application of mouse embryonic stem cells is in the creation of transgenic animals (Bradley et al., 1992). mESCs are relatively simple to genetically engineer, and when injected into embryos they can contribute to the germ line, leading to chimeric embryos and subsequently offspring that harbour mutations created in the stem cells (Capecchi, 2005). If the injected stem cells contributed to the germline, these animals can pass the mutations to their progeny, allowing a line to be established. This approach for creation of transgenic animals has been common and used very successfully since 1987, when a mouse with a mutation in the hypoxanthine guanine phosphoribosyl transferase (*Hprt)* gene was first engineered (Doetschman et al., 1987; Hooper et al., 1987; Kuehn et al., 1987).

In vitro cell culture differentiation of mouse embryonic stem cells is used as a model of early embryo development, including understanding pluripotency and exit from it to differentiation. They are much easier to obtain than cells from embryos or human embryonic stem cells. Additionally they proliferate quickly giving rise to large amounts of cellular material, which is needed for some types of experiments, such as ChIP-sequencing for example.

Furthermore, embryonic stem cells in combination with current gene editing technologies (such as CRISPR-CAS9) can be used to model human genetic variants associated with diseases to study the underlying molecular mechanisms (Merkle and Eggan, 2013).

1.7 Human embryonic stem cells

Human embryonic stem cells were only isolated in 1998 (Thomson et al., 1998), because their self-renewal seems to be regulated differently than in mESCs. Similarly to mESCs, hESCs express *POU5F1* and *NANOG* (Ginis et al., 2004). However, signalling *via* LIF and the STAT3 pathway is not important for pluripotency maintenance in hESCs (Dahéron et al., 2004; Reubinoff et al., 2000). A feeder layer of MEFs supplemented with bFGF or matrigel- or laminin-coated plates with addition of MEF-conditioned medium are used for culturing hESCs (Amit et al., 2000; Xu et al., 2001).

There is a notion that hESC are "later" in development than mESCs, and they are rather similar to epiblast stem cells (EpiSC) from the mouse (Tesar et al., 2007). EpiSC are clearly pluripotent, but when injected into a blastocyst stage embryo they do not colonize it (Brons et al., 2007; Huang et al., 2012). If hESCs are engineered to express *POU5F1*, *KLF4*, and *KLF2* transcription factors and are grown in the presence of LIF and the inhibitors GSK3β and

ERK1/2, they enter a different pluripotency state that resembles mESCs, suggesting that it is possible for hESCs to achieve naïve pluripotency (Hanna et al., 2010).

Human embryonic stem cells have huge potential for regenerative therapies. Differentiation of hESCs or induced pluripotent stem cells into tissues that are damaged or need replacing could be a solution to problems in transplant medicine, including the low number of organ donors and histocompatibility.

1.8 Sources and functions of cell-to-cell variability

For both mESCs and hESCs, cell-to-cell variability is an intrinsic feature of cells in cell culture. The function of heterogeneity within embryonic cell population is not very clear. It was proposed that it might be a result of cells transiently entering differentiation-primed states (Nimmo et al., 2015).

At the level of whole organisms, the key sources of heterogeneity are genetic differences. The genetic variation between organisms of the same species results in phenotypic variation, and is important for maintaining fitness of the population, especially in changing environments. Genetic variability is most visible and easily interpreted for simple Mendelian traits, such as blood type or Hemophilia A, but also for more complex traits including height (Wood et al., 2014) or susceptibility to type-2 diabetes (Morris et al., 2012).

Interestingly, monozygotic twins who have the same genetic make-up still exhibit considerable phenotypic differences. The discordance between monozygotic twins in both phenotype and behaviour is extensively studied in the context of health and disease. Monozygotic embryos start to differ even at

the early embryonic stages with, for example, differences in the initial number of cells in each embryo after division, or the position after implantation resulting in a slightly different environment (Machin, 1996). The discordance between monozygotic twins that arises during their lifetime has been attributed to differences in epigenetic marks that become increasingly divergent with time (Fraga et al., 2005). Epigenetic differences lead to differential gene expression, subsequent differences in protein amounts and activities, and ultimately to phenotypic variation between organisms.

As pointed out for mESCs and hESCs, the cells within one organism also differ. The most obvious differences between cells within an organism are encoded in the processes of development and differentiation to build tissues and cell types that perform different functions in the organism (Figure 1.8). Cell type is a poorly defined concept, but it is still used to describe these large functional differences between cells. A good example of a heterogeneous tissue with quite well defined cell types is an intestinal crypt, which is composed of stem cells and differentiated cells, including absorptive cells and several types of secretory cells such as Goblet and Paneth cells (Grun et al., 2015). The most important differentiation mechanisms involve the response of gene regulatory networks to signalling by growth factors or other molecules, and asymmetric divisions leading to the emergence of two different daughter cells (Morrison and Kimble, 2006). However, these processes are not entirely deterministic, and stochastic events are also an important factor (Losick and Desplan, 2008).

Apart from deterministic, hard-wired mechanisms that regulate cellular phenotypes, there are more subtle and stochastic sources of cell-to-cell variability (Figure 1.8). These are the main sources of heterogeneity within a

cell type or a seemingly homogeneous population of cells (Raser and O'Shea, 2005).

Firstly, cells differ due to the fact that each is in its own microenvironment with a particular level of nutrients, signalling molecules and environmental cues that affect cell state. Regional differences in the tissue, such as the amount of a particular signalling molecule, lead to slight differences in extracellular signalling, which influence intracellular signalling to different extents. Some signalling pathways are more robust to such changes than others. Similarly, the abundance of nutrients or oxygen, and interactions with other cells, shape cellular phenotype.

Secondly, the internal state of cells varies according to their individual histories. This means that the number and activity of molecules is often not exactly the same between cells. The transcriptomic state of a cell depends on its chromatin state and signalling state. For example, cells can differ in their cell cycle state. The cell cycle is a very dynamic process, and the expression of many genes depends on it. These include cell cycle regulators that are present at different points of the cell cycle, such as cyclins, and also other genes related to cell growth (Lim and Kaldis, 2013; Nurse, 2000; Vermeulen et al., 2003). For example transcription of histone mRNAs is upregulated in preparation for S phase when they are needed for packaging the new DNA strand. Globally, the level of all mRNAs increases during the cell cycle when the cell grows (Qiu et al., 2013). Other processes that play roles are for example uneven partitioning of mitochondria (Johnston et al., 2012; Mishra and Chan, 2014) and other molecules in the cell during cell division (Huh and Paulsson, 2011).

Thirdly, some variability emerges from the stochastic nature of biochemical processes. Many molecules within a cell are present as only a few copies, and

the reactions between them are infrequent. For example, the abundance of mRNA of a particular gene depends on the time at which it is measured: before or after a transcriptional burst. Transcription in eukaryotic cells does not happen at a constant rate, but in bursts. Over time, there are periods when the promoter of a gene is open, the transcriptional machinery is bound and the RNA molecules are synthesised in "bursts" or "pulses". These are followed by times when the gene is OFF and RNA is not synthesised. This behaviour can be quantified in terms of the average size of bursts and the frequency (*i.e.* how often these bursts occur). The extent of the cell-to-cell variability caused by stochastic transcription is related to the transcriptional burst size and frequency at the particular promoter. Mechanistically, expression bursts are dependent on the stochastic processes of transcription factors and RNA polymerase binding (Sanchez and Golding, 2013).

Finally, one has also to bear in mind the fact that within a living population of cells there are on-going somatic mutations that may contribute to the overall observed heterogeneity.

Before the development of high throughput single cell mRNA sequencing, variability between individual cells was measured by other means. For example, tagging a gene with a fluorescent protein and measuring the fluorescence of each cell using microscopy or FACS reveals cell-to-cell variation in the levels of particular proteins. In genetically identical cells taken from a homogeneous environment, heterogeneity (or "noise") can be measured using two fluorescent reporters, which allows one to discriminate between intrinsic and extrinsic noise (Elowitz et al., 2002; Swain et al., 2002). In the dual reporter system, intrinsic noise is defined as independent fluctuations

between the two marker proteins, while extrinsic noise are coupled fluctuations of both markers between cells.

From this example of an experimental definition of intrinsic *versus* extrinsic noise it follows that intrinsic noise is defined as noise within a single cell. Sources of intrinsic noise are usually the stochastic nature of cellular processes, the extent of which depends on the number of molecules involved (Rosenfeld et al., 2005). On the other hand, extrinsic noise describes cell-to-cell differences. Extrinsic noise can be caused by environmental factors or the state of the cell, such as the amount of particular transcription factor or cell cycle stage. Importantly, extrinsic noise may be global and affect all the genes in a cell or may affect only a subset, for example one signalling pathway.

Although we often use the word noise to describe variability between cells, it does not mean it is a meaningless and undesirable phenomenon. On the contrary, gene expression noise has been shown to have several functions in cell populations. Notably, noise in gene expression functions in gene regulatory circuits to create bistable switches between alternative cell fates. Amplifying noise can cause a cell to be randomly pushed towards one of two decisions. The decision that is made must be subsequently stabilized within the circuit. Networks containing bistable switches often exhibit a mechanism of hysteresis, which governs the kinetics of switching (Grimbergen et al., 2015; Veening et al., 2008). The existence of two alternative states of cells within the same environment is a basis for survival and a fitness strategy of bacteria known as bet-hedging. Bistable switches are common in prokaryotes but they are also present in eukaryotes (Palani and Sarkar, 2012; Shiraishi et al., 2010).

Heterogeneous gene expression is also implicated in developmental priming. Pluripotent or multipotent progenitor cells have the capability to

differentiate into different cellular fates. They do not lose this ability despite the stochastic expression of markers of lineages to which they have the potential to differentiate. It has been suggested that lineage priming makes it quicker and more efficient for cells to differentiate when the differentiation cues appear (Nimmo et al., 2015).

Increase in the heterogeneity of a population is often a vital part in complex cellular decision-making processes (Balazsi et al., 2011). Several transitions in cells have been shown to function in this way, such that there is an initial stochastic phase followed by a deterministic phase that ensures that cells move fully through the differentiation or developmental trajectory. This phenomenon occurs during reprograming of somatic cells to induced pluripotent cells (Buganim et al., 2012) and during polarisation of naive CD4+ T cells to Th1 and Th2 subtypes (Antebi et al., 2013; Fang et al., 2013).

In some cell types, for example neuronal cells or T helper cells, intercellular heterogeneity *in vivo* is large and there seems to be continuum of cell states with some metastable states that are more likely to be occupied by more cells (Zeisel et al., 2015). It has even been proposed by Sten Linnarsson to abandon the concept of cell type, as it is difficult to draw borders between states, and rather focus on describing the functions of each cell instead (oral communication).

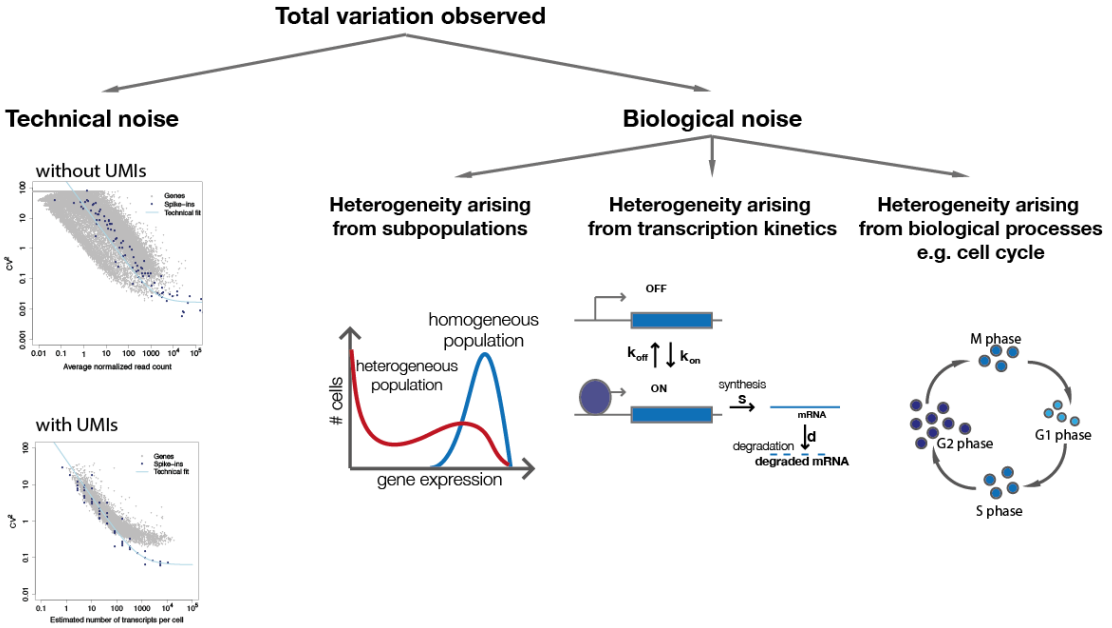

Figure 1.8 Contributors to noise

Decomposition of observed variation in scRNA-seq. Technical noise estimation based on synthetic spike-in molecules. Biological variation can be decomposed into (1) variation arising from the presence of subpopulations, (2) cell-to-cell variation in gene expression that can be estimated using the variance and from which transcription kinetic parameters can be modelled, and (3) biological variation due to cell function and biological processes such as cell cycle.

1.9 Single cell mRNA sequencing technologies

As mentioned above, heterogeneity in cell populations has been measured using fluorescent markers and microscopy or FACS for many years. FACS allows one to follow up to one or two dozen proteins at a time (Chattopadhyay et al., 2006), and mass cytometry increases the number of proteins to over 40 per cell (Bendall et al., 2011). Similarly, the proximity ligation assay (PLA) approach is limited to a predefined list of proteins for which antibodies are available (Soderberg et al., 2006).

For the detection of RNA, single cell qPCR (Bengtsson et al., 2008; Eberwine et al., 1992; Taniguchi et al., 2009; Warren et al., 2006) and single molecule FISH (Femino et al., 1998; Raj et al., 2006; Raj et al., 2008; Tyagi and Kramer, 1996) can be used to measure the amount of messenger RNA within a single cell. These approaches are also based on pre-selection of markers. Single cell mRNA sequencing revolutionised measurements of cellular heterogeneity, because it measures all highly and moderately expressed mRNAs in the cell and so does not require *a priori* knowledge about the genes of interest.

Each single cell mRNA sequencing experiment can be divided into the following steps: isolation of single cells, cell lysis, reverse transcription, amplification of cDNA, preparation of sequencing libraries and eventually sequencing (Kolodziejczyk et al., 2015a) (Figure 1.9).

The first and critically important step is to isolate single cells. Historically, in the first single cell mRNA experiments, single cells were selected and picked from the early embryo using micro pipetting (Grun et al., 2014; Tang et al., 2010; Tang et al., 2009). This method has an advantage that one can pick a cell from a particular position and virtually no cells are lost in the process.

Suspended single cells, such as blood cells, can be sorted into wells of a microtiter plate using FACS (Macaulay et al., 2016), they can be separated using microfluidic devices such as the Fluidigm C1 (Kolodziejczyk et al., 2015b; Mahata et al., 2014; Zeisel et al., 2015) or they can be encapsulated in nanoliter droplets (Mazutis et al., 2013). It is important to note that whereas many immune cell types naturally exist as single cell suspensions, other cells have to be dissociated from their tissue to become suspended. Dissociation is not trivial and requires enzymatic or mechanical approaches. Such treatment

may have an effect not only on the intactness and viability of cells, but also on their transcriptomes.

The key advantage of FACS is the possibility to sort for particular subpopulations that can be stained using surface markers. In addition, by index sorting, the intensity of the fluorescence as well as values for forward and side scatter can be recorded for each cell. This provides information about protein abundance, and cell size and granularity on top of the single cell transcriptomes (Hayashi et al., 2010). When dealing with known, rare cell types (e.g. blood stem cells) FACS can capture essentially all cells from the population of interest and sort them into individual wells. The main disadvantage of using FACS to sort single cells into microtiter plates are the microliter reagent volumes involved, which can be prohibitively expensive in large-scale experiments as compared to nanoliter volumes involved in microfluidics (Jaitin et al., 2014).

The Fluidigm C1 is a microfluidic platform that captures single cells (96 or 800 cells per chip) and performs reverse transcription and amplification of cDNA by PCR on chip. Since all these reactions are carried out in nanoliter volumes, this leads to lower reagent costs (Shalek et al., 2014; Trapnell et al., 2014; Treutlein et al., 2014). Importantly, this platform enables microscopic inspection of each cell upon capture, which allows identification of positions where multiple cells or debris were captured.

To capture 96 cells, one requires a starting population of at least 1000 cells, so this method is impractical for rare populations. An important limitation of this method is that cells being captured have to be homogeneous in size and compatible with one of the available capture site sizes (5–10, 10–17, and 17–25 microns in diameter). Nonspherical or sticky cells also do not capture well, but

at the same time, this capture method is much more gentle than FACS, and hence is better suited to delicate cell types such as neurons, megakaryocytes *etc*.

Recently, droplet-based microfluidics methods have been published, namely inDrop (Klein et al., 2015) and Drop-Seq (Macosko et al., 2015). These protocols encapsulate single cells in aqueous droplets within a surrounding oil phase. These droplets can be fused with other droplets to deliver reagents to perform lysis, reverse transcription and PCR. Reagent can also be delivered into droplets using picoinjection (Lee et al., 2014b). Several thousand cells can be analysed in one experiment using these methods. These methods will likely prove especially useful for surveying cells from different tissues to identify new cell types and cell functions.

Some less frequently used methods include laser capture microdisection (LCM), which is useful to pick cells from a particular position in a tissue. It is low throughput and does not necessarily guarantee that a single cell, rather than small group of cells is captured (Frumkin et al., 2008; Keays et al., 2005). Finally, nanoliter plates can be used for capturing single cells. Simply by adjusting the concentration of the cells in suspension, cells can be deposited and virtually every well will receive zero or one cell (Bose et al., 2015; Fan et al., 2015a).

To solve the problems caused by dissociation of cells from within tissues, methods for *in situ* transcriptome analysis are being developed, such as TIVA (Lovatt et al., 2014), FISSEQ (Lee et al., 2014a; Mitra et al., 2003) or padlock probe-based methods (Ke et al., 2013). These methods work for a limited number of genes and are also limited spatially by the resolution of the microscope.

In single cell mRNA sequencing and also other single cell protocols, the goal is to perform a single-tube reaction. Avoiding intermediate purification steps is crucial for avoiding nucleic acid losses, which reduce the sensitivity of the method. Captured cells are lysed by addition of lysis buffer containing detergent to disrupt the cell membrane. For plant or fungi cells, protoplasts must first be obtained by enzymatic or mechanical removal of the cell wall. Efficient cell lysis is important to release RNAs to the reaction and for the subsequent steps.

In the next step, RNAs are reverse transcribed, and this is a key step for achieving high sensitivity. A major goal of this stage is to avoid reverse transcribing rRNAs, which are high-abundance and would dominate any signal from the much lower abundance mRNAs. Due to the low abundance of mRNAs, common mRNA purification methods cannot be used. Most protocols (SmartSeq (Ramskold et al., 2012), Smartseq2 (Picelli et al., 2013), STRT-Seq (Islam et al., 2011), QuartzSeq (Sasagawa et al., 2013)) use polyT primers that bind to the polyA tail of mRNAs. This way only mRNAs and polyadenylated non-coding RNAs are reverse transcribed.

Alternatively, primers that are specifically designed not to bind to rRNAs have been used (Bhargava et al., 2013). The disadvantage of this approach is that there may be biases against some mRNAs. Finally, it was shown recently that random hexamer primers can be used (Armour et al., 2009; Fan et al., 2015b). Provided reverse transcription is performed at low temperature, most rRNAs are within folded ribosomes and are not transcribed. Moving beyond polyA priming would be useful for analyses of non-coding RNAs, such as circRNAs (Fan et al., 2015b), and also bacterial RNAs, which are of course not polyadenylated (Kang et al., 2011).

Second strand cDNA synthesis can be done using the template switching properties of the reverse transcriptase to minimize detection of partially transcribed species: this approach is used in SmartSeq (Ramskold et al., 2012). Alternatively polyA tailing and subsequent second strand synthesis priming from the polyA sequence can be used, but this leads to stronger 3′ bias of read coverage over transcripts, meaning that there are more reads mapping to the 3′ end of the transcript. This originates from incomplete reverse transcription, as in the first single cell sequencing protocol by Tang and colleagues and the QuartzSeq protocol (Sasagawa et al., 2013; Tang et al., 2009).

It is estimated that each cell contains around 10pg of mRNA (Ramskold et al., 2012), which will not produce sufficient cDNA for sequencing library preparation alone. Thus the cDNA must be amplified. This is done either by PCR or *in vitro* transcription followed by another round of reverse transcription. Most methods use PCR for amplification: SmartSeq(Ramskold et al., 2012), SmartSeq2 (Picelli et al., 2014), STRT (Islam et al., 2011), the Tang protocol (Tang et al., 2009), and SC3-seq (Nakamura et al., 2015). The main caveat of PCR is the fact that the exponential amplification that occurs may distort the relative amounts mRNA molecules. The alternative approach of *in vitro* transcription (IVT) was incorporated into the CEL-Seq (Hashimshony et al., 2012) and MARS-Seq (Jaitin et al., 2014) protocols. Amplification *via* IVT is linear but it leads to stronger 3′ biases due to the additional round of reverse transcription of the amplified RNA.

Sequencing libraries are prepared from amplified cDNA using the same protocols as for conventional bulk mRNA sequencing experiments and can be sequenced on any sequencing platform.

The optimal single cell RNA sequencing application depends upon the desired application. For discovery of new cell types, tag-counting droplet methods with high throughput are most advisable, while for analysis of allelic expression or splicing one must use a protocol that provides sequencing coverage of the entire length of mRNA molecules.

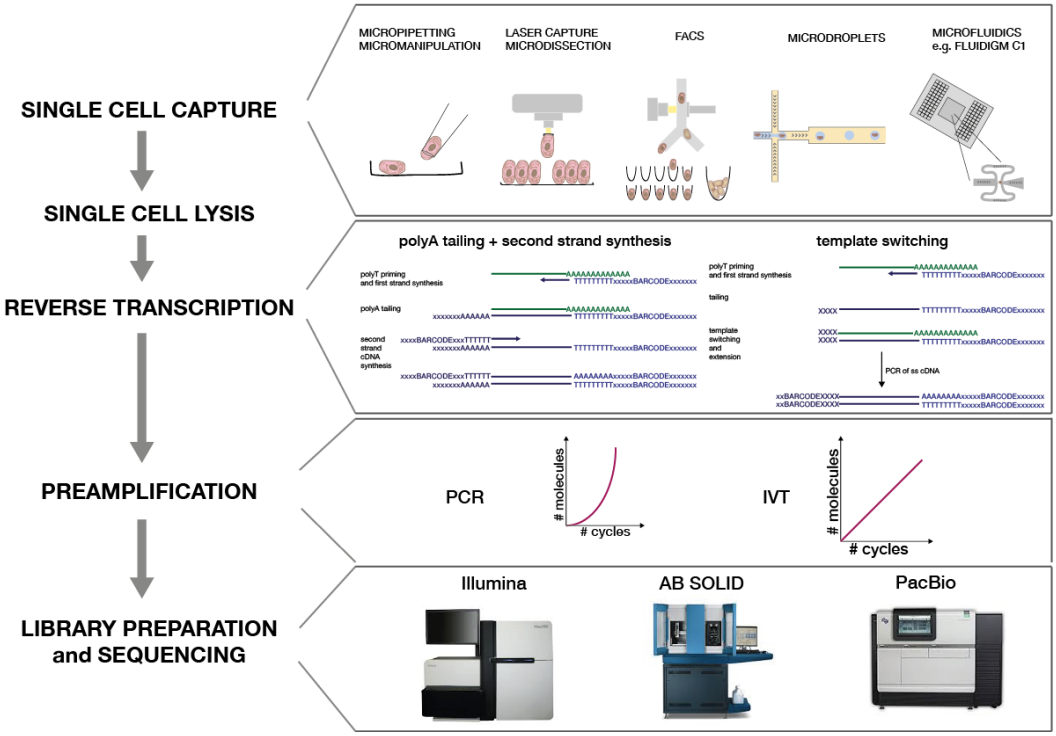

Figure 1.9 Single cell RNA sequencing workflow
On the left, steps common to all single cell experiments are shown, and on the right, different approaches that can be taken for each of them.

1.10 Technical variability in single cell mRNA-seq experiments

It is important to be aware that single cell RNA sequencing is subject to variation introduced by the experimental process rather than genuine biological differences between samples – technical noise.

Firstly, some technical noise originates from the reverse transcription step. The number of molecules in each cell is limited and it is estimated that only 10% of them are transcribed to cDNA with current technologies (Islam et al.,

2014). The molecules that are transcribed are selected stochastically. Due to Poisson sampling, the expression level estimation may not represent the original set of molecules from the cell, especially for lowly abundant mRNA species. Additionally, there may be a higher chance for some species of mRNA to be transcribed than others depending on their sequence and length of their polyA tails. These biases have not yet been systematically investigated.

Secondly, there is variation in the measurement from batch to batch. This may be due to differences between operators, batches of reagents or other factors.

Thirdly, single cell RNA sequencing data has the same biases as conventional RNA sequencing, such as PCR amplification bias, sequence bias during fragmentation and coverage biases. Importantly, more rounds of amplification are required than in bulk RNA sequencing providing more opportunities for the introduction of base substitutions. If amplification is performed using PCR, then PCR amplification biases are also present. It was also reported that reverse transcription with poly-dT priming leads to 3' bias in read coverage (Mortazavi et al., 2008; Ramsköld et al., 2012). This is also the case in bulk-level experiment that uses poly-dT priming.

To estimate some sources of bias and technical error it has proved very useful to add ('spike-in') an external standard into each cell prior to lysis. ERCC Spike-In is the most commonly used, commercially available set of control molecules and it consists of 92 synthetic polyadenylated mRNA species of different known concentrations (Jiang et al., 2011). These were designed so as to lack sequence similarity to any known eukaryotic genome. It allows one to measure the sensitivity and accuracy of each experiment, as well as perform correction of some batch effects. It is also used for estimation of the

extent of technical noise (Brennecke et al., 2013). ERCC spike ins can be used to produce a calibration curve to estimate the absolute number of molecules in each cell (Kivioja et al., 2012). It has to be noted that ERCC molecules do not go through cell lysis and are not associated with proteins, thus are not subjected to all the processes that cellular mRNAs are. Furthermore they are not capped, and they have very short polyA tails in comparison to endogeneous mRNAs.

In addition to ERCCs, one can use unique molecular identifiers (UMIs), which are highly diverse, random, unique barcodes for tagging each cDNA molecule generated during reverse transcription (Fu et al., 2011; Islam et al., 2014; Shiroguchi et al., 2012). They enable one to count molecules by counting the number of unique UMI sequences associated with each transcript instead of counting the number of sequencing reads that map to a particular transcript. This can ameliorate PCR biases (Kivioja et al., 2012). The main disadvantage of UMIs is that until now they have only been used for methods that count the 3' end of molecules. In addition, to estimate the number of molecules one has to sequence deeply, and UMI methods also tend to overestimate noise for highly expressed genes.

Technical variability within an experiment can be also estimated by performing pool and split experiments (Deng et al., 2014; Marinov et al., 2014) and using a known amount of standardized extracted RNA (Brennecke et al., 2013).

1.11 Single cell mRNA sequencing applications

Single cell mRNA sequencing is an unbiased and straightforward way to survey cellular populations to describe the cells that are present. Tissue functions depend on the identity and frequencies of cell types within the

tissue. By sequencing all cells in the tissue one can find new cell types that have not been described previously. For example, by sequencing all cells from intestinal crypts, a new secretory cell type was discovered (Grun et al., 2015). Once a new subpopulation of cells is identified, it is quite straightforward to identify a set of reliable cellular markers for this particular population using differential expression analysis, correlation analysis (Mahata et al., 2014) or random forest approaches (Macaulay et al., 2016) (Figure 1.10). We performed single cell mRNA sequencing on a population of differentiating mouse CD4+ T-helper 2 cells and identified LY6C1/2 as a cell surface marker for a population within these cells that produces steroid and appears to be immunosuppressive (Mahata et al., 2014). Similarly, mitotic markers of radial glia that allow staging them according to their cell cycle progression were identified (Pollen et al., 2014).

The identification of groups of cells that have similar transcriptomes is a challenge (Figure 1.10). The choice of clustering approach and the similarity measure that is used depends on the particular biological system, the composition of the population and relative differences between cells. Thus, several approaches have to be tested to find the optimal one with good separation and compactness of clusters and that accurately represents the biological system under study. One of the indicators can be the compactness of clusters, measured by the sum of squares within groups, which should be significantly lower than that of randomly permuted data (Treutlein et al., 2014). Usually only moderately and highly expressed genes are used, because lowly expressed ones have a high level of technical noise that interferes with clustering. Alternatively, one can use a set of highly variable genes for clustering (Jaitin et al., 2014). They can be identified by calculating their

coefficient of variation, or preferably by identifying genes that are more variable than is expected by chance by modelling technical noise using the spiked-in standards (Brennecke et al., 2013). Validation of the clusters is usually done by examining expression of particular cell markers and assigning them to clusters.

Other commonly used methods for identification of subpopulations are dimensionality reducing visualisation methods such as principal component analysis (PCA) (Figure 1.10). Using PCA it was shown that to be able to separate cells from different tissues, namely as blood, epidermal, and pluripotent cells and neurons one needs only very shallow sequencing, and expression levels of 500 most expressed genes, when cells were sequenced to 10,000 reads per cell is enough (Pollen et al., 2014).

A nonlinear dimension-reduction method, t-distributed stochastic neighbour embedding (tSNE) (Van der Maaten and Hinton, 2008) is a machine-learning algorithm that models the data in such a way that similar cells are placed near each other. Importantly the distances on this plot, unlike on PCA do not correspond to how similar points are to each other. Initially, this method was slightly modified and very successfully used on mass cytometry data from bone marrow cell samples (Amir et al., 2013) and subsequently it has been adopted to single cell mRNA sequencing data to show subpopulations in differentiating mouse embryonic stem cells (Klein et al., 2015), 39 subpopulations of cells from retina (Macosko et al., 2015) or nine major classes of cells from mouse cortex (Zeisel et al., 2015).

Single cell mRNA sequencing data often have many zero values due to dropout events (Lun et al., 2016), which may lead to misleading results in methods such as PCA. To address this problem a dimension-reduction

approach called Zero Inflated Factor Analysis (ZIFA) was established. This method uses a latent variable factor analysis model and models the dropout rate to accommodate zeros within the data (Pierson and Yau, 2015).

SNN-Cliq is method bases on the shared nearest neighbour (SNN) similarity measure. Rather than using numerical values of gene expression it uses ranking of similarities between gene expression values (Xu and Su, 2015).

Other approaches for reducing the dimensionality of scRNA-seq data include self organizing maps (SOMs) (Kim et al., 2015a), circular *a posteriori* projection (CAP) (Jaitin et al., 2014), BackSPIN clustering (Zeisel et al., 2015), single-cell clustering using bifurcation analysis (SCUBA) (Marco et al., 2014). New methods are published regularly.

Provided that a sufficiently large number of cells is surveyed it is possible to find rare or outlier cells within a population. Although rare, these cells are often involved in important functions and are biologically relevant. These include stem cells within tissues, secretory cells and rare cell populations within tumours, which may convey resistance to a particular drug. Once identified using single cell sequencing they can be enriched for using cell surface markers discovered in the singe cell mRNA sequencing data (Grun et al., 2015).

Furthermore, single cell sequencing opens an avenue for sequencing unicellular organisms that cannot be cultured in conventional media and cannot be obtained in large quantities (Marcy et al., 2007; Gawad et al., 2016; Proserpio et al., 2016). Similarly, single cell mRNA sequencing was applied to profile early human embryos (Yan et al., 2013; Petropoulus et al., 2016), which are very limited and one could not easily obtain enough cells to sequence them using conventional methods.

Single cell transcriptomic data aid understanding processes where cells traverse from one state to another and where cellular decisions are being made. The transition between states can be binary or gradual and may or may not involve discrete intermediate states. Analysis of gene expression changes throughout the transition can give an insight into transcriptional waves that often accompany them. Key genes and transcription factors that act as switches to drive the process can be identified from such analyses.

Although single cell mRNA sequencing provides only a snapshot of a population in given time, one can take advantage of the fact that cells are not synchronized and so order them along the process they undergo such as development or differentiation. This ordering, places the cells along an axis referred to as 'pseudotime'. These approaches provide temporal resolution without performing time course experiments, or allow additional information to be extracted from time course data. Ordering cells along the process is performed by several algorithms developed for this purpose. The first method that was developed to serve this purpose was Monocle (Trapnell et al., 2014), it first uses independent component analysis (ICA) for dimensionality reduction and subsequently constructs a minimal spanning tree (MST) through the data points. The longest possible path through the MST is taken to represent pseudotime. An important limitation to Monocle is that one has to specify number of bifurcations that occur in the data. Waterfall is similar to Monocle but it uses clustering and PCA for dimensionality reduction instead of ICA, and then it also draws an MST to find the longest path through the cells (Shin et al., 2015). Moreover diffusion maps were successfully used for defining developmental trajectories (Angerer et al., 2015; Haghverdi et al., 2015; Julia et al., 2015).

All above-mentioned methods assume that the process being analysed is directional, but there are phenomena in biology, which are oscillatory, and the most important example is cell cycle. For analysis of such processes Oscope was developed (Leng et al., 2015). It uses gene co-expression to identify, which genes oscillate and using them orders cells in a cyclic fashion.

If genetic information of maternal and paternal alleles is known, as in the case when two genetically distinct mouse strains, such as BL6 and CAST are crossed, single cell mRNA sequencing can give information about expression of genes at allelic resolution. This gives more information than just identifying monoallelic and imprinted genes (Deng et al., 2014). The heterogeneity of the ratio between alleles in each cell gives us information about gene expression noise and allows dissection of the noise between intrinsic cellular processes and extrinsic stimuli (Kim et al., 2015b).

Knowing the composition of noise and heterogeneity of each allele allows modelling of gene expression kinetics at each promoter. Kinetics of transcription factor binding, which result in specific burst sizes and frequencies can be fitted to the noise level at each promoter. If additional factors such as degradation rates of mRNA are known they can be incorporated into such models (Kim et al., 2015b).

Finally, single cell mRNA sequencing enables investigation of gene regulatory networks in naturally perturbed systems. Gene regulatory modules can be identified by calculating correlations or by clustering cells. In such networks, transcripts of genes are nodes and co-expressions of these genes are the edges. To analyse how genes interact with each other the networks must be perturbed. Cells in the population can be undergoing transitions such as differentiation, or they can respond to an extracellular signal that affects their

transcription. The weighted gene co-expression network analysis (WGCNA) approach was developed for bulk samples (Zhang and Horvath, 2005) but it was also successfully used for analysis of single cell data (Moignard et al., 2015; Xue et al., 2013).

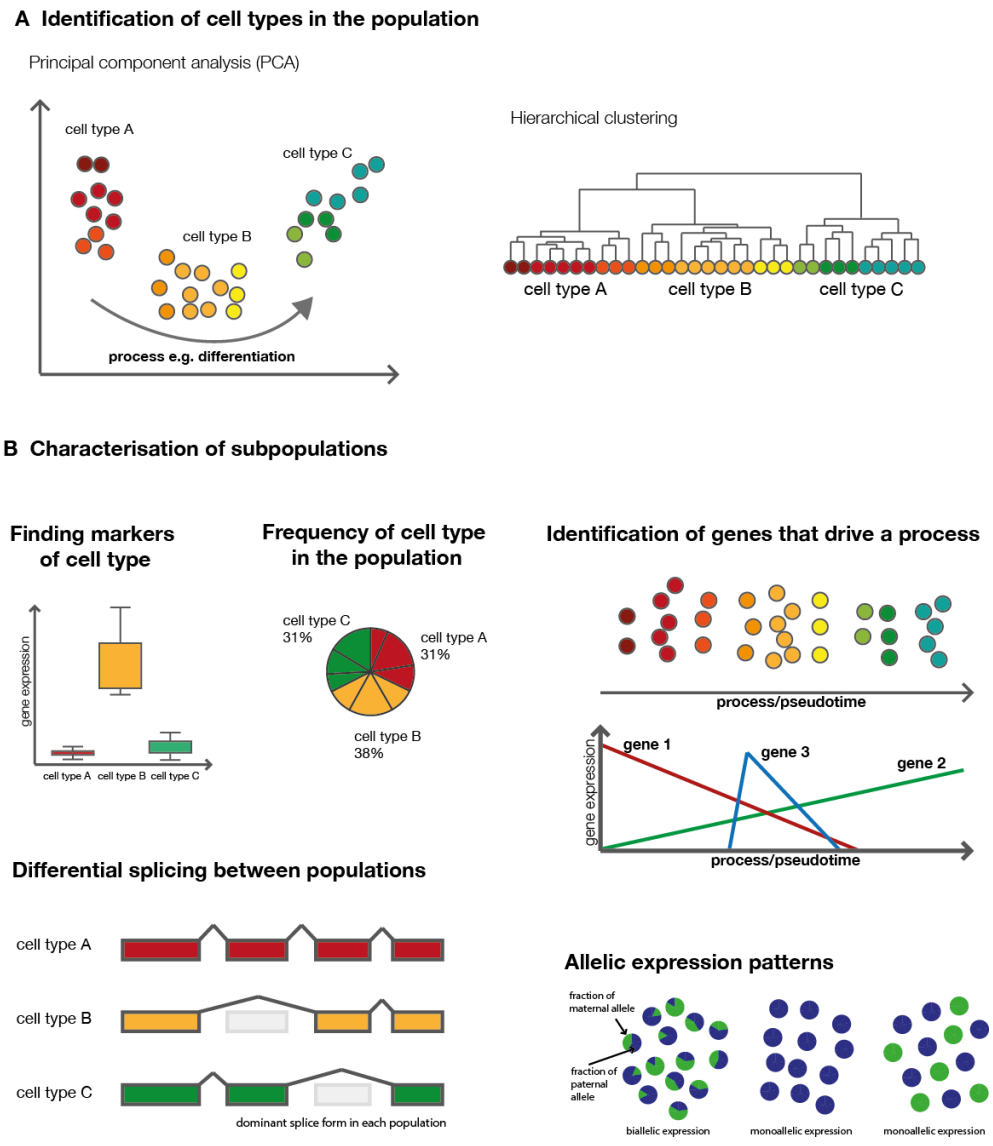

Figure 1.10 Identification and Characterization of Cell Populations

(A) Identification of cell populations can be performed using principal component analysis (PCA) or hierarchical clustering. (B) Different approaches to subpopulation characterization: finding markers of cell types by analysing differential expression between different groups of cells; frequency of cell populations; identification of genes that have particular patterns during a process such as development or response to stimuli: genes that either increase or decrease expression throughout the process,

but most interestingly genes that are expressed transiently in the intermediate cell types, as these genes may be important for the process to proceed; differential splicing analysis: differential splice variants may divide population of cells in to subpopulations; and analysis of allele-specific expression patterns: if a sample of heterogeneous genetic background, such as a cross of mice from two genetically distant inbred lines is provided, imprinted and monoallelically expressed genes can be identified.