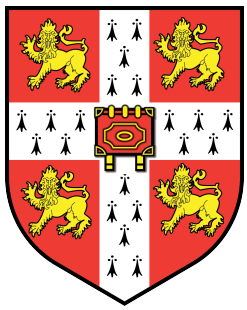
Single cell mRNA-sequencing of mESCs reveals cell-to-cell variation in pluripotency and cell cycle genes



Aleksandra Anna Kołodziejczyk

Trinity College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
May 2016

Summary

Cell culture conditions for embryonic stem cells are important for their selfrenewal capacity and for them to maintain pluripotency. Depending on the media that cells are cultured in, they exhibit different morphology and gene expression patterns. It was shown that ES cells cultured in 2i versus serum results in cells with more homogeneous morphology and more uniform Nanog expression.
I analysed the transcriptomes of over 700 individual mESCs cultured in three conditions (serum, 2i and alternative 2i) using full-transcript single cell RNAsequencing to understand the causes of culture medium-dependent differences in gene expression variability. I aimed to quantify and dissect the cell-to-cell variation in the three conditions in an unbiased way by high-throughput single cell mRNAsequencing and statistical data analysis in a way that was not possible before.
Firstly, I found that global levels of intercellular heterogeneity in gene expression are indistinguishable between conditions. At the same time, specific groups of genes (pluripotency genes in serum, cell cycle genes in 2i) do differ in their noise levels across culture conditions. The heterogeneity of pluripotency genes in the serumcultured mES cells is a consequence of subpopulations of cells that are differentiating away from the pluripotent state. In 2i and a2i-cultured cells, the transcriptomic heterogeneity originated in gene expression signatures of different cell cycle stages.
Secondly, I showed that the transcriptomic signatures of cells grown in the three media are distinct, with cells grown in 2i medium being most similar to the blastocyst cells of the early embryo.
Additionally, I found that differences in cell cycle genes' noise profiles correlate with proliferation rate, where slowly-cycling cells have broader, more noisy expression profiles and clearer separation between cells in G1/S and G2/M phases.
Moreover, I observed a previously described but poorly understood 2C-like population in 2i-cultured cells. I characterized this population in detail and compared it to in vivo data from early stages of mouse embryo development to determine whether it truly is equivalent to the embryonic 2-cell stage. I observed that these cells globally are more transcriptionally similar to blastocyst cells than cells from the 2-cell stage of the embryo.
Finally, I investigated the pluripotency gene regulatory network by analyzing correlations between transcription factors and chromatin-associated genes in the mouse ES cell data. I found two major clusters: pluripotency factors and differentiation regulators. In the pluripotency cluster, I identified new putative pluripotency regulators (Ptma, Zfp640, Zfp710). I validated these by knockdown with CRISPR repression technology, and demonstrated that even partial depletion of these genes causes a shift towards a more differentiated state.
Single cell RNA sequencing allowed me to look at cell populations and genes in the dataset to unravel cell identities and genes that regulate processes in these cells. This work highlights the power of single cell sequencing whilst providing data and analytical approaches that will be a useful resource for further study.

Declaration

The work presented in this dissertation was carried out at the MRC Laboratory of Molecular Biology, the EMBL European Bioinformatics Institute and the Wellcome Trust Sanger Institute between October 2012 and May 2016. This thesis is the result of my own work except when explicitly stated in the main text and is unlike any work I have previously submitted for any other qualification. This thesis does not exceed the word limit of 60,000 words required by the University of Cambridge School of Biological Sciences.

Aleksandra A. Kołodziejczyk
May 2016

Acknowledgements

Although the last four years were challenging, overall my time in Cambridge has made me really happy both personally and professionally and I have achieved more than I hoped for when I started my PhD. There are many people who made this happen and I feel I should thank.

First and foremost, I am deeply grateful to my supervisor Sarah Teichmann for guidance, support, patience, and trusting me. I was really lucky to be able to work in the exciting and emerging field of single cell genomics, which together with Sarah's optimism and insights allowed me to make new discoveries. Naturally, having a supervisor who taught me how to do scientific research both in and outside of the lab will impact my whole future career.

Secondly, I would like to thank John Marioni for his advice and discussions throughout our fruitful collaboration. I extend my gratitude to my collaborators: especially to Jong Kyoung Kim, for his invaluable help with bioinformatics and statistics; and to Jason Tsang, Xuefei Gao and Pentao Liu for all of the discussions about the biology of stem cells and for their help with the experimental parts of my project.

I would like to thank all current and previous members of the Teichmann lab at LMB, EBI and Sanger. I have to mention Tomislav Ilicic for assistance with various computational aspects of the project and his friendship, and Xiuwei Zhang for patience, teaching me basics of R and all Ragusa chocolate she brought from Switzerland. I am grateful to Tina Perica, Valentina Proserpio, Liora Haim-Vilmovsky and Bidesh Mahata for warm welcome and showing me around when we were still at the LMB. Furthermore, I am thankful to Kedar Natarajan for our discussions about stem cells, to Alex Tuck for sharing his data with me and to Johan Henriksson for turning my dataset into a database.

This thesis would not read well without the editing assistance and the 'unPolishing' of my English by Mike Stubbington.

This PhD would not have been possible without the support, love and hours on Skype with mom, dad and Paweł.

I would like to thank my friends: Łukasz Kopeć, Filip Szczypiński, Monika Folkierska-Żukowska, Julia Majewska and Kasia Wojtczak for their support, for visiting me, for telling me I do not have nine lives, for Cake/Vodka Mondays, and for words like "doktormatka". My friends from Trinity College: Rebecca Berrens, Ana Casanova, Annette LaRocco, Sun Lee, Tilman Flock, James Kane, Andreas Jakowetz, Matthew Dunstan, Janina Voigt and Tobias Schmidutz, who were always there for me and who helped to provide much needed fun in Cambridge and during our trips in Spain, Germany and Poland.

My life in Cambridge has been enriched by activities outside of my doctoral research, especially with my friends from CU Polish Society and Federation of Polish Societies in UK: Michał Włodarski, Dominika Kampa, Dominika Wolańska, Kasia Doniec, Tomek Cebo, Ola Pędraszewska, Marta Tondera, Czarek Łastowski and Kasia Rachuta. Thanks for the conferences, cultural events and fun we had organizing and running the society. Special thanks go to Tamás Sztanka-Tóth, for our Polish-Hungarian friendship.

I also need to thank Victor Sourjik (and all Sourjik lab members) and Dmitry Veprintsev, who inspired and encouraged me. Without them on my path I probably would not have attempted to pursue a PhD in the first place.

Finally, I would like to thank the BBSRC, Abcam and Wellcome Trust for funding to support my research.

Summary

Cell culture conditions for embryonic stem cells are important for their self-renewal capacity and for them to maintain pluripotency. Depending on the media that cells are cultured in, they exhibit different morphology and gene expression patterns. It was shown that ES cells cultured in 2i *versus* serum results in cells with more homogeneous morphology and more uniform *Nanog* expression.

I analysed the transcriptomes of over 700 individual mESCs cultured in three conditions (serum, 2i and alternative 2i) using full-transcript single cell RNA-sequencing to understand the causes of culture medium-dependent differences in gene expression variability. I aimed to quantify and dissect the cell-to-cell variation in the three conditions in an unbiased way by high-throughput single cell mRNA-sequencing and statistical data analysis in a way that was not possible before.

Firstly, I found that global levels of intercellular heterogeneity in gene expression are indistinguishable between conditions. At the same time, specific groups of genes (pluripotency genes in serum, cell cycle genes in 2i) do differ in their noise levels across culture conditions. The heterogeneity of pluripotency genes in the serum-cultured mES cells is a consequence of subpopulations of cells that are differentiating away from the pluripotent state. In 2i and a2i-cultured cells, the transcriptomic heterogeneity originated in gene expression signatures of different cell cycle stages.

Secondly, I showed that the transcriptomic signatures of cells grown in the three media are distinct, with cells grown in 2i medium being most similar to the blastocyst cells of the early embryo.

Additionally, I found that differences in cell cycle genes' noise profiles correlate with proliferation rate, where slowly-cycling cells have broader, more noisy expression profiles and clearer separation between cells in G1/S and G2/M phases.

Moreover, I observed a previously described but poorly understood 2C-like population in 2i-cultured cells. I characterized this population in detail and compared it to *in vivo* data from early stages of mouse embryo development to determine whether it truly is equivalent to the embryonic 2-cell stage. I observed that these cells globally are more transcriptionally similar to blastocyst cells than cells from the 2-cell stage of the embryo.

Finally, I investigated the pluripotency gene regulatory network by analysing correlations between transcription factors and chromatin-associated genes in the mouse ES cell data. I found two major clusters: pluripotency factors and differentiation regulators. In the pluripotency cluster, I identified new putative pluripotency regulators (*Ptma*, *Zfp640*, *Zfp710*). I validated these by knockdown with CRISPR repression technology, and demonstrated that even partial depletion of these genes causes a shift towards a more differentiated state.

Single cell RNA sequencing allowed me to look at cell populations and genes in the dataset to unravel cell identities and genes that regulate processes in these cells. This work highlights the power of single cell sequencing whilst providing data and analytical approaches that will be a useful resource for further study.

Publications

1. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA (2015) Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 17 (4), 471-485.

2. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The Technology and Biology of Single-Cell RNA Sequencing. Molecular cell 58 (4), 610-620.

3. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy D, Marioni JC, Teichmann SA. (2016) Classification of low quality cells from single cell RNA-seq data. Genome Biology 17 (1), 1.

4. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun 6, 8687.

5. Tsang JCH, Yu Y, Burke S, Buettner F, Wang C, Kolodziejczyk AA, Teichmann SA, Lu L, Liu L. (2015) Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biology 16 (1), 1-16.

6. Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, Hebenstreit D, Dingler FA, Moignard V, Göttgens B, Arlt W, McKenzie ANJ, Teichmann SA (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep. 22;7(4):1130-42.

7. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 10: 1093–1095.

Contributions

This thesis is the result of my own work except:

- CV^2 plots showing technical noise on Figure 1.8 were generated by Dr Jong Kyoung Kim.
- Microscopy pictures on Figure 3.5 were taken from Roeder and Radtke, 2009.
- Cell culture was done in collaboration with Dr Jason Cheuk-Ho Tsang and Dr Xuefei Gao.
- NPC differentiation dataset was kindly shared with us by Dr Alex Tuck.
- mRNA sequencing libraries of bulk controls (not CRISPR experiment) were done by Wellcome Trust Sanger Institure sample preparation pipeline team.
- dCas9-Krab and hyPBase plasmids as well as PB-gRNA-BsaI plasmid backbone are kind gift of Dr Xuefei Gao.
- Gene expression heterogeneity measurement using DM was developed by Dr Jong Kyoung Kim.
- Plots showing DM level on Figure 3.8 and Figure 3.9 were generated by Dr Jong Kyoung Kim.
- Single cell sequencing data batch correction was done by Dr Jong Kyoung Kim.

Table of contents

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).

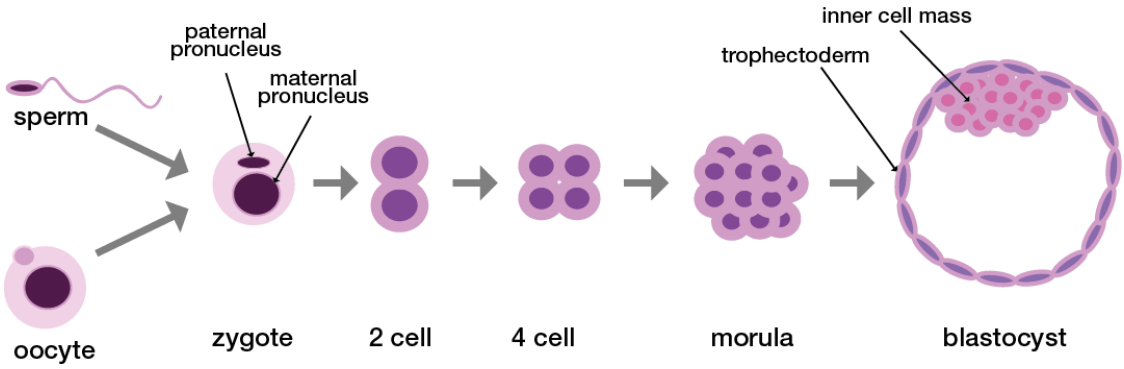


Figure 1.1 Early embryo development
During fertilisation sperm and oocyte combine to form a zygote. It divides giving rise to more totipotent cells. The first two lineages are formed at the blastocyst stage where some cells form a trophectoderm layer which encapsulates the second type of cells -inner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

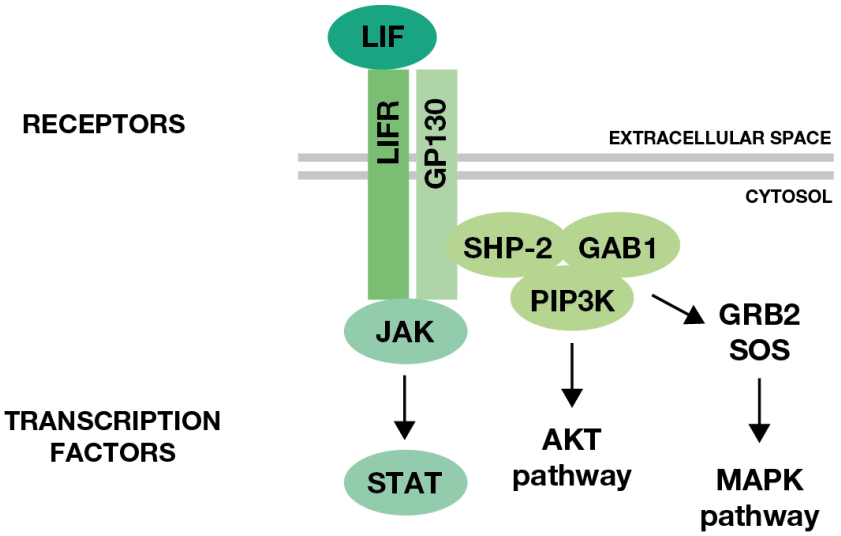


Figure 1.2 LIF signalling

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).

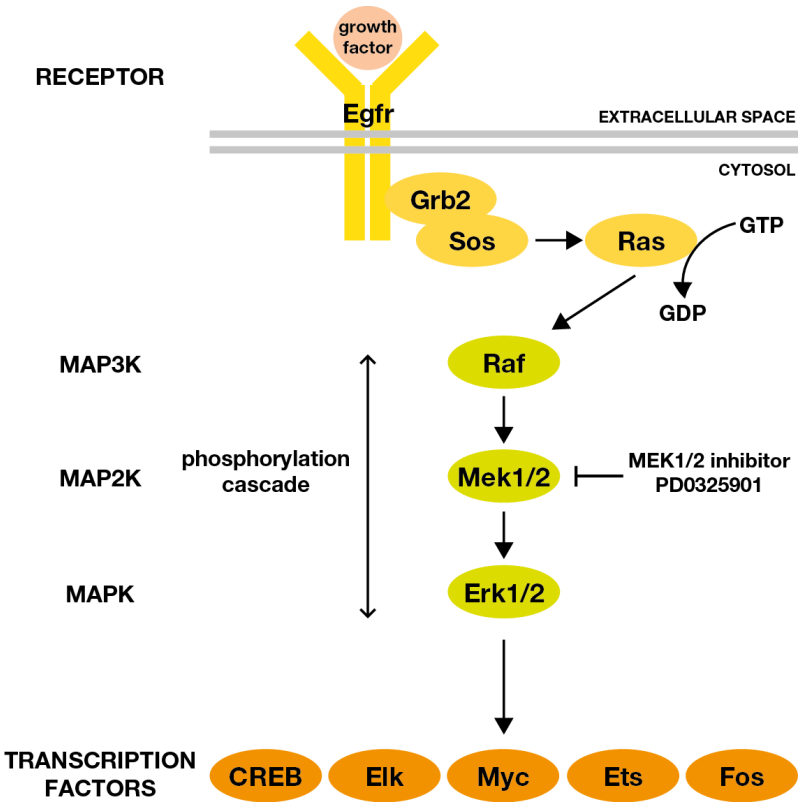


Figure 1.3 MAPK signalling

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).

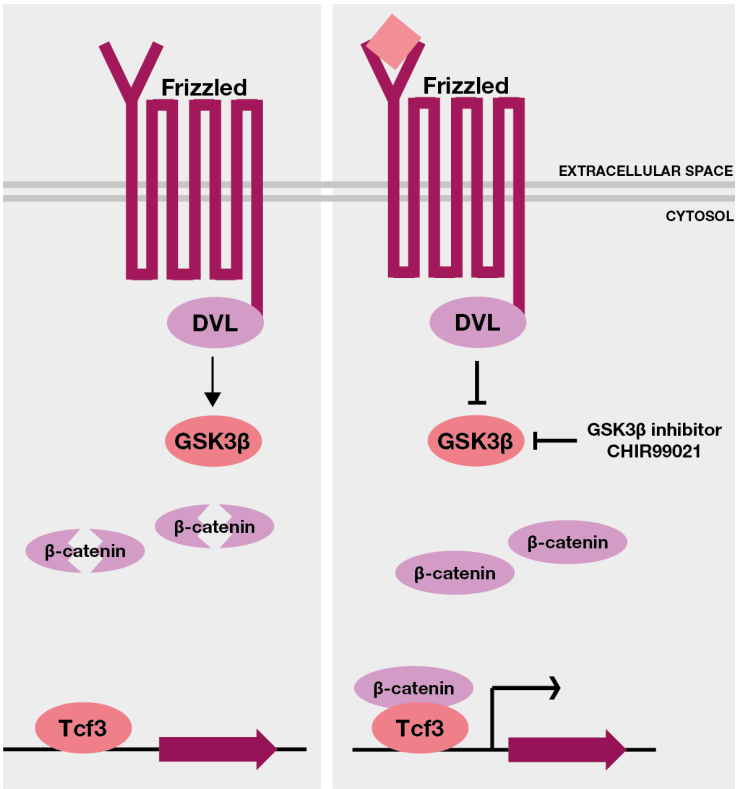


Figure 1.4 Wnt signalling

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).

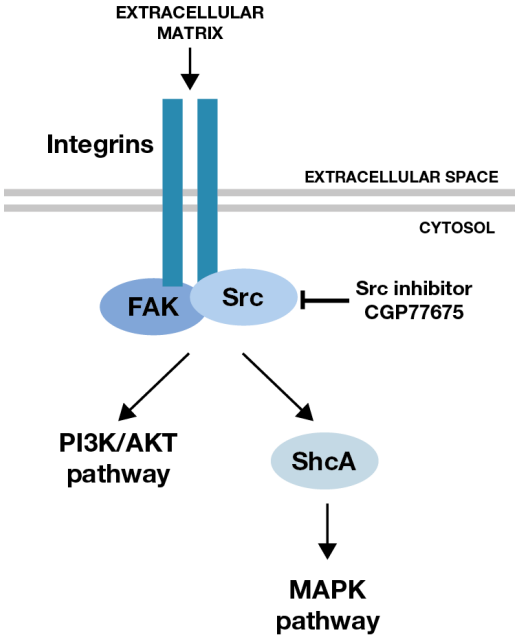


Figure 1.5 Src signalling

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).

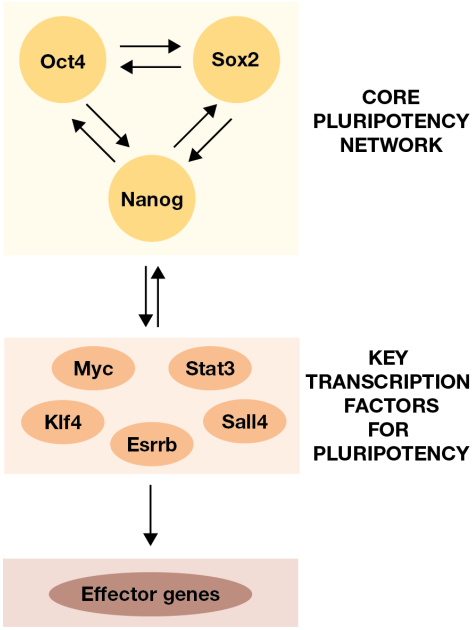


Figure 1.6 Pluripotency network

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).

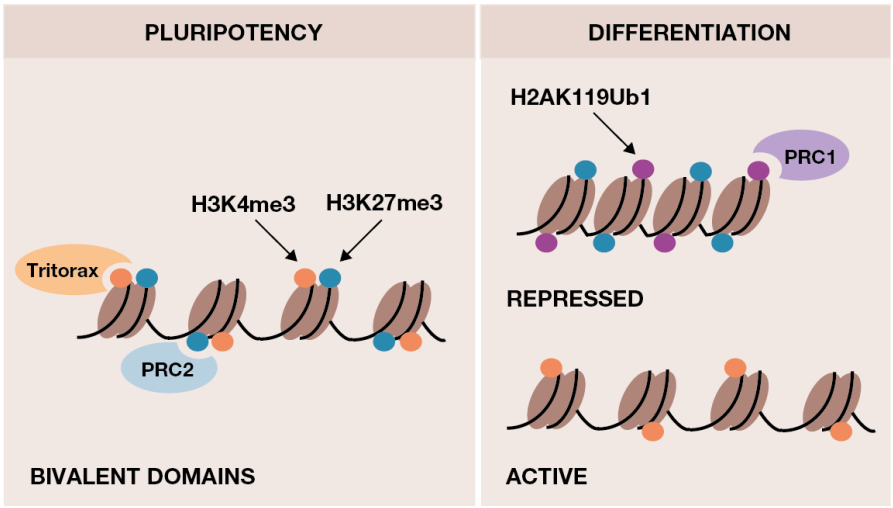


Figure 1.7 Histone modifications in pluripotent and differentiated cells
The promoters of tissue-specific genes and pluripotency genes include both active (H3K4me3) and repressive (H3K27me3) marks deposited by the Trithorax and PRC2 complexes respectively. Upon differentiation, these domains either loose 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).

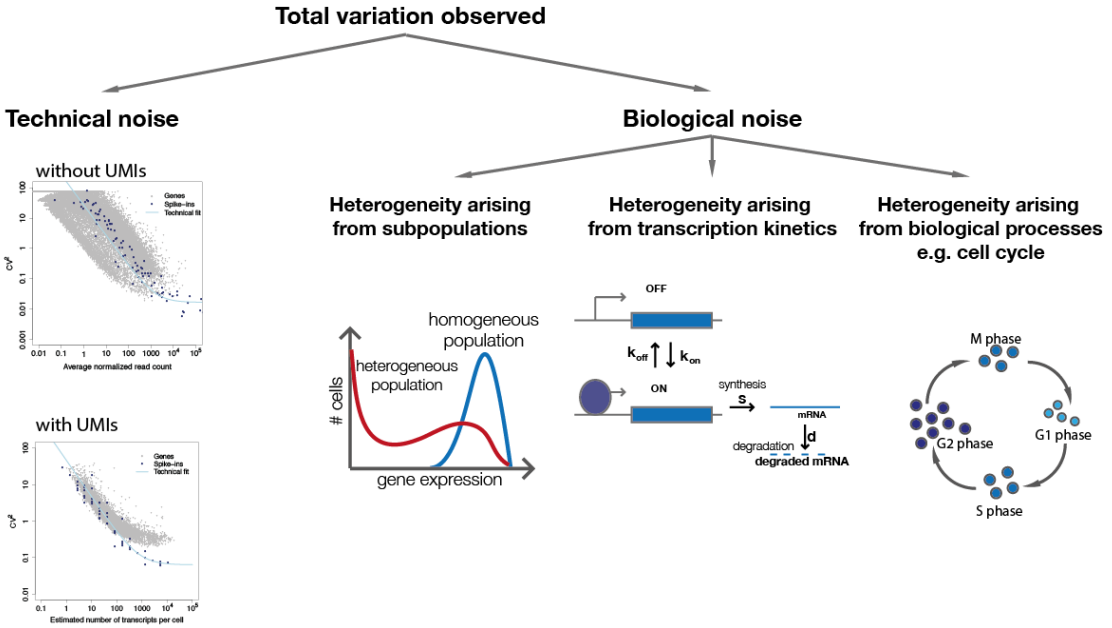


Figure 1.8 Contributors to noise

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.

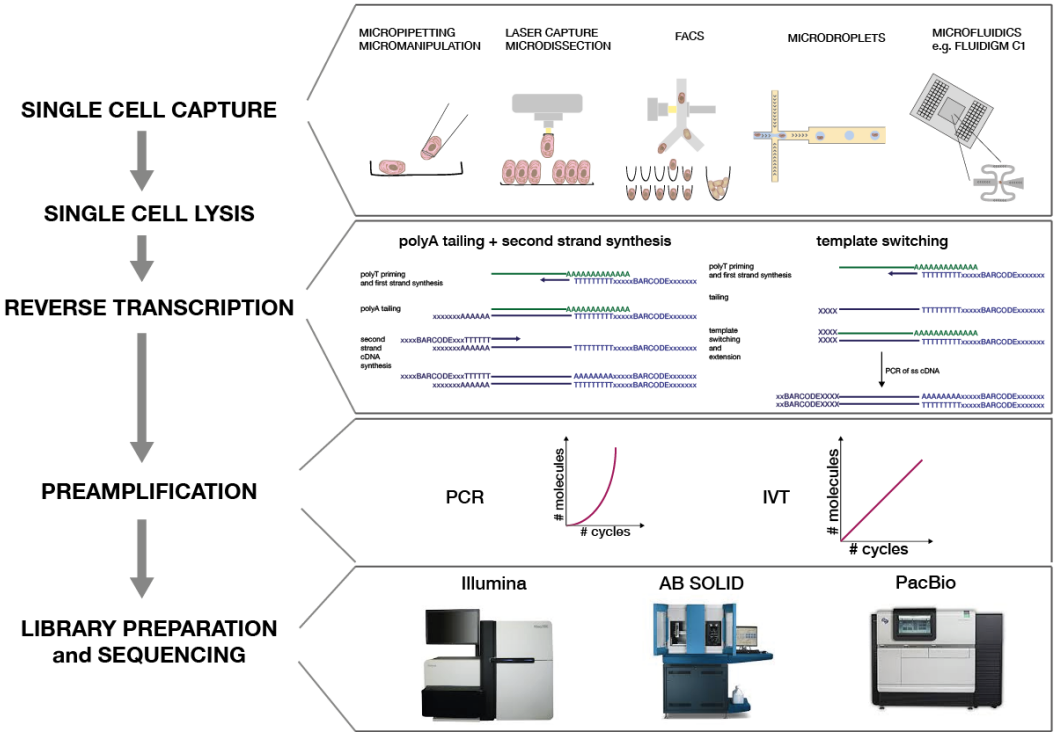


Figure 1.9 Single cell RNA sequencing workflow
On the left, steps common to all single cell experiments are shown, and on the right, different approaches that can be taken for each of them.

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).

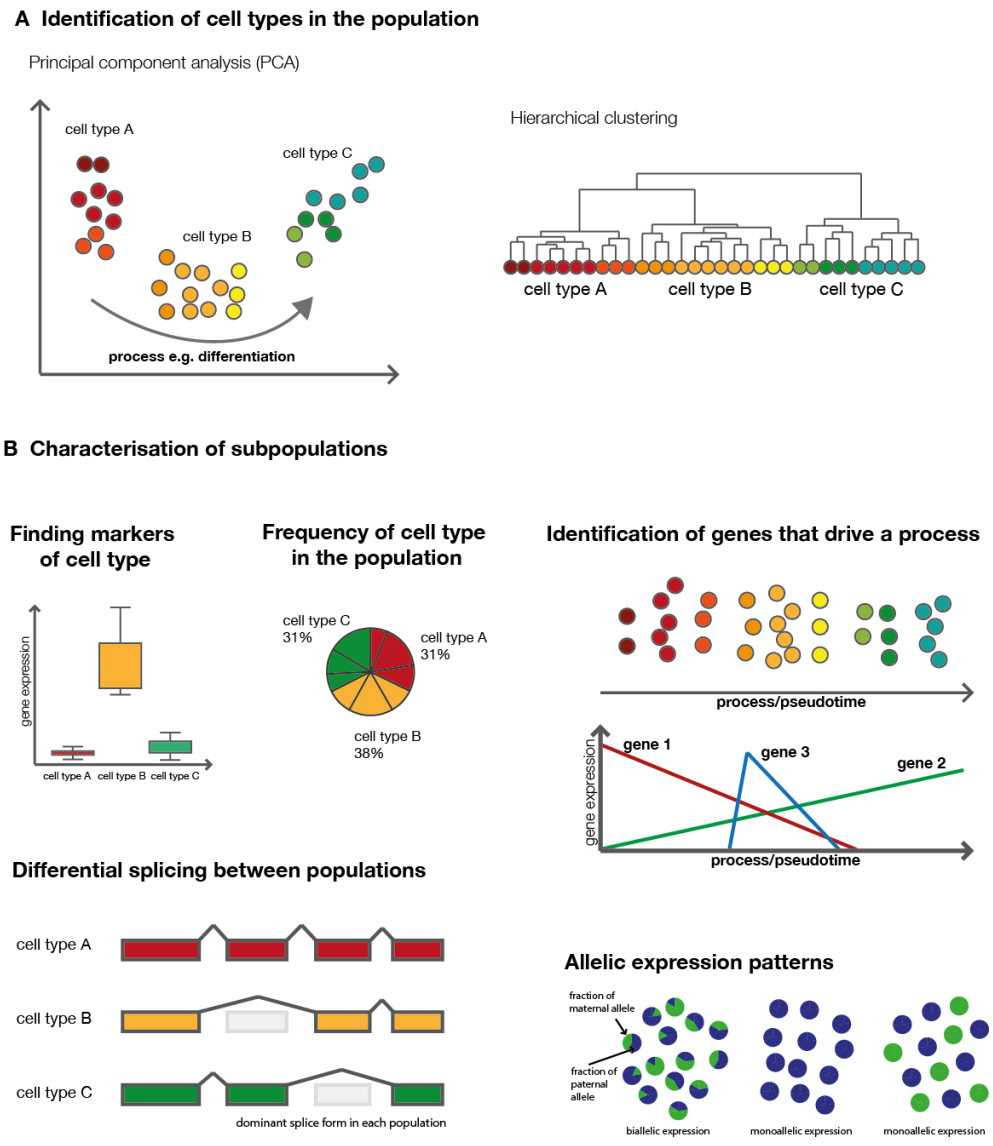


Figure 1.10 Identification and Characterization of Cell Populations

(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.

Chapter 2

Materials and Methods

2.1 Cell culture conditions

Cell cultures were done in collaboration with Dr. Jason Tsang. The G4 (C57BL/6Ncr x 129S6/SvEvTac) mouse hybrid (George et al., 2007) embryonic stem cells were obtained from Mount Sinai Hospital and were maintained on STO feeders in serum-containing media at 5% CO2 and 37°C. They were sub-cloned, and a line with normal karyotype was selected based on spectral karyotyping analysis performed at the Molecular Cytogenetics core facility at the Sanger Institute for further analysis. The cells were split onto gelatinized plates (10cm, Corning) and expanded in serum-containing media or chemically defined media (standard 2i or alternative 2i) for at least three passages.

The three media are as follows:

1) Serum-containing media: Knockout DMEM (Gibco), 1X penicillin-streptomycin-glutamine (Gibco), 1X non-essential amino acids (Gibco), 100

U/ml recombinant human leukaemia inhibitory factor (Millipore), 15% foetal bovine serum (HyClone), 0.1 mM β-mercaptoethanol (Sigma).

2) Standard 2i media: N2B27 basal media (NDiff 227, StemCells), 100 U/ml recombinant human leukaemia inhibitory factor (Millipore), 1 μM PD0325901 (Stemgent), 3 μM CHIR99021 (Stemgent).

3) Alternative 2i media: N2B27 basal media (NDiff 227, StemCells), 100 U/ml recombinant human leukaemia inhibitory factor (Millipore), 1 μM CGP77675 (Sigma), 3 μM CHIR99021 (Stemgent).

Dr. Alex Tuck performed NPC differentiation time course using protocol published by Bibel et al., 2007 and harvested cells at day 6 and day 8. He prepared libraries for single cell mRNA sequencing using the protocol described in the section 2.2 and these samples were sequenced 150bp paired end on Illumina HiSeq2000. Mapping and downstream analysis was preformed as described in the section 2.5.

2.2 Single cell mRNA-seq using SmartSeq and Fluidigm C1

2.2.1 Single cell suspension preparation

Cells were harvested by trypsinisation (0.05% trypsin/EDTA, Gibco) for 10 minutes, when they reach 70-80% confluence for single cell capture. Subsequently they were inspected under the microscope to assure the cells are a single cell suspension, counted and diluted to 1.3×10^{-6} cells per millilitre.

2.2.2 cDNA synthesis and amplification

For each culture condition, 4000 cells were loaded on to a 10-17 micron Fluidigm C1 Single-Cell Auto Prep IFC, and cell capture was performed according to the manufacturer's instructions. The capture efficiency was determined using a microscope to exclude samples from the analysis with no or more than one cell captured or samples where in addition to cell there was cellular debris visible. Upon capture, reverse transcription and cDNA preamplification were performed in the 10-17 microns Fluidigm C1 Single-Cell Auto Prep IFC using the SMARTer PCR cDNA Synthesis kit (Clontech) and the Advantage 2 PCR kit (Ramskold et al., 2012).

Within the C1 cells are first lysed to release RNA using Triton-X 100 in the lysis buffer. Subsequently reverse transcription mix is added to perform reverse transcription. Importantly template-switching mechanism is used to avoid additional steps of adapter ligation and second strand synthesis.

The yield of the cDNA from a single cell is low, so it needs to be amplified before library preparation can be performed. During reverse transcription, adaptors are incorporated within the primers to allow amplification of full-length transcript by PCR. Reverse transcription is primed using a poly-T oligonucleotide, which allows selection of polyadenylated RNA species i.e. mRNA and some lncRNAs; this avoids sequencing abundant rRNAs. Full-length amplified cDNA was harvested, assessed and quantified using High Sensitivity DNA Kit (Agilent) and stored at -20°C.

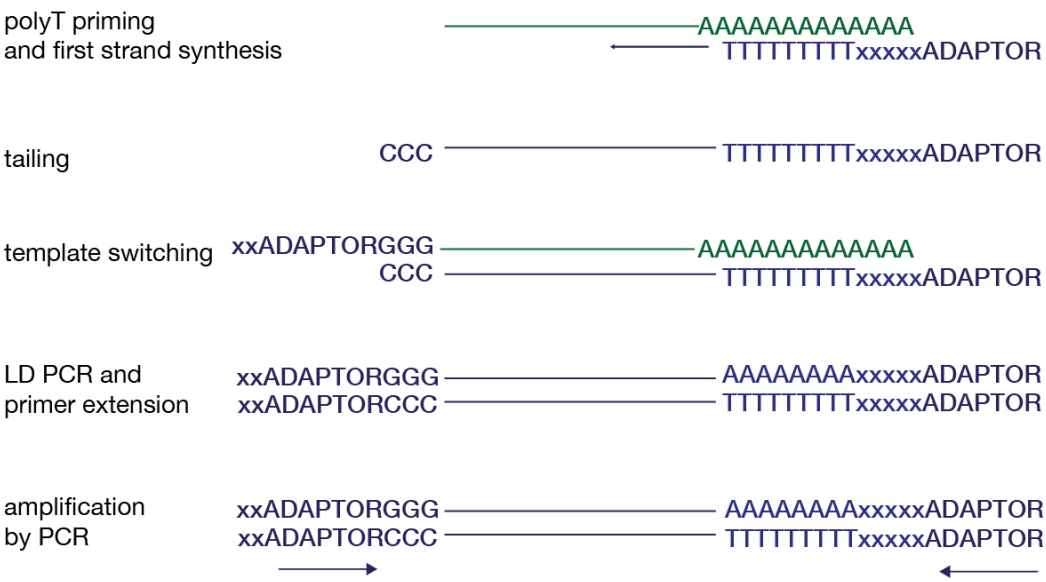


Figure 2.1 Schematic of cDNA synthesis and amplification
Polyadenylated RNAs are selected by reverse transcription with polyT primer; second strand is synthesized with template switching reaction. cDNA is amplified by PCR.

2.2.3 Illumina library preparation using Nextera XT

cDNA was diluted to a range of 0.1-0.3 ng/μl and Nextera libraries were prepared using the Nextera XT DNA Sample Preparation Kit and the Nextera Index Kit (Illumina) following the instructions in the Fluidigm manual "Using the C1™ Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing". Libraries from one chip were pooled, and paired-end 100bp sequencing was performed on 4 lanes of an Illumina HiSeq2000.

2.3 mRNA sequencing of bulk controls

Bulk mRNA sequencing libraries were prepared and sequenced using the Wellcome Trust Sanger Institute sample preparation pipeline with the TruSeq RNA Sample Preparation v2 kit (Illumina). RNA was extracted from 1-2 million cells using the Qiagen RNA Purification Kit on a QiaCube robot. The quality of the RNA sample was checked using gel electrophoresis. For library

preparation, poly-A RNA was purified from total RNA using oligo-dT magnetic pull-down. Subsequently, mRNA was fragmented using metal-ion catalysed hydrolysis. The cDNA was synthesized using random hexamer priming, and end repair was performed to obtain blunt ends. A-tailing was done to enable subsequent ligation of Illumina paired-end sequencing adapters, and samples were multiplexed at this stage. The resulting library was amplified using 10 cycles of PCR, substituting the Kapa Hifi polymerase for the polymerase in the Illumina TruSeq kit. Samples were diluted to 4nM, and 100bp paired end sequencing carried out on an Illumina HiSeq2000. The Sanger sequencing facility performed Sequencing Quality Control.

2.4 Candidate gene expression downregulation using CRISPR repressor

2.4.1 CRISPRi plasmids and cloning

Expression of candidate pluripotency regulators was downregulated with CRISPRi technology. I obtained three plasmids necessary for genome integration and expression of dCas9-KRAB and gRNA from Dr. Xuefei Gao. Two plasmids were used, one bearing gRNA linked to mCherry (Figure 2.2) and the second one dCas9-KRAB linked to BFP (Figure 2.3). Both expression cassettes are within LTR sites that are integrated into the genome using the hyperactive piggyBac transposase (Yusa et al., 2011) expressed from the third plasmid (Gao et al., 2014) (Figure 2.4).

Oligonucleotides targeting sites at promoters of candidate genes were ordered from Sigma-Aldrich (Table 2.1). I diluted the oligos to 1 mM in water and mixed them 1:1. I took 10 μl of oligo mix and heated it up to 98°C in the

thermo-cycler and then lowered the temperature by 1°C every minute until it reached 20°C to anneal the oligos and create sticky ends for the ligation to the backbone. pPB-gRNA-BsaI backbone was designed in a way that there are two BsaI cutting sites in the position where annealed oligos need to be ligated.

I performed restriction digestion of the plasmid using BsaI enzyme from New England Biolabs for 2h at 37°C. In 50 μl reaction I digested 2 μg of plasmid using 20U of the enzyme in 1x CutSmart buffer. Subsequently I ran a 2% agarose gel, cut the band corresponding to the double cut plasmid and purified the DNA using Qiagen Gel Extraction kit. Ligation was performed for each insert in the same way. 0.05 μg of plasmid was mixed with 5 μl of 5 mM annealing product, 1U of T4 DNA ligase from Thermo Fisher in 20 μl reaction containing 1x ligation buffer. Ligation was done for 1h at room temperature. 1 μl of ligation reaction was used for heat shock transformation of 25 μl of DH5α cells. Cells were plated on ampicillin for selection of successfully transformed cells and subsequently colonies were picked and grown in LB media and then I purified plasmids using MiniPrep kits from Qiagen. To check if ligation was successful I performed test digestions with BglII and XhoI (if successful 0.5k, 1.7kb and 3.9kb fragments were observed, if not: 0.9kb, 1.7kb and 3.9kb fragments) and subsequently sent plasmids for Sanger sequencing.

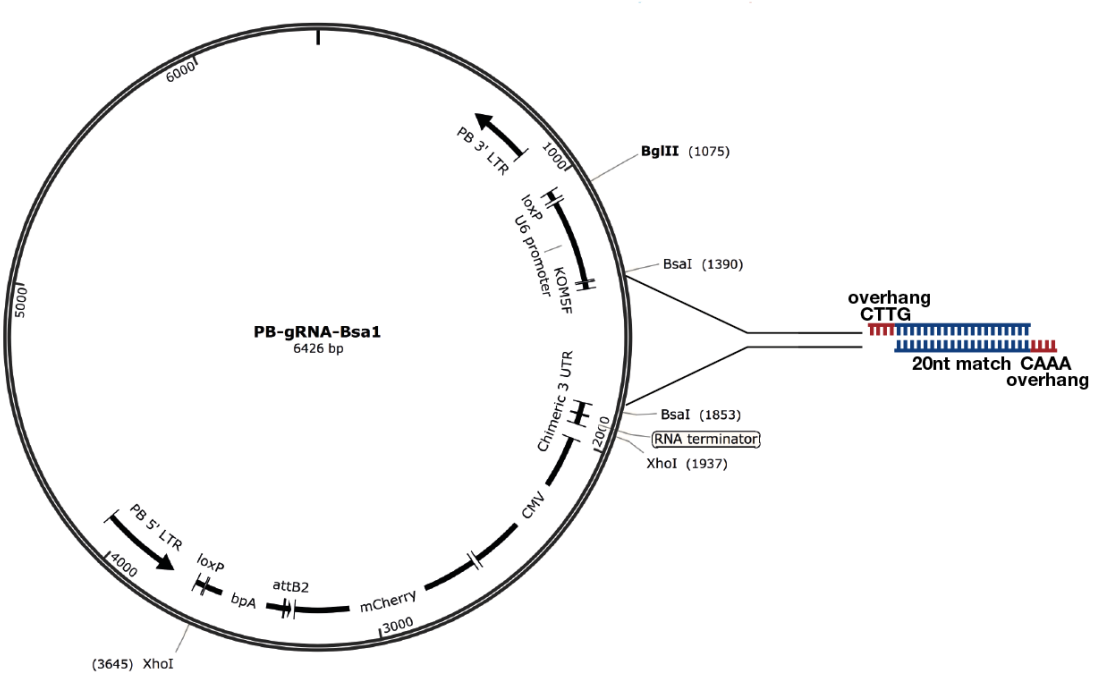


Figure 2.2 Schematic of gRNA plasmid

Column	Name	PRIMER 1	PRIMER 2
Ptma	Ptma-1	cttggcgccgcgtgagtcccccac	aaacgtgggggactcacgcggcgc
	Ptma-2	cttgcaatagcgccgggactaggg	aaacccctagtcccggcgctattg
	Ptma-3	cttgctgcgctcagccaatagcgc	aaacgcgctattggctgagcgcag
	Ptma-4	cttgttcggaatcgagccaatgag	aaacctcattggctcgattccgaa
	Ptma-5	cttggcgcagcgcgcgccaagccg	aaaccggcttggcgcgcgctgcgc
Set	Set-1	cttgtgctgattggagggagggcg	aaaccgccctccctccaatcagca
	Set-2	cttgtcaaagaagtttctgctgat	aaacatcagcagaaacttctttga
	Set-3	cttggccgccccctttctccatcgc	aaacgcgatggagaaggggggcggc
	Set-4	cttgcccggcgcgcctgcgctctg	aaaccagagcgcaggcgcgccggg
	Set-5	cttggccgggggcgggacttgcgc	aaacgcgcaagtcccgccccggc
	Set-6	cttgacggcgcgagcctctccggc	aaacgccggagaggctcgcgccgt
	Set-7	cttgggggagcaccgcgcgggggc	aaacgcccccgcgcggtgctcccc
Zfp710	Zfp710-1	cttgggagagcagggaagtgtggg	aaaccccacacttccctgctctcc
	Zfp710-2	cttggatgagaaggggtggagcca	aaactggctccacccccttctcatc
	Zfp710-3	cttgtgtgggaggaattgatgaga	aaactctcatcaattcctcccaca
	Zfp710-4	cttgccagggagagcagggaagtg	aaaccacttccctgctctccctgg
	Zfp710-5	cttgcctctgcgagcaggcttagg	aaaccctaagcctgctcgcagagg
	Zfp710-6	cttggaaaacaaaagagagataaa	aaactttatctctcttttgttttc
	Zfp710-7	cttgaagaagaaaaatcctctctg	aaaccagagaggatttttcttctt
	Zfp710-8	cttgtccaggcttgcaattcgagt	aaacactcgaattgcaagcctgga
Zfp640	Zfp640-1	cttgcaagatcactgtggctgtgc	aaacgcacagccacagtgatcttg
	Zfp640-2	cttggacaaagaggcgggatcttc	aaacgaagatcccgcctctttgtc
	Zfp640-3	cttgggaagcaaactttaacatta	aaactaatgttaaagtttgcttcc
	Zfp640-4	cttgactggccaatcaagttcgcc	aaacggcgaacttgattggccagt
Kat6b	Kat6b-1	cttggggctctgtgcgctgcagcc	aaacggctgcagcgcacagagccc
	Kat6b-2	cttgcctcccctgagggcggtgag	aaacctcaccgccctcaggggagg
	Kat6b-3	cttgcgggtgacggacagacccgt	aaacacgggtctgtccgtcacccg
	Kat6b-4	cttgggcatccccgccctcccctg	aaaccaggggagggcggggatgcc
Etv5	Etv5-1	cttgccggaggccggcgcgcagag	aaacctctgcgcgccggcctccgg
	Etv5-2	cttggacgtgtgtgctctgggctg	aaaccagcccagagcacacacgtc
	Etv5-3	cttgcggggatggccgccgaccaa	aaacttggtcggcggccatccccg
	Etv5-4	cttgcaagaggtgatgggcagccg	aaaccggctgcccatcacctcttg
	Etv5-5	cttgaaggtggctacacaggcaag	aaaccttgcctgtgtagccaccttt
	Etv5-6	cttgtttttcagtgcaagtaaggg	aaaccccttacttgcactgaaaaa
	Etv5-7	cttgggcttttgtggtagacaggc	aaacgcctgtctaccacaaaagcc
	Etv5-8	cttgttggttggttttggcttttg	aaaccaaaagccaaaaccaaccaa
Dpy30	Dpy30-1	cttggtctgctgcccgcgggggtg	aaaccacccccgcgggcagcagac
	Dpy30-2	cttgcgacgaggacggccagtcgg	aaacccgactggccgtcctcgtcg
	Dpy30-3	cttgccgagcctcgcgatgcgacg	aaaccgtcgcatcgcgaggctcgg
	Dpy30-4	cttgtcctcccaccgctacatcct	aaacaggatgtagcggtgggagga
	Dpy30-5	cttgatttgcctcaagtctgtaaa	aaactttacagacttgaggcaaat
	Dpy30-6	cttgatacatacttcttgaacaat	aaacattgttcaagaagtatgtat

Table 2.1 Sequences of oligonucleotides used to construct insert gRNA plasmid

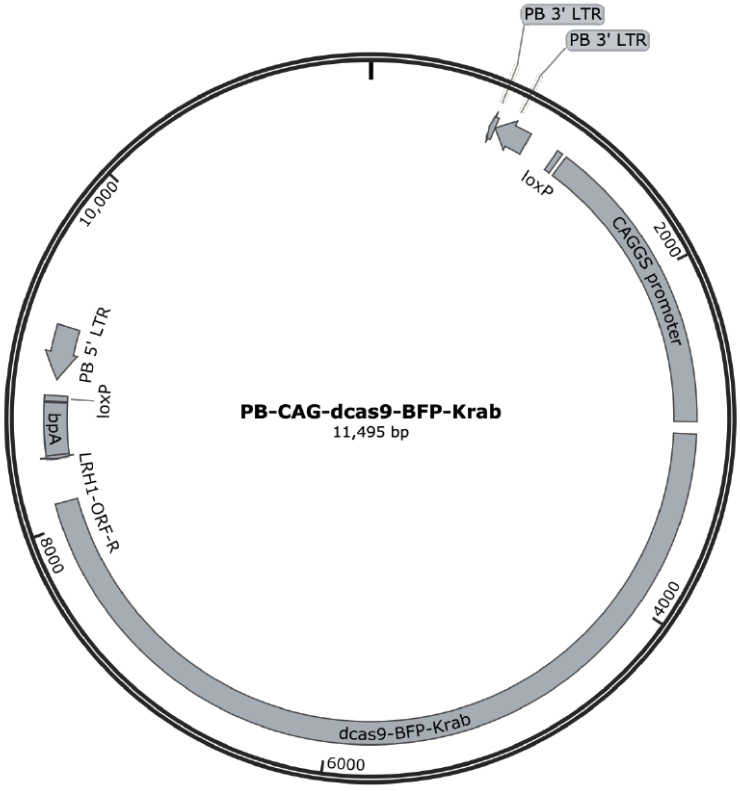


Figure 2.3 Schematic of dCas9-Krab plasmid

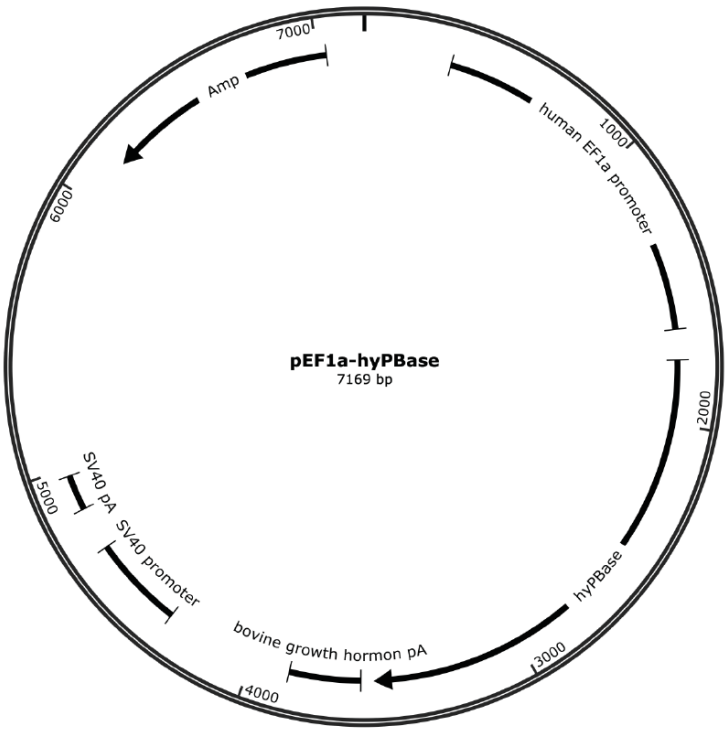


Figure 2.4 Schematic of piggyBac plasmid

2.4.2 Downregulation of target gene expression and cell sorting

GFP-OCT4 reporter strain ES cells (Silva et al., 2008) were grown by Dr. Xuefei Gao in 6-well plates and transfected with plasmids (1) 5 μg of dCas9-KRAB-BFP plasmid (2) 1 μg of hyPBase plasmid and (3) 1 μg cocktail of gRNA plasmids (Gao et al., 2014) targeting the gene of interest in a 1:1 ratio using Lipofectamine2000, Life Technologies. Subsequently, cells were cultured in knockout DMEM (Gibco) medium containing 15% serum and 100 U/ml LIF for 4 days. After 4 days cells were harvested from the culture dish using trypsin (0.05% trypsin/EDTA, Gibco) and the Cytometry Core Facility at Sanger Institute sorted BFP and mCherry double positive cells.

2.4.3 Library preparation

RNA was extracted using Qiagen RNeasy Mini kit from 10,000 mCherry and BFP positive cells that were sorted for each sample in triplicates. Modified SmartSeq2 protocol was used for reverse transcription and amplification of cDNA (Picelli et al., 2014), because the amount of RNA from 10,000 cells is not sufficient for conventional bulk library preparation protocols, which involve polyA species enrichment where a substantial amount of material is lost. Sequencing libraries were prepared using Nextera XT kit according to manufacturers guidelines, barcoded with Nextera XT Dual Index kit and sequenced on an Illumina HiSeq2500 in rapid mode.

2.5 Data analysis

2.5.1 Sequencing reads alignment

For each cell, 100bp paired-end reads were aligned to the *Mus musculus* genome (GRCm38) using GSNAP (version gmap-2014-05-15_v2) with default options (Wu and Nacu, 2010). To detect splice junctions in reads, I used a set of known splice sites from the GTF file for GRCm38 provided by Ensembl (release 73). Only reads uniquely mapped to the genome were counted for each gene using htseq-count and the same GTF file (Anders et al., 2014).

Dr. Jong Kyoung Kim additionally applied location and scale adjustments to the normalized read counts to remove technical variation among multiple batches as described below.

2.5.2 Normalisation and batch correction

To remove technical variation across multiple batches, Dr. Jong Kyoung Kim applied location and scale adjustments to the normalized read counts by using the ComBat function of the sva package of R with default options (Johnson et al., 2007). He first $\log_{10}$-transformed the normalized read counts (after removing lowly expressed genes whose mean normalised read counts are less than 10) and after adding a pseudo count of 1. Secondly, he adjusted for batch effects using ComBat with the known batch covariate and sample conditions. Finally, he re-transformed the batch-adjusted expression values x back to the original scale (10^{x-1}). If the re-transformed values were less than 0 or the original read counts are 0, we set the re-transformed values to 0.

2.5.3 Quality control of cells

To exclude poor quality libraries from downstream analysis, first I removed cells that correspond to empty capture sites, capture sites with multiple cells, or capture sites containing cell debris on the C1 chip by visually inspecting them under microscope. Second, it has been known that some cells suffer from cell rupture during the process of microfluidic cell capture (Islam et al., 2014). To identify these abnormal cells, I calculated two quantities for each cell: the number of reads mapped to exons, and the proportion of reads (of all reads mapped to exons) mapped to 37 genes on the mitochondrial chromosome. I identified two populations of cells in terms of the above two quantities and most of the cells corresponding to empty cells or cell debris are in one of the two populations. Biologically when cell is ruptured cytoplasm leaks out and there is a relative increase in abundance of transcripts that are enclosed within the mitochondria. Based on this, I set the following criteria to remove abnormal cells:

1) Cells that have fewer than 500,000 reads mapped to exons

2) Cells that have greater than 10% reads mapped to mitochondrial genes

Finally, I compared the normalised read counts of genes between cells in the same condition, and found that in one cell (cell "85" in the first replicate of serum) there was a problem in library preparation and many genes were abnormally amplified (Figure 2.5). I removed the cell from further analysis. In summary, I have the following number of cells for the analysis: 81, 90 and 79 for serum replicates; 82, 59, 72 and 82 for 2i replicates, 93 and 66 for a2i replicates, where the total number of cells across conditions is 704.

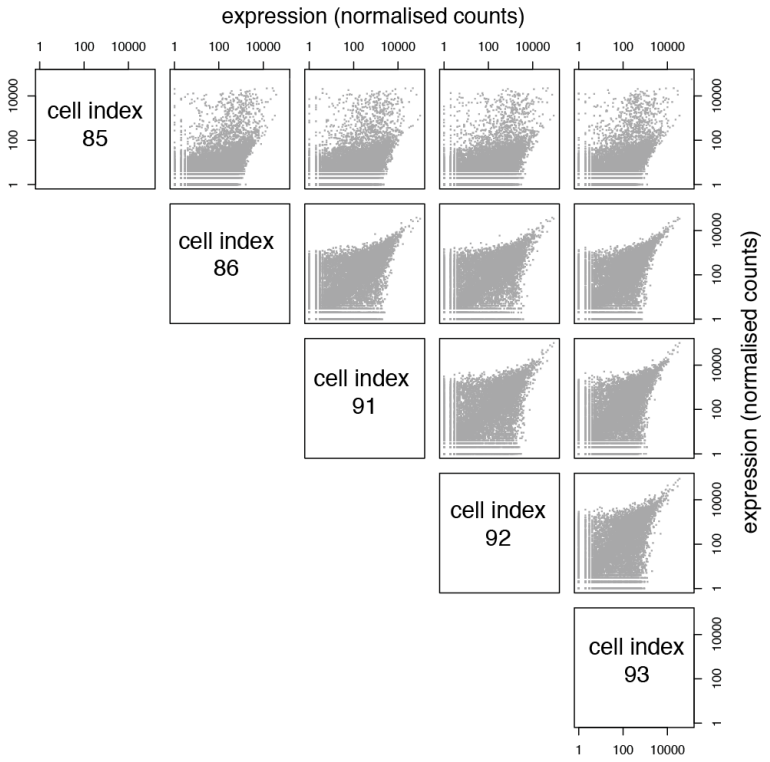


Figure 2.5 Correlation of gene expression levels in single cells.
Expression levels correlate with each other as shown representatively for cells index 86, 91, 92 and 93 (serum 1). Cell 85, as it is substantially different than any other cell, suggesting failure of the experimental protocol.

2.5.4 Calculating DM as a measure of noise

To account for the confounding effects of gene length and mean expression level on the CV, Dr. Jong Kyoung Kim computed the DM values for each gene using rolling medians of the squared CV. First, he computed gene lengths by taking the union of all exons within a gene based on the Ensembl annotation. He excluded all exons annotated as "retained_intron". He also removed lowly expressed genes whose mean normalised read counts (reads per million) are less than 10, since we cannot distinguish biological noise from technical noise for these genes. Second, he computed rolling medians from the scatter plot between the mean normalised read counts and the squared CV values, where

the x- and y-axis are log10 transformed. Third, we calculated the mean-corrected residual of the squared CV of gene i to its corresponding rolling median $f(i)$ such that

$$r(i) = \log_{10} CV(i)^2 - f(i).$$

Finally, to correct for the effect of gene length on the mean corrected residual, he calculated the difference between the mean corrected residual of the squared CV of gene i and its expected residual by using the following formula

$$DM(i) = r(i) - g(i),$$

where $g(i)$ is the rolling median of gene i from the scatter plot between $r(i)$ and $\log_{10}$ transformed gene lengths. To compute the rolling medians, he used the rollapply function of the zoo package of R (Zeileis and Grothendieck, 2005) and the following parameters: the number of genes in the window is 50 and the number of overlapping genes between adjacent windows is 25. This relative noise measure, which is referred to as DM, does not depend on either gene expression levels or gene lengths (Spearman's ϱ=0.0200 for gene expression levels and ϱ=0.0206 for gene length in the serum condition) (Kolodziejczyk et al., 2015b).

2.5.5 Testing the absolute level of cell-to-cell variation of a functional category within a culture condition

To test whether genes belonging to a defined functional category have a high or low level of expression heterogeneity within a culture condition, Dr. Jong Kyoung Kim performed gene set enrichment analysis using the Piano package of Bioconductor (Varemo et al., 2013). He used the DM values for gene-level statistics and calculated the mean DM values as a gene-set statistic

for each GO term. The associations between Ensembl gene IDs and GO terms were obtained from the biomaRt package of Bioconductor (Kasprzyk, 2011). Since gene set enrichment analysis tends to bias towards large or small categories in terms of their number of genes, he considered only gene sets with between 3 and 2,000 genes. The P-value for each GO term was then computed by randomly taking a set of genes of the same size as in the GO term, and by repeating this 10,000 times.

2.5.6 Testing the relative difference in expression heterogeneity of a functional category across culture conditions

To explore further the difference of the three culture conditions in terms of gene expression noise, Dr. Jong Kyoung Kim compared two sets of DM values for each GO term between two culture conditions using the two-sided paired t-test. He only considered GO terms with at least 2 genes having DM values. The associations between GO terms and their offspring terms were obtained from the GO.db annotation package of Bioconductor

(http://www.bioconductor.org/packages/release/data/annotation/html/GO.db.html).

2.5.7 Differential expression analysis

I identified differentially expressed genes from bulk data and single cell data using the DESeq package (Anders and Huber, 2010). I considered genes that differed in expression by two-fold and with a multiple testing adjusted *p*-value was < 0.05 to be differentially expressed. For single cell differential expression analysis I used each as a replicate of the condition it came from and I removed genes that had mean expression below 50 counts.

2.6 Doubling time estimation of mouse embryonic stem cells in different conditions

Fifty thousands G4 mouse ES cells were plated by Dr. Jason Tsang in single wells on gelatinized 6-well plates, and maintained in the three culture conditions of interest (total 12 wells for each culture condition): serum-containing media, standard 2i media and alternative 2i media. Three wells were harvested and quantified on a haemocytometer every 24 hours for 4 days to estimate the doubling time of mouse ES cells in each condition.

2.7 Datasets

	Generated by:	Data accession numbers
Single cell mRNA seq data of mESC cultured in three conditions (2i, a2i, serum)	Kolodziejczyk et al., 2015, Cell Stem Cell	Array Express E-MTAB-2600
Single cell mRNA seq data of early mouse embryo development	Deng et al., 2014, Science	Gene Expression Omnibus GSE45719
2C-like cell gene expression profiles (microarray data)	Macfarlan et al., 2012 Nature	DE count tables from Supplementary Table 4
NPC differentiation time course	Dr. Alex Tuck (unpublished)	unpublished

Chapter 3

Cell-to-cell gene expression variation associated with mESC culture conditions.

3.1 Introduction

Despite their shared hallmarks of biological origin, mouse embryonic stem cells propagated in different *in vitro* environments are morphologically distinct and possess characteristic transcriptional and epigenetic profiles (Ficz et al., 2013; Marks et al., 2012). Depending on how the pluripotency of mESCs is maintained in culture, they exhibit different characteristics. Cells cultured in serum/LIF are flattened, grow in a monolayer and are well-attached to the surface, while cells in 2i/LIF and a2i/LIF form compact three-dimensional colonies and tend to attach to each other more than to the surface. Furthermore, serum/LIF maintained mESCs are morphologically more heterogeneous (Marks et al., 2012; Shimizu et al., 2012; Ying et al., 2008).

It was shown using bulk RNA sequencing that transcriptomes of cells cultured in 2i and serum differ. Several developmental, metabolic and cell cycle related genes are differentially expressed between conditions, further illustrating the importance of cell culture condition in determining phenotype (Marks et al., 2012). The reason for the distinct transcriptomes may lie in different epigenomes of these cells (Angermueller et al., 2016; Ficz et al., 2013; Smallwood et al., 2014). Cells grown in 2i/LIF are globally hypomethylated in comparison to cells grown in serum/LIF (Habibi et al., 2013), and also they exhibit different histone modification patterns (Marks et al., 2012).

The morphological heterogeneity of cells grown in serum/LIF led to attempts to understand this property of the population. Certain pluripotency factors such as *Nanog* (Chambers et al., 2007; Kalmar et al., 2009), *Dppa3* (Hayashi et al., 2008) and *Rex1* (Zfp42) (Toyooka et al., 2008) exhibit transcriptional fluctuations, meaning that within the population there is a group of cells that express these genes at a low level and another subpopulation that expresses them highly. Cells that express low levels of *Nanog* can change their expression to high and vice versa, and these populations remain in a dynamic equilibrium (Kalmar et al., 2009). It was shown that cells that express low levels of NANOG are less pluripotent, and this led to the hypothesis that this population represents the differentiation-poised states and is instrumental in regulating exit from pluripotency (Chang et al., 2008).

Importantly, others have expressed concern that the phenomenon of fluctuations may originate from the use of fluorescent reporter systems (Chang et al., 2008; Faddah et al., 2013; Reynolds et al., 2012). It was suggested that *Nanog* is randomly monoallelically expressed i.e. cells stochastically switch off

one of the alleles (Miyanari and Torres-Padilla, 2012). In cases when one of the alleles of Nanog is fused to fluorescent reporter protein, the population of cells will divide into two subgroups, cells with low levels of fluorescence, where the fluorescent reporter protein tagged allele is switched off, and the second population with high fluorescence from the active reporter allele. It is worth noting that some groups have shown that *Nanog* is expressed from both alleles (Faddah et al., 2013; Filipczyk et al., 2013) and this points to the conclusion that fluctuations are not an artefact of reporter system, but a biological phenomenon.

The presence of transcriptionally heterogeneous subpopulations, prevalent bivalent chromatin domains, increased methylation content and reduced RNA polymerase pausing in serum compared to 2i mESCs has led to the notion that serum-maintained mESCs exist in a metastable pluripotent state (Marks et al., 2012), implying a higher transcriptional cell-to-cell variation compared to the uniform ground states exhibited by the chemically defined "2i" conditions (Klein et al., 2015; Kumar et al., 2014).

In this chapter I aimed to characterize in detail heterogeneity of mouse embryonic stem cells in different culture conditions by quantification of gene expression variability and comparison between three culture conditions: serum/LIF, 2i/LIF and alternative 2i/LIF (Shimizu et al., 2012; Ying et al., 2008). Subsequently, I set out to understand the biological context of the observed variability. In more detail, the questions that I wanted to address involve understanding heterogeneous Nanog expression at the mRNA level and surveying if there are other genes that exhibit such variability. Furthermore, I wanted to identify transcriptionally similar subpopulations of cells in serum and to investigate whether Nanog-high cells from serum are

similar to 2i-cultured cells. I then aimed to compare the whole transcriptome heterogeneity between conditions to find whether it is higher in serum in comparison to 2i and to find genes that contribute to this heterogeneity. Finally, I wanted to analyse if culturing cells in the alternative 2i media leads to similar transcriptomes to 2i, as is suggested by their similar morphologies (Shimizu et al., 2012). I used single cell RNA sequencing to overcome limitations of previous transcriptomic analyses and to provide a high-resolution analysis of cellular heterogeneity.

3.2 Experimental design

To examine gene expression variability and understand how serum-grown mESCs differ from those grown in 2i media, an F1 hybrid (C57BL/6Ncr male x 129S6/SvEvTac female) male mESC cell line (George et al., 2007) was cultured in three different conditions: (1) three replicates of serum + LIF, (2) four replicates of 2i + LIF, and (3) two replicates of "alternative 2i" + LIF, which are henceforth referred to as serum (serum1, serum2, serum3), 2i (2i1, 2i2, 2i3, 2i4) and a2i (a2i1, a2i2) (Figure 3.1). I characterized cells in these three conditions by single cell RNA-sequencing using the Fluidigm C1 system. The cDNA from each 96-cell chip was sequenced on four lanes of a HiSeq2000. Reads were aligned to the *Mus musculus* genome (GRCm38) using GSNAP and subsequently reads mapped to each gene were counted using HT-Seq.

culture condition	serum	2i	alternative 2i (a2i)
components of medium	DMEM 15% fetal bovine serum + leukemia inhibitory factor (LIF)	N2B27 basal media inhibitors of: GSK3β (CHIR99021) Mek1/2 (PD0325901) + LIF	N2B27 basal media inhibitors of: GSK3β (CHIR99021) Src (CGP77675) + LIF
cell characteristics	more differentiation permissive more heterogeneous	ground pluripotent state more homogeneous	not well characterised
references	Pease et al., 1990 Xu et al., 2001	Ying et al., 2008 Li et al., 2008	Shimizu et al., 2012
number of cells	chip 1 - 81 cells chip 2 - 90 cells chip 3 - 79 cells	chip 1 - 82 cells chip 2 - 59 cells chip 3 - 72 cells chip 4 - 82 cells	chip 1 - 93 cells chip 2 - 66 cells

Figure 3.1 Experimental schematic of hybrid mESCs in three culture conditions. Table of experimental setup and cell culture conditions used in our study.

3.3 Quality control

Single cell mRNA sequencing experiments work with fragile cells and very small amounts of material. Thus it is essential to perform quality control to remove from analysis samples containing broken or dead cells as well as those exhibiting technical problems, such as pipetting errors or poor quality of sequencing library preparation (Ilicic et al., 2016).

Three criteria were used to remove poor quality cells. First, I excluded samples that upon microscopic inspection (20x light microscope), appeared empty, contained double or multiple cells or showed some debris within capture sites of the C1 chip. Second, samples with fewer than 500,000 reads mapped to exons were discarded. Low numbers of reads mapping to the transcriptome may suggest contamination or failure in one of the steps of the protocol: cell lysis, reverse transcription, cDNA amplification or library preparation. Third, I removed cells where more than 10% of reads mapped to the genes encoded by the mitochondrial genome. A high percentage of reads mapping to the mitochondrial genome is a good indication of low quality cells. One possible explanation is that when the cell is broken, cytoplasm leaks out

during washing steps, but membrane enclosed parts of the cell such as mitochondria and their contents remain intact. This leads to an apparent enrichment of transcripts from the mitochondrial genome, as they are enclosed within mitochondria and are not washed out (Figure 3.2).

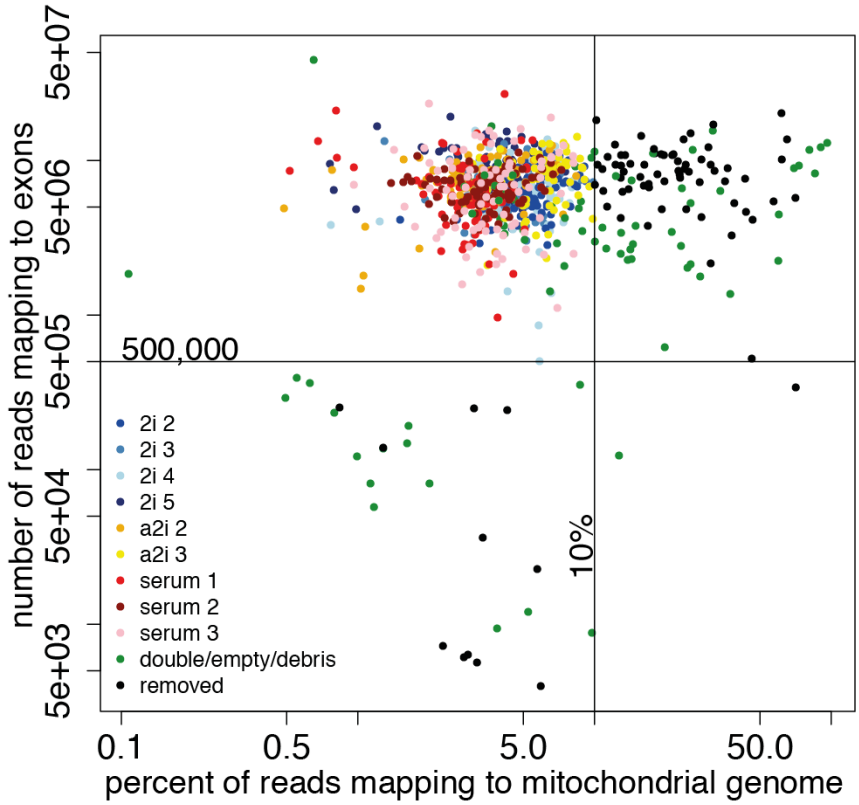


Figure 3.2 Quality control of cells
Quality control metrics were the number of reads mapping to exons (y axis), and the proportion of reads mapped to mitochondrial genes (x axis). Lines represent the thresholds used. Green points represent cells excluded upon microscopic examination of the C1 chip and black points represent cells that did not pass the thresholds.

After removing poor quality cells (18.5% of all cells), 295 2i cells, 159 a2i cells and 250 serum cells remained. On average, I sequenced over 9 million reads per cell. Over 80% of reads mapped to the *Mus musculus* genome and over 60% to exons (Figure 3.3). I also performed standard bulk RNA sequencing using at least a million cells per sample for each condition to compare to single cell sequencing data of the same samples. Bulk data were

obtained from the same cell culture as 2i 1, serum 1, serum 2 and a2i 1, thus the only difference between single cell experiment and respective bulk are technical.

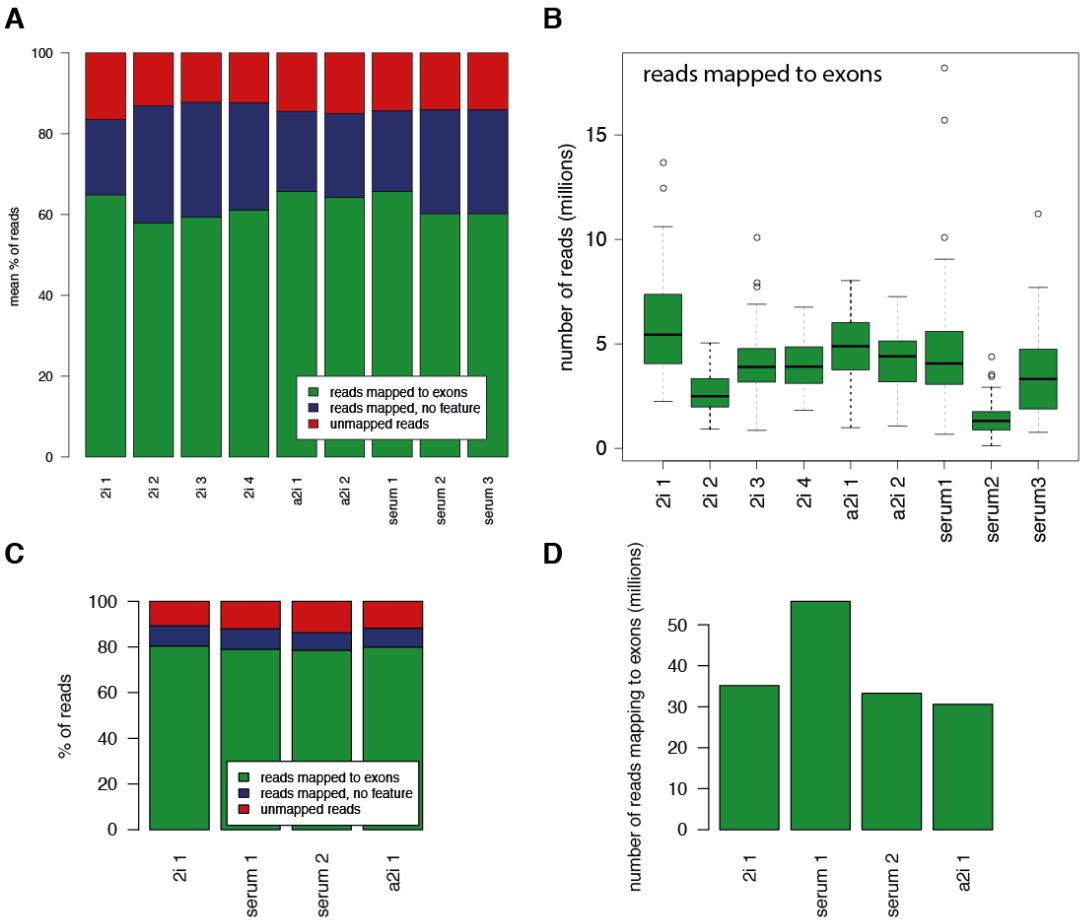


Figure 3.3 Mapping statistics for each cell for each sample.
The mean percentage of reads mapping to Ensembl exons (green), to the genome outside Ensembl annotated regions (blue) and unmapped reads (red) for each of nine experiments. (A) and (B) show results for single cell experiments while (C) and (D) for accompanying bulk.

To assess if the single cell RNA-seq data was in agreement with the results from bulk experiments, I averaged gene expression levels across the single cells profiled in each condition and compared with bulk RNA sequencing of cells from the same culture. I observed that the mean expression levels of all genes recapitulated the bulk gene expression levels with a Spearman rank

correlation coefficient of 0.88 for 2i, 0.89 for a2i, 0.91 for serum 1 and 0.90 for serum 2, and all *p*-values are smaller than 10^{-15} (Figure 3.4). It is worth noting that for lowly expressed genes there is less correspondence, as these genes are not detected in all single cells, due to lower sensitivity of single cell methods and technical noise.

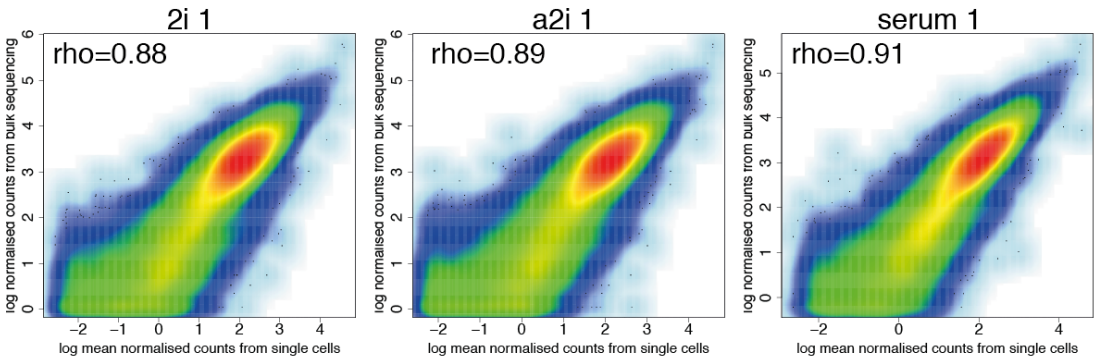


Figure 3.4 Comparison of gene expression levels between bulk and single cells.
2D kernel density estimation of scatter plot between expression level in bulk experiment and mean of gene expression from single cells in each condition. Value of Spearman rank correlation coefficient (rho) between bulk and mean of single cells is indicated in the top left corner.

3.4 Variability of gene expression

An advantage of the single cell approach is that I can investigate gene expression in more detail by focusing not only on mean expression values, but also by studying the distribution of expression levels across the population, capturing cell-to-cell variability in gene expression (Grun and van Oudenaarden, 2015).

It was shown previously that some genes have higher heterogeneity than others in cells cultured in serum (Canham et al., 2010; Kalmar et al., 2009; Kumar et al., 2014). For example Roeder and Radtke (2009) showed that protein levels of OCT4 are relatively more homogeneous within a culture in

comparison to levels of NANOG (Roeder and Radtke, 2009). This prompted me to see how this compares to the mRNA expression of these genes. Indeed I observed that *Nanog* is more heterogeneously expressed than *Oct4* (Figure 3.5). Coefficient of variation of gene expression for *Nanog* is 0.75 while for *Oct4* it is 0.68.

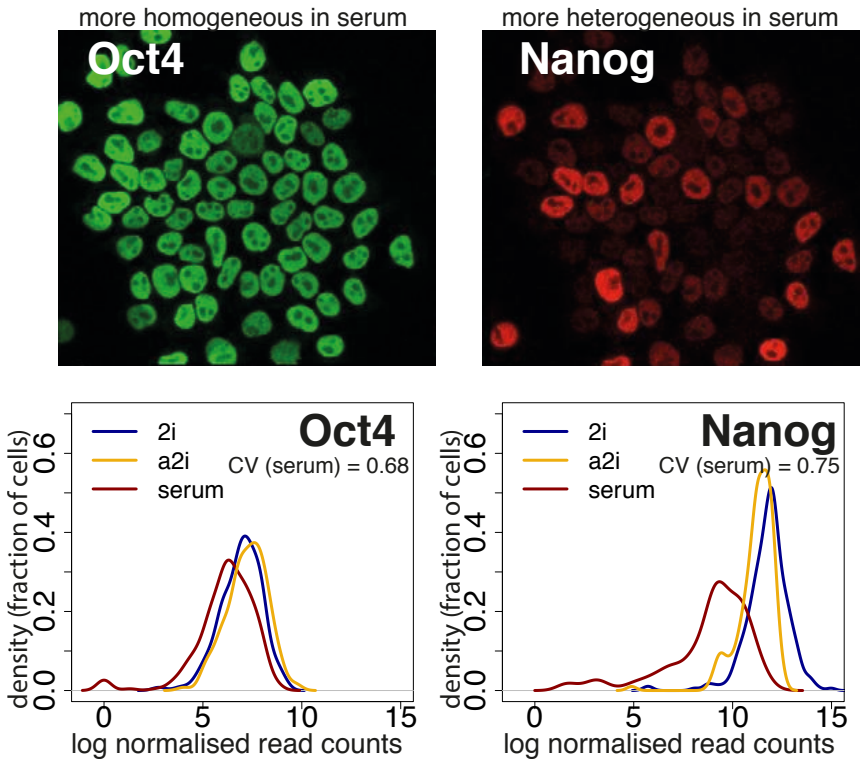


Figure 3.5 Variability of expression of Oct4 and Nanog.
Microscopy pictures showing fluorescently labelled Oct4 and Nanog are from Roeder and Radtke, 2009, and plots below show expression heterogeneity of Oct4 and Nanog in three culture conditions plotted using single cell mRNAseq data.

Subsequently I investigated if there was a difference in heterogeneity depending on the culture condition that the cells originated from. Upon inspection of gene expression distributions of several genes it was striking to me that some genes like *Tcerg1* do not have significantly different expression profiles between culture conditions (the two-sided Kolmogorov–Smirnov test

(KS test) p-value for 2i and a2i comparison is 0.82, and for 2i and serum 0.16). By contrast, some genes are more heterogeneous in one of the conditions, such as *Ccnb1*, which is more heterogeneous in 2i ($P=7\times10^{-4}$ by two-sided sided KS test between 2i and serum). Other genes, such as *Nanog*, *Klf4* or *Nr0b1*, are more heterogeneous in serum ($P<10^{-15}$ by the two-sided KS test between 2i and serum for genes mentioned above) (Figure 3.6). The null hypothesis of the KS test is that data in both samples are from the population with identical distribution. It compares cumulative distributions of two samples testing for different median, different variance or different distribution without making assumptions about the type of the distribution. Low *p*-value suggests that data were sampled from two populations, which have different distributions.

Many pluripotency associated genes are heterogeneous in serum, but in 2i. There is exception to this pattern. More specifically, Utf1 is a pluripotency factor implicated in regulation of bivalent genes (Jia et al., 2012), which is more heterogeneously expressed in 2i than in a2i and serum.

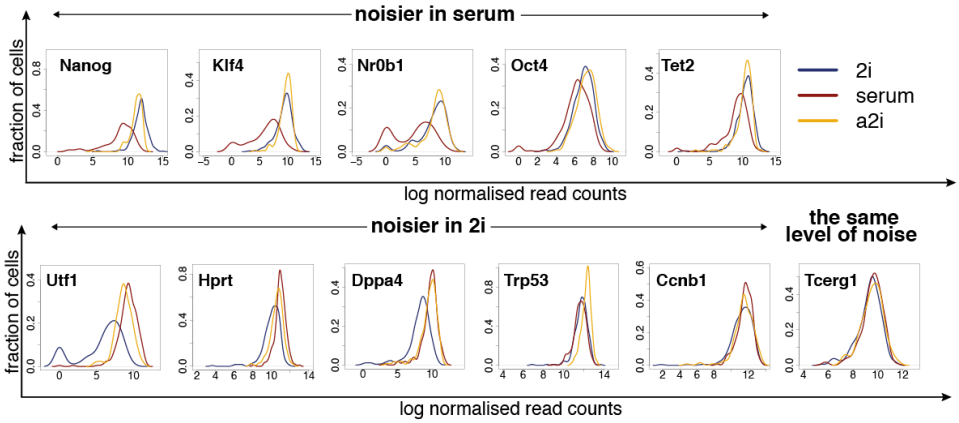


Figure 3.6 Gene expression distributions across cells

Gene expression distributions of genes, which are noisier in 2i than serum, which are noisier in serum than 2i and that have similar noise profiles in serum (red), 2i (blue), a2i (yellow). Distributions of gene expression were smoothed using the kernel density estimation function in R with default parameters

3.5 Transcriptome-wide gene expression variability measurement

Comparison of gene-expression variation was performed previously for selected genes using single molecule RNA-FISH and at the protein level with FACS with a few genes at a time (Raj et al., 2008). The strength of single cell RNA sequencing is that it allows us to investigate variability of all moderately and highly expressed genes at the same time from one population of cells.

To compare the global levels of gene expression heterogeneity between the three different culture conditions we did not use coefficient of variation (CV) of the normalized read counts, because the CV of a gene depends strongly on its mean expression level and length of the gene, which makes it difficult to interpret the noise difference of a gene between conditions. In collaboration with Dr. Jong Kyoung Kim, to account for the confounding factor of expression level, we used the distance between the squared CV of each gene and a running median as a measure of cell-to-cell variation. This is derived from the scatter plot of the mean normalized read counts versus the squared CV values, as in (Newman et al., 2006). We refer to this expression-level normalized measure of noise as distance to the median (DM). To calculate DM genes are divided into three groups depending on their length, because longer genes tend to have higher CV^2 in comparison to short genes. Subsequently for each of these groups rolling median of CV^2 depending on gene expression is calculated. And finally for each gene the median CV^2 for the expression bin this gene falls in is subtracted from the CV^2 of this gene (Please refer to Chapter 2 for details).

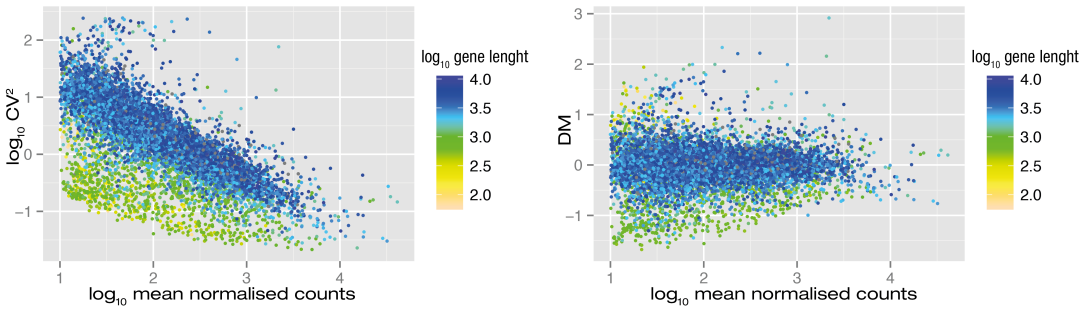


Figure 3.7 Gene expression variability measured with coefficient of variation (CV) and distance to the median (DM)

Plots show that there is a linear relationship between CV^2 and the level of gene expression, while this bias is not present for DM. Colours of dots indicate length of each gene.

Using DM, transcriptome-wide cell-to-cell variation is similar across the three culture conditions and I found that transcriptome-wide DM values are not significantly different across the three culture conditions (P=0.6252 by the Friedman rank sum test) (Figure 3.8). To compare three culture conditions at the same time we had to use the Friedman rank sum test, which is a nonparametric version of ANOVA. It is used to find different samples within 3 or more groups when data points are paired.

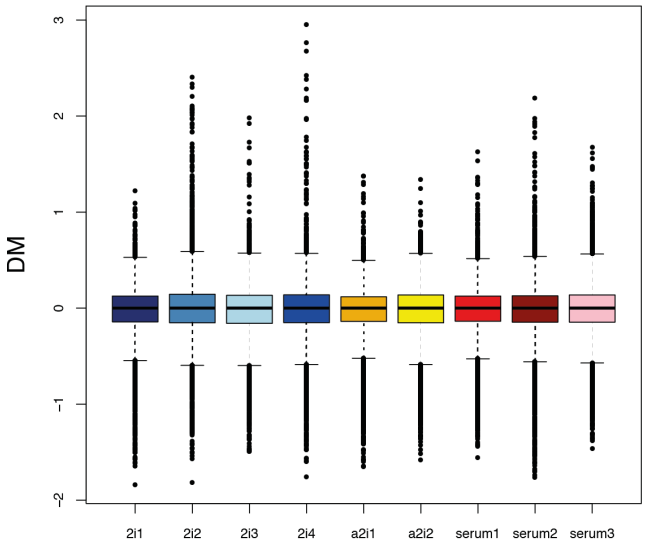


Figure 3.8 Gene expression variability across cells in different conditions measured with DM

Comparison of global gene expression variability by showing DM distribution of all expressed genes in all conditions, not including 2C-like cells.

Cells cultured in serum are more morphologically heterogeneous than cells cultured in 2i (Marks et al., 2012; Toyooka et al., 2008) and exhibit more variable expression of pluripotency factors, such as *Nanog* and *Zfp42* (Canham et al., 2010; Hayashi et al., 2008; Kalmar et al., 2009; Martinez Arias and Brickman, 2011; Singh et al., 2007). Hence, I expected that global gene expression variability would be higher in cells grown in serum compared with 2i. There were no reports on heterogeneity in a2i, but as morphologically a2i is similar to 2i, I anticipated that they would also be transcriptomically similar due to morphological similarities between these cells and those grown in 2i.

I observed that expression of pluripotency genes such as *Nanog* or *Nr0b1* is more heterogeneous in serum than in 2i or a2i. If these genes were to be more heterogeneous in serum, other genes might be more heterogeneous in 2i and a2i. These heterogeneous genes in 2i and a2i would balance heterogeneously

expressed pluripotency genes in serum leading to similar global heterogeneity. This prompted us to ask whether the gene expression heterogeneity levels of genes belonging to individual functional categories are the same or different between conditions.

To explore the relative difference in gene expression heterogeneity levels for each functional category between the culture conditions, we first compared the DM values of genes in pairs of culture conditions for each Gene Ontology (GO) term (excluding 2i replicates containing 2C-like cells; for discussion of 2C-like cells see chapter 4). We used paired t-test for comparison of DM between GO categories to show that a GO category and its child terms have more noise consistently in one condition compared to another. We did not perform an adjustment of the *p*-values for several reasons. The conventional FDR/FWER adjustment procedures can give very conservative *p*-values in this case, which means that the power of detecting GO categories showing true noise differences between two conditions will be too low. Additionally, we were interested in the consistent noise differences of a GO category and its child terms. In this case, the tests for GO categories are not independent and the multiple testing methods cannot be applied directly.

We found that 712 GO terms (out of a total of 19,107 terms) exhibit a significant difference in their noise levels in at least one pairwise comparison ($P<0.01$). For example, the expression of genes involved in "organ development" ($P=3.3\times10^{-4}$) and "cell adhesion" ($P=4.8\times10^{-4}$) are noisier in serum than in the inhibitory conditions (2i and a2i). These terms contain many of the pluripotency factors that were observed to display noisy expression patterns (Figure 3.9).

In contrast, genes involved in "cell cycle" ($P=5.4\times10^{-3}$) and "nuclear division" ($P=5.9\times10^{-6}$) have higher levels of noise in 2i compared to serum. When we included 2i replicates containing 2C-like cells, the conclusions are still valid, but marginally significant ($P<0.1$), possibly due to the presence of 2C-like cells (2C-like cells identification and characterization is described in chapter 5).

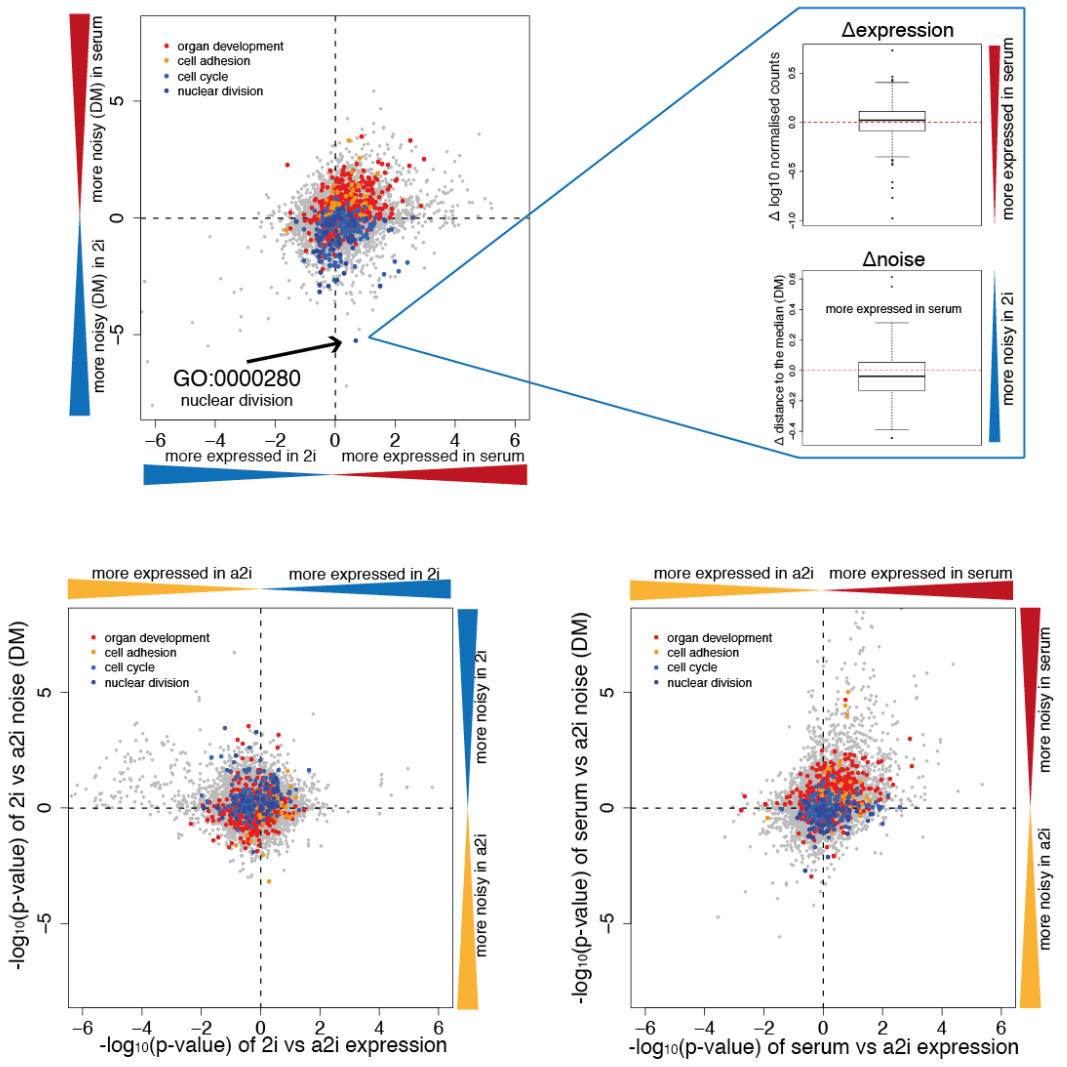


Figure 3.9 Gene expression heterogeneity of functional categories of genes
Comparison of the levels of gene expression and noise for gene ontology (GO) categories between the culture conditions (excluding 2i replicates containing 2C-like cells). The logarithm (log10) of P-values from two-sided paired t-test applied to mean

normalized read count (x-axis) and DM (y-axis) was computed for each GO category and plotted against each other by multiplying the sign of the t-statistic. Boxplots show an example of a GO category (GO:0000280, nuclear division) that is noisier in 2i and is similarly expressed between the two conditions.

3.6 Subpopulations of differentiating cells in serum

Fluctuations of gene or protein expression in serum were reported previously for some of the genes such as *Nanog* (Faddah et al., 2013; Kalmar et al., 2009; MacArthur et al., 2012; Singh et al., 2007), *Esrrb* (van den Berg et al., 2008) and *Zfp42* (Toyooka et al., 2008). Our data recapitulate these observations. Moreover, I found new genes to be noisy, such as *Nr0b1* or *Tet2* (Figure 3.6).

Genes that show noisy expression, especially those with obvious bimodal expression patterns like *Nanog*, *Klf4* or *Nr0b1*, may indicate the existence of underlying subpopulations. Indeed, hierarchical clustering of subsets using expression of known pluripotency genes and differentiation markers (Boyer et al., 2006; Cole et al., 2008; Kunath et al., 2007; Ng and Surani, 2011; Xu et al., 2014; Young, 2011) reveals that serum-grown cells split into three distinct groups. These three groups differ in the expression levels of pluripotency factors as well as other genes. In both inhibitory conditions, *Nanog* and other pluripotency factors are less noisy than in serum. Neither 2i nor a2i populations contain a subpopulation structure similar to serum-cultured cells. All 2i cells and all a2i cells (except two) cluster separately from serum, and intermingle with each other. This indicates that 2i and a2i cultured cells are similar with respect to their expression of pluripotency genes (Figure 3.10).

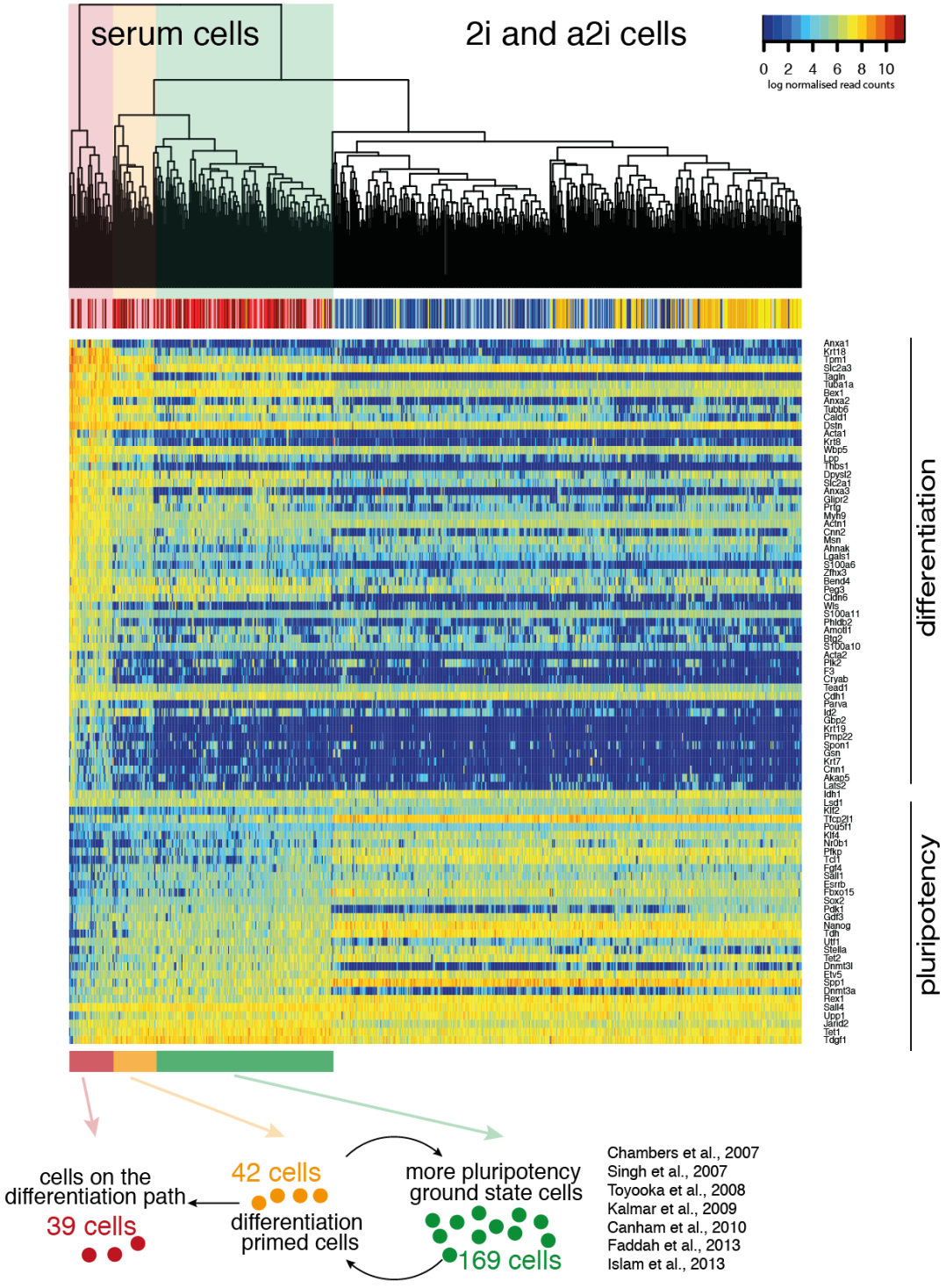


Figure 3.10 Subpopulation structure of cells cultured in serum

Clustering of cells in three culture conditions using a panel of pluripotency factors and differentiation markers. Correlations between cells and genes were calculated using Spearman correlation. Below the heatmap I show a model of the subpopulations of cells grown in serum. The schematic shows cells that express differentiation markers (red), cells that are primed for differentiation while remaining pluripotent (orange) and cells that are closest to ground state of pluripotency (green).

The first subpopulation of cells from serum consists of 39 cells (15%) that express higher levels of markers of differentiation, for example *Fos* or *Hes1*, and high levels of cytoskeletal genes, such as keratins (*Krt8*, *Krt18*), actins (*Acta1*, *Acta2*) and annexins (*Anxa1*, *Anxa2*, *Anxa3*). At the same time, these 39 cells have low levels or no expression of transcription factors involved in maintenance of pluripotency (e.g. *Nanog*, *Sox2* and *Oct4*). This suggests that these cells have exited pluripotency and committed to differentiation. The second group consists of 42 cells (17%) with somewhat lower expression levels of some pluripotency genes, such as *Zfp42* and *Sox2*, and some expression of differentiation genes, yet high expression of *Oct4* and *Dppa3*. These cells may correspond to a previously described "differentiation permissive" set (Chambers et al., 2007; Islam et al., 2014; Kalmar et al., 2009). Finally, the largest group of 169 cells (68%) expresses the highest levels of pluripotency factors and very low expression of keratins or actins (Figure 3.11).

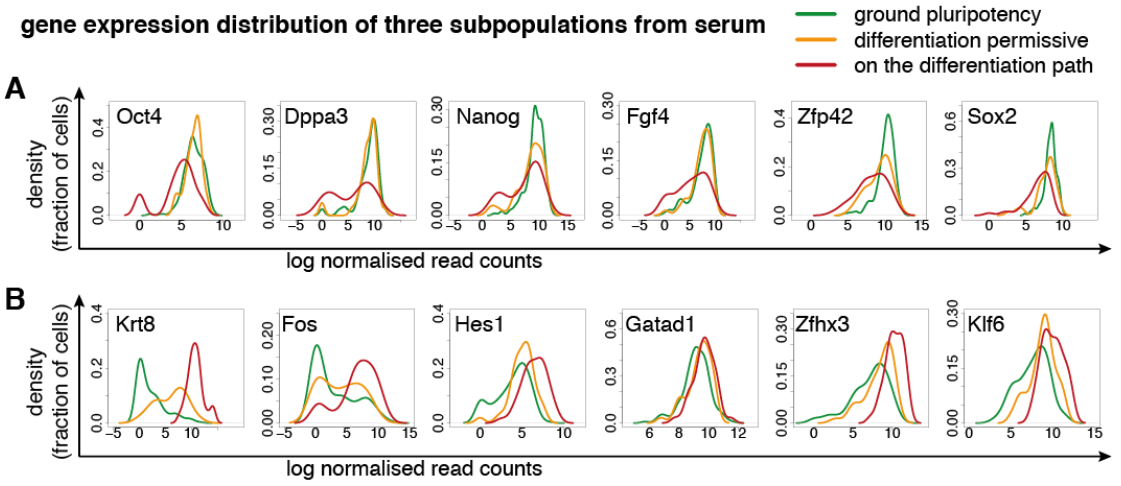


Figure 3.11 Gene expression differences between three clusters of cells in serum
Gene expression distributions of genes that become downregulated (A) and upregulated (B) upon differentiation. Expression is shown as $\log_2$ size factor normalized counts. Oct4 expression is similar in cells closer to the ground state of pluripotency (green) and cells that are primed for differentiation (yellow), and is lower in cells I defined as moving towards differentiation (red).

To examine if cells I identified as 'on the differentiation path' are indeed doing so, I decided to compare them to the cells that differentiate towards neuronal progenitor cells (NPCs). It is known that if signals for pluripotency maintenance are removed, mESCs spontaneously differentiate towards the neuronal lineage (Ying et al., 2003b). I predicted that there would be a similarity between these subpopulation of cells from serum and cells on the NPC differentiation pathway. I used single cell RNA-seq data generated by Dr. Alex Tuck from mESC cultured in serum and the same cells at day 6 and day 8 of an NPC differentiation time course (Bibel et al., 2007). I performed principal component analysis of Spearman's rank correlation coefficient between all the cells and I observed that cells belonging to the Nanog-low subpopulation lie between the more pluripotent cells and these that are differentiating towards NPCs (Figure 3.12). This strongly supports our earlier hypothesis that these cells are indeed progressing down a differentiation pathway.

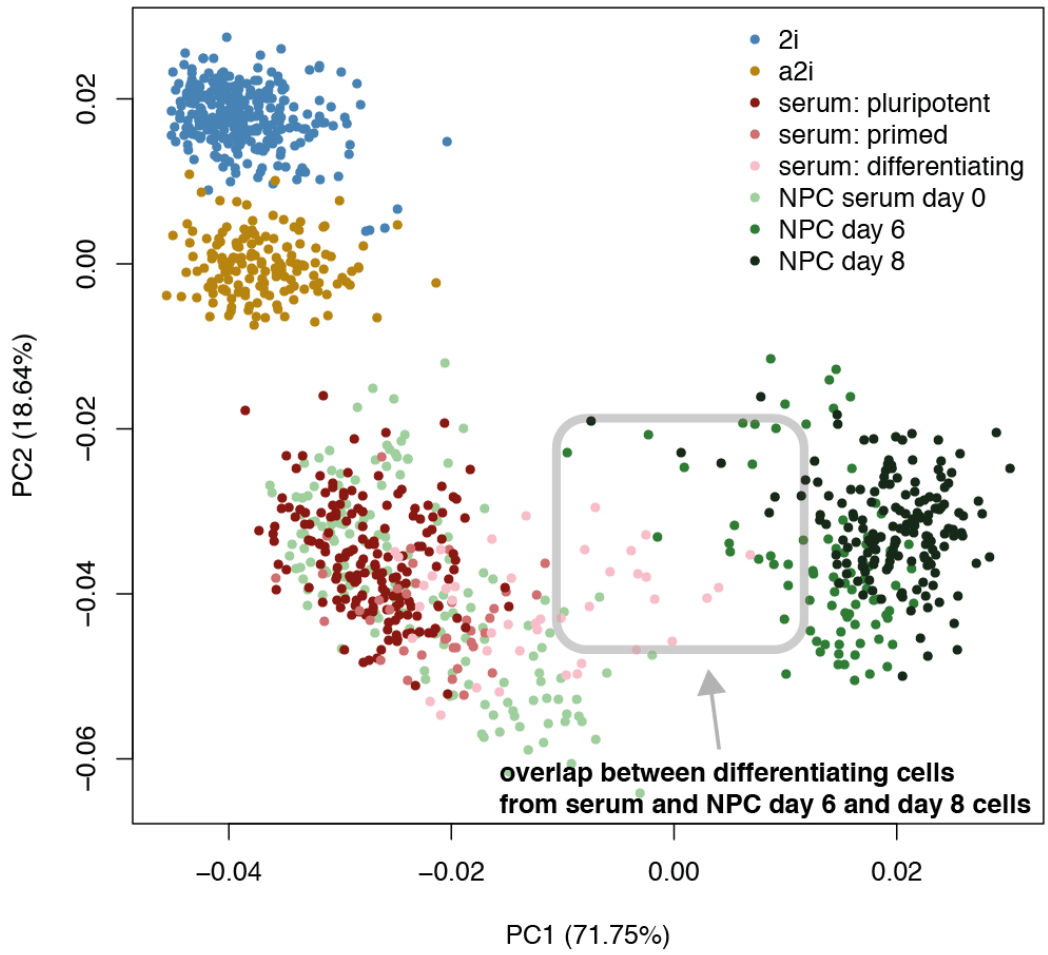


Figure 3.12 Principal component analysis of expression data from serum and cells progressing towards NPC fate.

All genes with mean normalized counts larger than 50 were considered and PCA was performed on the Spearman's rank correlation matrix between cells.

Identification of a pluripotent mESC population in serum, led me to ask if these cells are the same as the ground pluripotent state cells found in 2i condition. I performed PCA to see if there is overlap between these populations, but observed that cells cultured in each condition cluster separately, meaning that they have distinct transcriptomic states. PC1 separates the culture conditions and genes that contribute the most to this

separation are genes involved in development as well as metabolism. Notably, cells from replicates of each culture condition cluster together showing that the separation of three culture conditions is due to biological difference rather than to batch effect (Figure 3.13).

I performed GO term analysis of genes that contributes most to PC1, which separates the conditions (Figure 3.13 BC). GO term "positive regulation of mesenchymal cell proliferation" among others contains genes from WNT and Sonic Hedgehog pathways, several fibroblast growth factors and transcription factors from Forkhead family, "lung development" also contains members of WNT pathway, several types of growth factors including leukaemia inhibitory factor and transcription factors including for example *Nodal*. Similarly terms "ossification", "neuron projection development", and "positive regulation of vasoconstriction" contain genes that function also in early development or in development and signalling in general. Appearance of "inactivation of MAPK activity" term is probably related to the fact that in 2i and a2i, MAPK is inhibited using drug. "Cell-cell adhesion" related genes are differently affected in a2i, in which SRC is inhibited and one of SRC functions is phosphorylation of focal adhesion kinase (FAK) (Meyn and Smithgall, 2009; Shimizu et al., 2012). Genes related to metabolism "glycolysis", "ribosomal subunit assembly", "translation" may reflect different metabolic states between serum and 2i as well as differences that come from different growth rates.

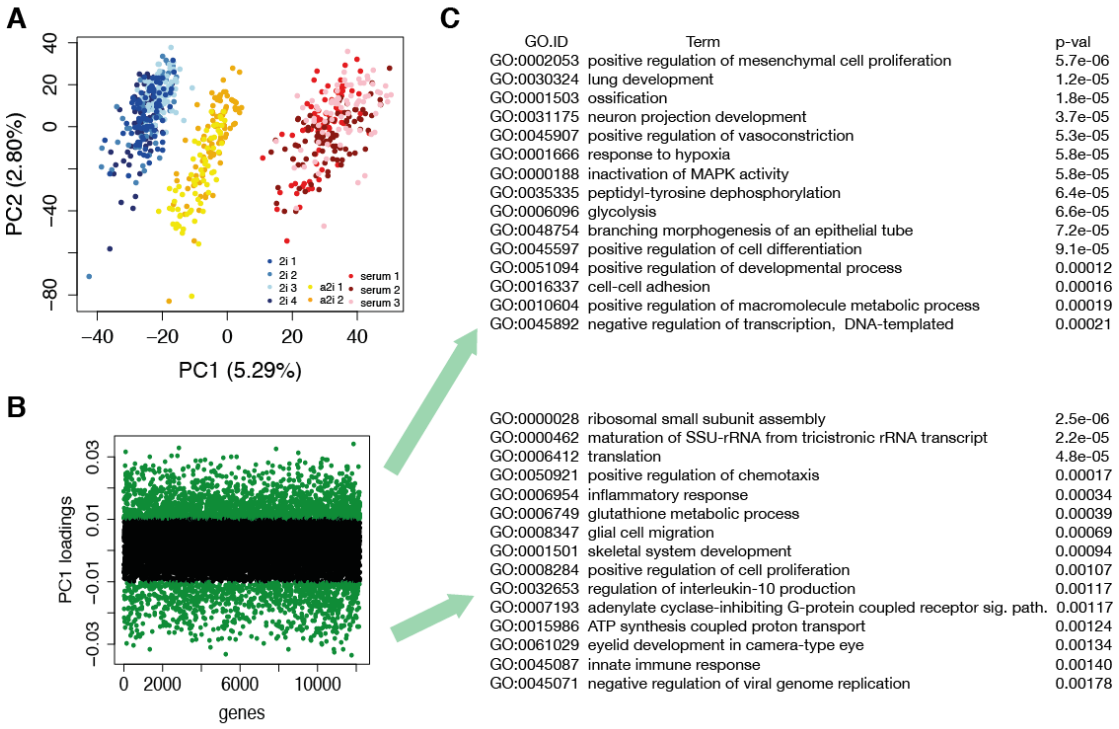


Figure 3.13 Clustering of mESCs grown in serum, 2i and a2i media
All cells (n=704) grown in the three different culture conditions are projected onto the first two principal components. All genes with mean normalized read counts larger than 10 were considered and principal component analysis (PCA) was performed. (B) Distribution of genes contributing to PC1. (C) Gene ontology enrichment analysis of genes most strongly contributing to PC1 separation.

3.7 Cell cycle variability in 2i and alternative 2i cultures

When we compared gene expression heterogeneity of different functional gene categories it was unexpected to see that cell cycle genes will have lower gene expression variability in serum than in the inhibitory conditions, because all of these cells cycle (Figure 3.8). To understand where this difference comes from I decided to analyse cell cycle gene expression of cells in three culture conditions. I used Cyclebase.org database, which uses experimental data from synchronized cells to rank genes from these that show the most consistent and pronounced cycling pattern (Santos et al., 2015). I selected 20 genes that have most pronounced cycling behaviour in their expression with peak in G2 or M

phase and found their mouse orthologs. When clustering cells based on these genes only, I found that 2i and a2i cells separate more clearly into two groups: one with high expression of G2 and M genes and the other with lower expression of these genes, suggesting that these remaining cells are in G1 or S phases of cell cycle (Figure 3.14).

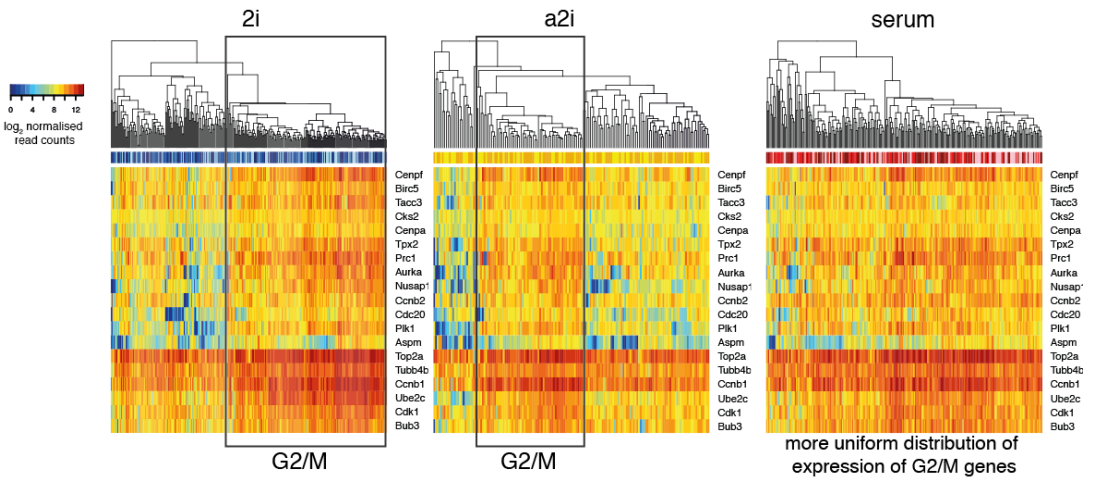


Figure 3.14 Cell cycle gene heterogeneity and cell cycle phase assignment
Heat maps showing the expression of cell cycle related genes in serum, 2i and a2i, with a distinct separation into G1/S versus G2/M cells in 2i and a2i, with less distinction between individual cells in serum.

To confirm that this annotation of cell cycle phases to cells is correct, I estimated mRNA content of cells using ERCC spike-ins (Consortium, 2005). Each cell was spiked with exactly the same amount of ERCCs and thus the ratio of reads mapping to ERCCs to reads mapping to all mouse genes depends only on the amount of transcripts in the cell and the higher it is the lower mRNA content of the cell. To make sure that lysis buffer spiked with ERCC is exactly the same in all samples, for this analysis I used only batch 3 of the data, which was done on one day in parallel. As expected, cells in the G1 and S phases in both 2i and a2i have significantly higher ratio of reads mapping to ERCCs to reads mapping to all mouse genes, meaning they have

less mRNA. There is significantly more mRNA in cells identified to be in G2/M phase in comparison to G1/S phase cells in both 2i and a2i. As the cells in these populations are not normally distributed I used the non-parametric Wilcoxon test (Figure 3.15).

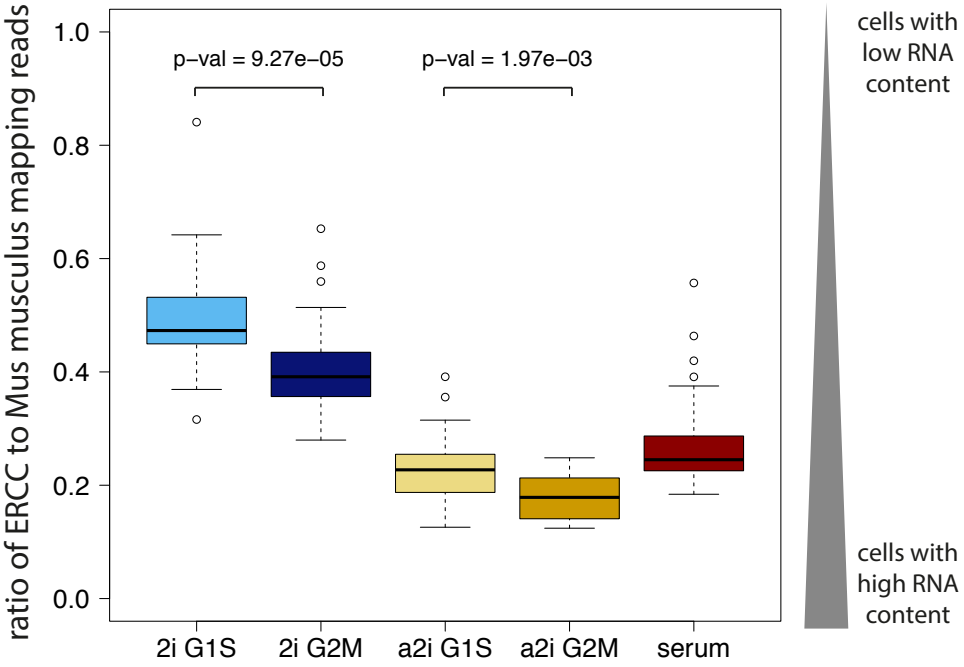


Figure 3.15 mRNA content in cells at different cell cycle stages
Comparison of mRNA content in cells using ratio of reads mapping to ERCCs (constant number of molecules spiked in in three conditions) to all exon mapped reads.

Another measure to check if the assignment is correct would be to see if cells from G1 and S phase have higher expression of histones. During S phase cell needs to double the amount of histones to package newly synthesized DNA, thus in G1 and S phase cell should have more histone transcripts. Indeed I observe that pattern in both 2i and a2i, suggesting that our classification of cell cycle phases is correct (Figure 3.16).

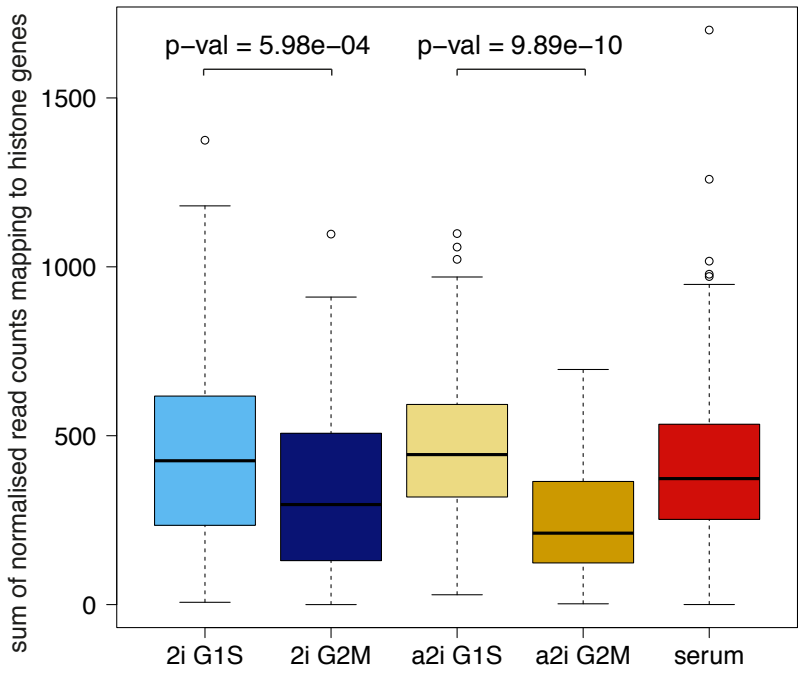


Figure 3.16 Histone mRNA expression in cells at different cell cycle stages
Comparison of histone mRNA content in cells from different cell cycle stages across culture conditions.

Cyclone is a machine learning based approach for cell cycle phase assignment; it can distinguish G1, S and G2/M phases (Scialdone et al., 2014). I used it for cell cycle phase prediction and it is in a good agreement with the assignment I made by clustering, 88% for 2i cells and 90% for a2i cells. In 28 cases (9.5%) in 2i, and 11 cases (7%) Cyclone identified cells to be in S phase, and I in G2/M. Only one cell in 2i was identified as G1 by Cyclone and G2/M by clustering. And 6 cells (2%) in 2i and 5 cells (3%) in a2i were assigned by Cyclone as G2/M and clustering identified it as G1/S.

3.8 Speed of cell cycle estimation from single cell mRNA sequencing data of cell population

To understand the source of the difference between 2i/a2i and serum with respect to the cell cycle I examined doubling rate of these cells and found that cells in serum and 2i showed different doubling kinetics (Figure 3.17). Within the first 24h the growth rate was faster in 2i than serum but later, at day 2, it slows down. At the time of harvest (48 hours after plating), the doubling time is 25 hours for 2i cells and 11 hours for serum, indicating that cells grown in 2i are more slowly cycling, probably due to a longer G1 phase. Degradation rates of mRNAs in serum and in 2i are similar, and average mRNA half time is about 7h, but many cell cycle genes have longer half lives (Sharova et al., 2009). The correspondence of lengthening doubling time and increasing cell cycle associated gene expression noise demonstrated the robustness of single cell transcriptomic 'snapshots' of specific biological process in a cell population.

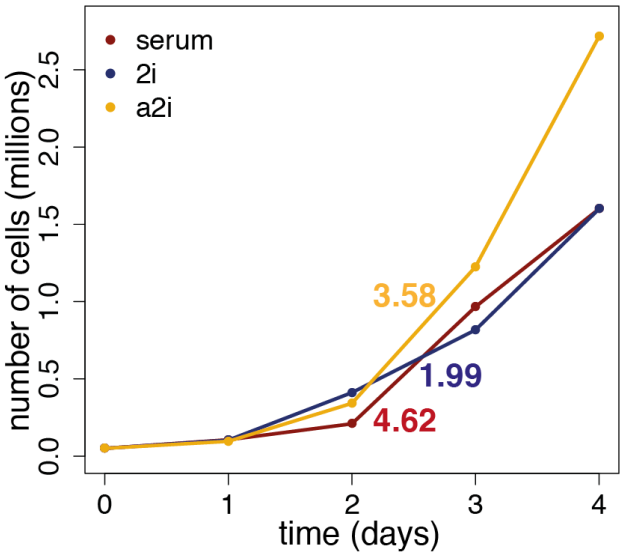


Figure 3.17 Growth kinetics of cells in three culture conditions.
Numbers shown are how many times cells grew between second and third day of culture, i.e. when cells were harvested for scRNA-seq experiment. At this point in culture cells cultured in serum grew slowest.

Additionally, I observed that the 39 and 42 cells from serum culture, which have begun to move forward on the differentiation pathway, have noisier expression of cell cycle genes. A shift in the distribution of the expression of G2/M genes, such as *Cks2* or *Cdc20* toward lower levels suggests that there are relatively more G1/S cells in these two groups (Figure 3.18). I inferred that more differentiated cells have a relatively longer G1 phase, as I sample more cells in G1 from this subpopulation in comparison to more pluripotent cells. This indicates that cells that I identified as differentiating have a longer cell cycle, and are proliferating more slowly than Nanog-high ground state pluripotent cells.

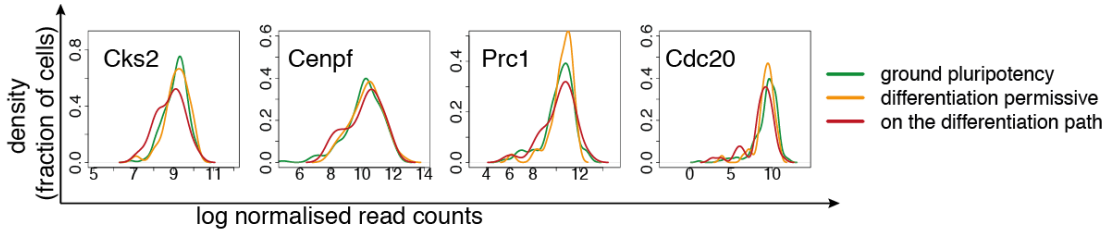


Figure 3.18 Gene expression distributions of cell cycle genes in subpopulation of cells cultured in serum.
Plots show distribution of cell cycle gene expression in cells from three subpopulations from serum. Cells that are on the differentiation path (red) are more heterogeneous than cells that are in the more pluripotent state (green).

To support and demonstrate further the fact that differentiating cells that start to cycle more slowly have more heterogeneous cell cycle gene expression distribution I used the NPC differentiation time course data. The distributions of the expression of cell cycle genes are significantly more heterogeneous in differentiating cells. For some genes, such as *Cdc20*, one can observe bimodal distribution in NPC differentiated cells from day 6 and day 8.

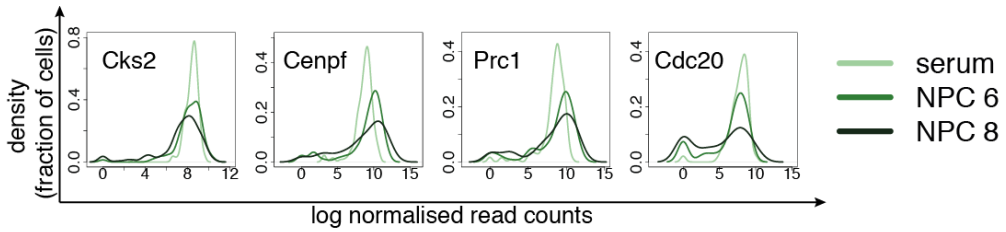


Figure 3.19 Gene expression distributions of cell cycle genes in cells from NPC differentiation time course

Plots show distribution of cell cycle gene expression in cells from NPC differentiation time course. Cells that are not differentiated (serum, light green) are more homogeneous than cells that are 6 or 8 days on the NPC differentiation path (darker green).

3.9 Cell Cycle Rank for measurement of cell cycle speed

Cell cycle gene expression is heterogeneous and this heterogeneity does not come only from the fact that cells are in different cell cycle phases and from the speed of cell cycle, but also from the heterogeneity due to the stochastic nature of gene expression, by bursts rather than continuously. This additional noise makes it difficult to see significant differences between populations, if few cells were sampled. For example the differences between gene expression distributions of cell cycle genes in subpopulation of cells cultured in serum are subtle if one looks at a single gene (Figure 3.18).

To overcome this problem I developed a measure called Cell Cycle Rank, which allows overcoming the effects caused by stochasticity of gene expression. To calculate the Cell Cycle Rank, 20 genes that have highest cyclic expression pattern and peak at G2 or M phases were selected from cyclebase.org and for each of these genes cells were ranked depending on how highly this gene is expressed. Subsequently ranks for these 20 genes were summed up for each cell. Cells that have high Cell Cycle Rank, express all 20 genes highly suggesting that they are likely to be G2/M cells, while those with

low rank are in G1/S phases. By summing the ranks I do not take under consideration the level of gene expression, so more highly expressed genes do not influence the result more than lowly expressed genes.

I calculated Cell Cycle Ranks for cells differentiating to NPC and plotted the distributions and as expected they are more heterogeneous for cells that are more differentiated (Figure 3.20 A). More interestingly, when I apply this method to the subpopulations of cells from serum, I can clearly see that cells identified as differentiating have a broader distribution of Cell Cycle Ranks in comparison to more ground state cells (Figure 3.20 B).

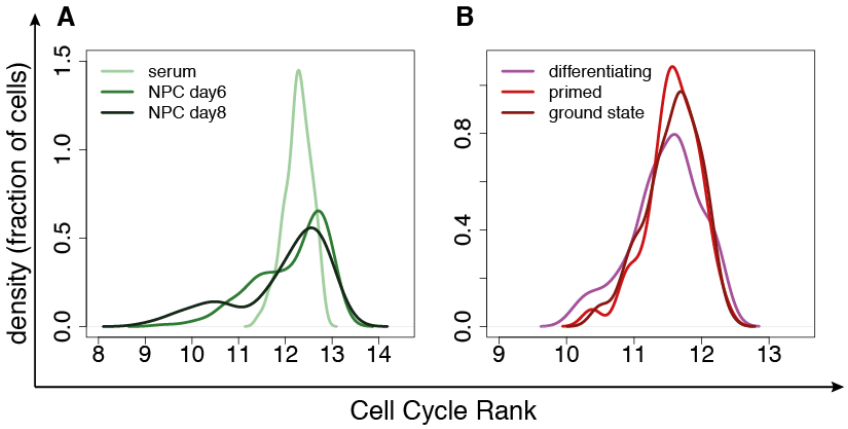


Figure 3.20 Cell Cycle Rank distribution
Distribution of Cell Cycle Ranks for (A) cells from NPC differentiation time course and (B) subpopulation of cells cultured in serum.

3.10 Conclusions

To quantify cell-to-cell heterogeneity in gene expression levels, for the first time in single cell RNA sequencing analysis we applied distance to the median, a measure of noise that is independent of gene expression level. Surprisingly, we found that on a global level, cells grown in 2i, a2i and serum are indistinguishable in terms of transcriptome-wide heterogeneity. It was assumed, based on expression of a small number of pluripotency markers, that

cells grown in serum are more heterogeneous. I have shown, however, that the noise composition of specific subsets of genes is different between the culture conditions. The noise in 2i was not captured previously, because it involves different gene sets than these that display heterogeneous expression in serum. Cells grown in serum, as observed previously, have more heterogeneous expression of pluripotency factors. This derives from the existence of subpopulations that differ in the expression of these genes.

Within the serum population I find that there are three clusters of cells, which likely correspond to different states of pluripotency versus differentiation. Previously, subpopulations of cells in serum were reported based on FACS analysis of proteins with heterogeneous abundance such as NANOG (Kalmar et al., 2009; Singh et al., 2007). Cells with low expression levels of *Nanog* were separated from those expressing *Nanog* at high levels, and microarray analysis of the transcriptomes of these two subpopulations was performed (Singh et al., 2007). This work showed that *Rex1* (*Zfp42*), *Sox2* and *Pou5f1* are more highly expressed in Nanog-high cells, a pattern I also observe.

Recently, single cell RNA sequencing of serum-grown mESCs (Islam et al., 2014) showed a subpopulation with low *Nanog* expression. In another large-scale study, using droplet microfluidics it was shown that there exist subpopulations of cells cultured in serum (Klein et al., 2015). In this study the authors sequenced several thousands of cells and were able to find precursors of different lineages in the embryo. Additionally, a qPCR study using a panel of 48 pluripotency markers showed that cells cultured in serum exist in two distinct states, with a small number of cells appearing to reside in an intermediate state (Papatsenko et al., 2015). I extended this analysis, and found three clusters, one of which represents differentiation-committed cells, one

represents an intermediate state and one represents a self-renewal state. I speculate that the first subpopulation has committed to differentiation with clear down-regulation of *Pou5f1* and *Sox2*, suggestive of irreversible commitment. In contrast, "differentiation primed" cells with higher expression of *Pou5f1* and *Sox2* could still revert to "pluripotent" cells. Additionally, the proportion of cells in G1 or S phase of the cell cycle increases in the "differentiated" cells, suggesting that their cell cycle is slower and that they do not expand as quickly as the more pluripotent populations. Importantly, I found that cells that express high levels of *Nanog* in serum are not similar to 'ground pluripotency state' 2i cells.

Our results show that mESCs partition into transcriptomically distinct cell populations according to the growth medium (serum, 2i or a2i). Cells cultured in 2i and a2i are similar to each other. When compared to single cells from different stages of mouse embryonic development, all three sets of cultured mESCs are closest to cells from the blastocyst stage, which is the stage from which the cells were extracted originally. The 2i and a2i cultured ESCs seem more similar to the blastocyst cells than serum cells. This is in agreement with previous findings showing that cells cultured in 2i are hypomethylated due to inhibition of Gsk3β and MEK. Similar low level of methylation is observed in the preimplantation epiblast, suggesting that these cells are in the naïve pluripotent state (Leitch et al., 2013). Regarding metabolic state, cells cultured in 2i have lower expression levels of glycolysis enzymes in comparison to serum.

Importantly cell cultured in 2i are not identical to blastocyst cells. This is expected because *in vitro* conditions are non-physiological especially in case of 2i media where pluripotent state is achieved by use of kinase inhibitors.

Additionally, I observed that 2C-like cells are globally more similar to blastocysts than to 2-cell stage embryonic cells.

A2i medium has been described as an alternative ground state that can be achieved through the use of a different inhibitor (Shimizu et al., 2012). As expected, a2i is not identical to 2i, but I believe that it is rightfully called an alternative ground state: on the transcriptome level, especially with respect to pluripotency genes, a2i cells are similar to 2i and *in vivo* blastocyst cells. In 2i and a2i media, there are no subpopulations of differentiating cells, hence the pluripotency genes are expressed more homogeneously. Despite these similarities, it is intriguing to note that a2i cells have a cellular RNA content similar to serum-cultured cells, while 2i cells contain about half as much RNA on average, independent of cell cycle stage. It should be noted that *Myc* is differentially up-regulated in a2i cells compared to 2i cells. As *Myc* has recently been shown to behave as a transcriptional amplifier of active genes (Lin et al., 2012; Nie et al., 2012) it provides a potential mechanic basis for the elevated of RNA content in a2i cells.

More generally, I observed a relationship between variability in the expression levels of cell cycle genes and the length of the cell cycle. Cells cultured in serum have the lowest level of noise, cells in a2i medium and cells in 2i the highest, which correlates negatively with doubling times in culture (doubling times quickest for serum and slowest for 2i). For dividing populations where the cell cycle is very slow, such as HSCs, it is possible to assign cells to one of four cell cycle stages, but this is more challenging for that cycle more quickly (Tsang et al., 2015).

In summary, single cell transcriptomics has allowed us to gain deep insights into the subpopulation structure within mES cell cultures. These results

emphasize the power of transcriptomics at single cell resolution for understanding multiple biological processes.

3.11 Further research

Results and conclusions of this study lead to new questions about biology of stem cells and pluripotency.

Self-renewal is a defining feature of stem cells and there are links between pluripotency and cell cycle, for example *via Myc* (Singh and Dalton, 2009), but it is not entirely clear what role cell cycle has in the pluripotency maintenance. In 2i medium cell cycle is targeted by inhibition of MAPK pathway, suggesting that this is essential for keeping cells pluripotent (Orford and Scadden, 2008). Additionally, LIF signalling *via* STAT3 is linked to the cell cycle regulatory pathways (Burdon et al., 2002). Furthermore, others and I observed that cells that differentiate start cycling slower, suggesting that there is a change in cell cycle. The link between cell cycle and pluripotency can be unravelled using single cell mRNA sequencing as one can assign cell cycle phases to cells and simultaneously monitor their pluripotency state.

Measuring cycling speed of cells is important especially for understanding cancerous cell populations. It is difficult to measure it without performing several time course measurements and additionally in very complex populations as in tumours it may be particularly difficult. By performing single cell mRNA sequencing one can first identify cell cycle populations of which the tumour is composed and subsequently identify cell cycle profiles of these cells and measure cell cycle heterogeneity. This will give an insight into, which cells are multiplying faster and thus predict which population will proliferate most aggressively. The ultimate goal could be finding an absolute

rather than relative measure of cell cycle speed using the heterogeneity of cell cycle genes and cell cycle phase profile.

Chapter 4

Characterization of 2C-like cells

4.1 Introduction

Mouse embryonic stem cells (mESCs) are derived from the inner cell mass of the blastocyst, which is already separated from the trophectoderm lineage that becomes part of the placenta. If injected into an embryo mESCs contribute to all tissues of the foetus, but are extremely inefficient at colonizing extraembryonic tissues (Bradley et al., 1984). It was suggested that these rare cases of contribution to extraembryonic lineages comes from either contamination with trophectoderm cells or from a subpopulation of so-called "2C-like cells", which have the potential to differentiate into trophectoderm (Macfarlan et al., 2012).

2C-like cells are described as a very rare cell population and express some markers of the 2-cell stage of embryonic development such as *Zscan4* family genes, which contain a zinc finger domain that mediates DNA binding and a SCAN domain responsible for oligomerisation. In the context of mouse

embryonic stem cells, *Zscan4* genes were suggested to function in the maintenance of pluripotency and genome integrity (Zalzman et al., 2010). Other proposed markers of 2C-like cells include several transcription factors, some with no clearly identified function, such as *Zfp352* and *Zswim2* and some that are factors signalling to the MAPK pathway, such as heparin-binding growth factor 1 (Fgf1) and keratinocyte growth factor (FGF7). 2C-like cells are also characterized by higher expression of pluripotency factor *Fbx15* (Tokuzawa et al., 2003) and higher *Tcstv1* and *Tcstv3* that were shown to function in telomere elongation in mouse embryonic stem cells (Zhang et al., 2016). Additionally these cells are also characterized by expression of the MuERV-L endogenous retrovirus and chimeric transcripts that arise *via* retroviral insertion in different places in the genome (Macfarlan et al., 2012).

In addition to being expressed in 2C-like cells, *in vivo* MuERV-L expression is initiated during S-phase of the cell cycle of zygote, which is the onset of zygote genome activation (ZGA), peaks at the two cell stage and then is efficiently downregulated (Kigami et al., 2002).

MuERV-L is an endogenous retrovirus, a type of transposable element that can duplicate and reinsert into the genome (Bénit et al., 1997). The structure of MuERV-L consists of long terminal repeats (LTRs) at 5′ and 3′ of the element and *Gag* and *Pol* genes in between them. Importantly, in contrast to retroviruses that can be horizontally transferred to other cells, MuERV-L does not contain the *env* gene that codes proteins that make up the capsid thus cannot form viral particles (Peaston et al., 2004; Schlesinger and Goff, 2015).

Figure 4.1 Structure of MuERV-L transposable element

LTRs contain presumptive TATA box and polyadenylation signals for expression of their genes, but most interestingly LTRs serve as alternative promoters for several developmental genes, if the virus inserts upstream of the gene. In this case chimeric transcripts between MuERV-L genes and peripheral genes are formed. This mode of gene regulation seems to be important in early development when due to epigenetic reprogramming and massive demethylation, transposable element sequences are derepressed (Macfarlan et al., 2011; Peaston et al., 2004).

It was suggested that expression of 2C marker genes is regulated by LTRs located upstream that lead to formation of chimeric transcripts. In adult tissues MuERV-L and other transposable elements are silenced *via* methylation, or histone modifications, including H3K27 methylation and H3 and H4 acetylation, mediated by KDM1A, G9A, KAP1 and HDACs (Macfarlan et al., 2012; Maksakova et al., 2013; Schlesinger and Goff, 2015).

2C-like cells have downregulated expression of key pluripotency markers *Pou5f1*, *Sox2* and *Nanog* at the protein level, but at the mRNA level they are indistinguishable from the rest of the population when it comes to the expression of these factors, suggesting regulation at the translation or protein degradation levels (Macfarlan et al., 2012; Schlesinger and Goff, 2015).

To explore evidence for the existence of this rare cell type I aimed to identify these cells in the single cell mRNA-seq data I collected (for details please refer to Chapter 3) and subsequently to characterize their transcriptomic profiles. Furthermore, as these cells were thought to resemble cells of the 2 cell embryo I wanted to compare the transcriptome of 2C-like cells to transcriptomes of cells from early stages of development.

4.2 Identification and characterization of 2C-like cells in 2i medium

To identify 2C-like cells in our samples, I examined the expression profile of genes shown previously to have at least 10-fold enrichment in 2C-like cells in comparison to the remaining mESCs (Macfarlan et al., 2012). Hierarchical clustering suggested the presence of 10 2C-like cells in 2i, and none in the a2i or serum culture conditions (Figure 4.2). Frequency of 2C-like cells within mESC culture is normally very low, often below 1%, thus the fact that I did not identify any 2C-like cells from a small number of cells, is not a proof of their absence. Most likely they are still present but at a very low rate.

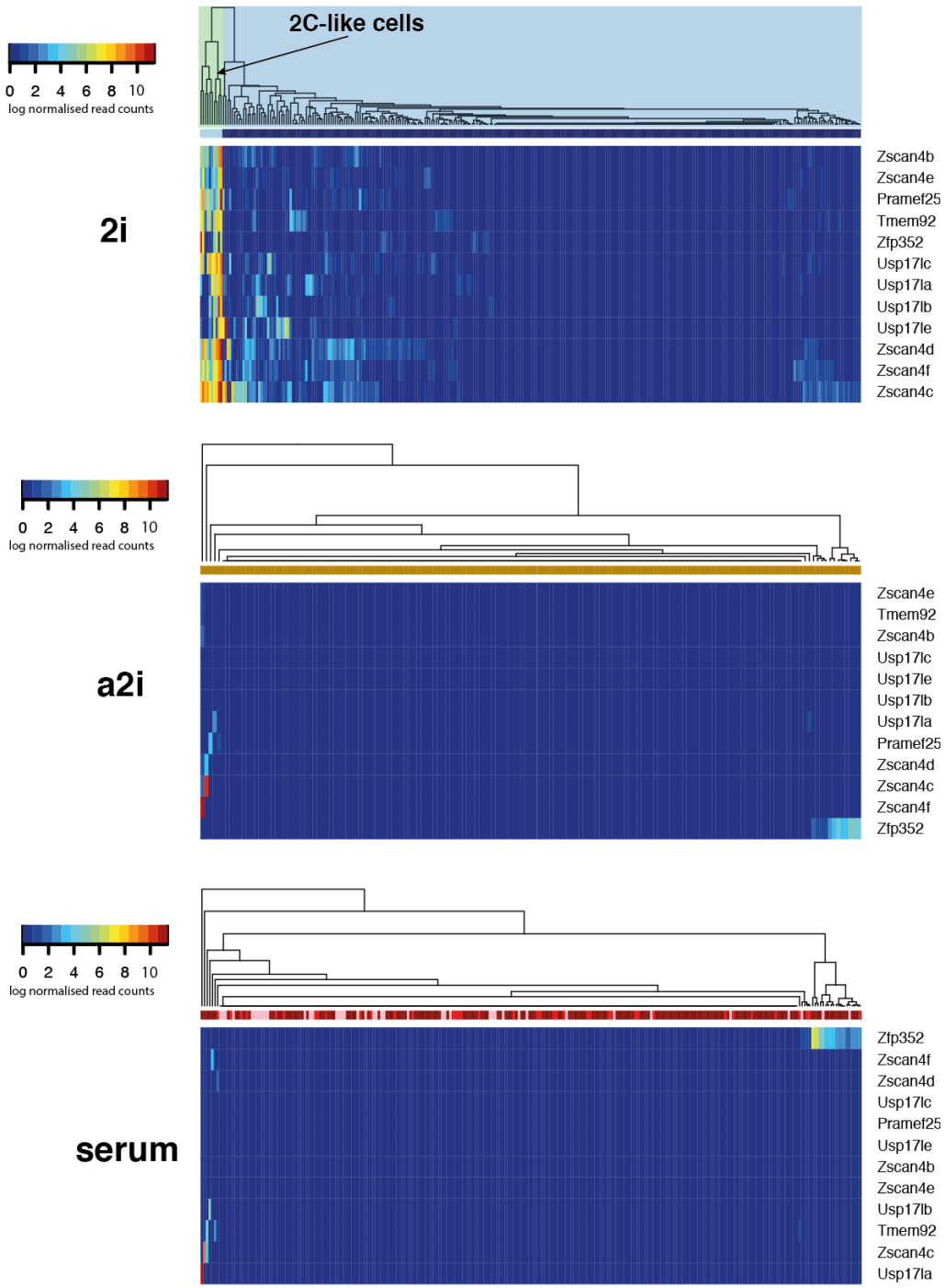


Figure 4.2 Identification of 2C-like cells

The first heatmap shows clustering of cells grown in 2i using markers of 2C-like state (Macfarlan et al., 2012). The dendrogram divides cells into two groups, one of which contains 10 cells expressing 2C-markers. The heatmaps below show no clearly defined subpopulations in a2i and serum.

I observed that, globally, the transcriptomes of 2C cells are altered, and only about 50% of reads on average map to exons, in comparison to 60% in the remaining population in 2i (Figure 4.3A). I hypothesized that this was due to greater transcription from unannotated MuERV-L sequences. I also considered the number of sequencing reads mapped to the MuERV-L reference sequence. I do indeed observe MuERV-L expression in 2C-like cells and no expression in the remaining cells (Figure 4.3B).

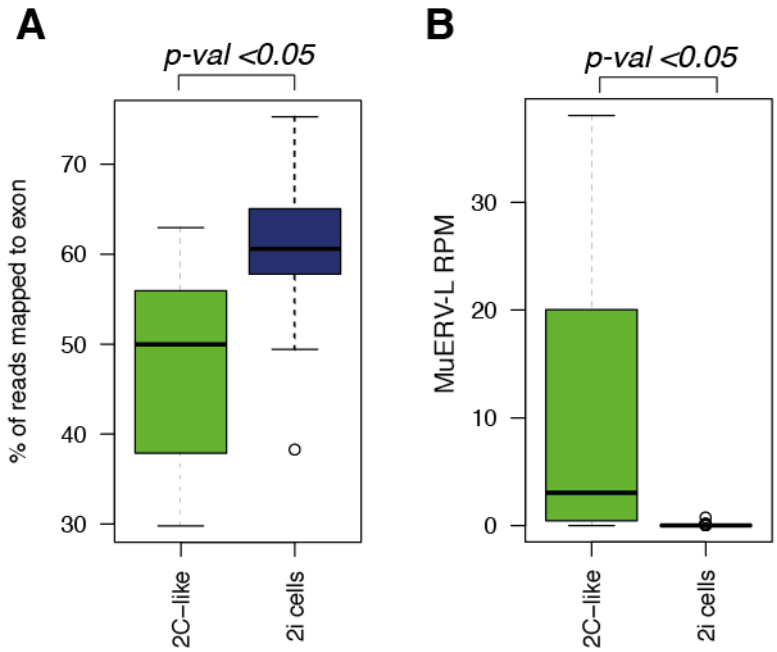


Figure 4.3 Exon mapping reads and endogenous retrovirus expression
(A) Boxplot showing % of reads mapping to the exons in both subpopulations of cells in 2i. P-value was calculated using Wilcoxon test. (B) Boxplot showing RPM (reads per million) mapping to the MuERV-L retrovirus in both subpopulations of cells in 2i. P-value was calculated using Wilcoxon test.

As a further means of assessing whether this population corresponds to a 2C-like state, I calculated mean expression of the genes identified by MacFarlan, (2012) as differentially expressed in 2C-like cells (Macfarlan et al., 2012). I observed that most of the genes that were shown previously to be

enriched in 2C-like cells are also enriched in 2C-like cells in our experiment (Figure 4.4).

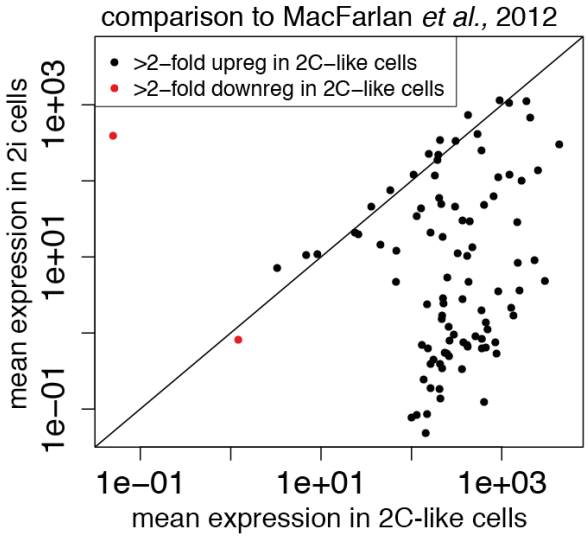


Figure 4.4 Comparison of differential expression results to MacFarlan
Mean expression of genes reported to be at least 2-fold upregulated or downregulated in 2C-like cells (Macfarlan et al., 2012) in cells that I identified as 2C-like cells and in the remaining 2i cells.

4.3 2C-like cells characterization

To characterize the 2C-like cells identified within this dataset I preformed differential expression analysis using DESeq and found 234 genes that are significantly upregulated in 2C-like cells in comparison to the rest of 2i cells, and 60 genes that are downregulated (Figure 4.5A, for the full list refer to appendix). Gene Ontology (GO) enrichment analysis did not reveal any significant terms within the downregulated gene set, but showed that there is some enrichment in upregulated genes related to metabolism (Figure 4.5B).

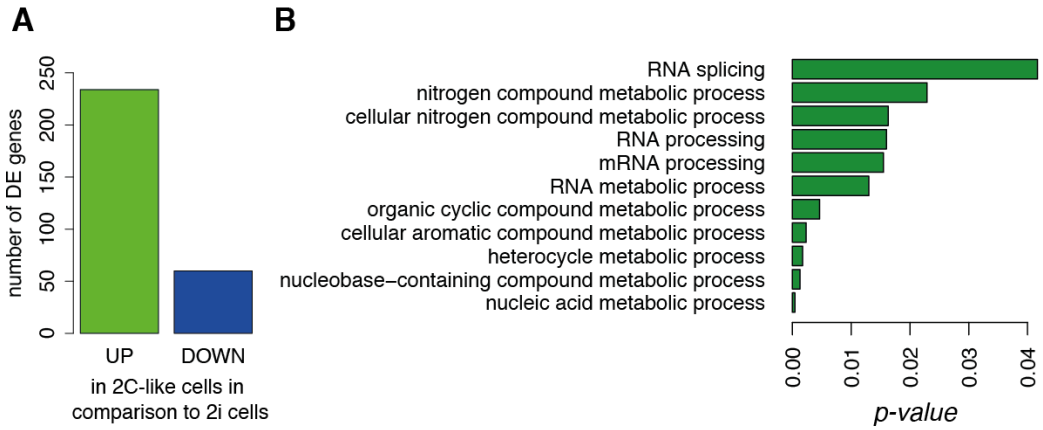


Figure 4.5 Differential expression analysis between 2C-like cells and 2i cells
(A) Bar plot showing the number of significantly (DESeq, adjusted *p*-val < 0.05) upregulated and downregulated genes in 2C-like cells. (B) Plot shows most significantly enriched gene ontology terms. *p*-value is corrected for multiple hypothesis using Benjamini-Hochberg method.

In addition to performing GO analysis I inspected all the genes that were upregulated in 2C-like cells to identify those that could bring some insight about the biology of these cells. There are several tens of genes without known function in this group and many that relate to RNA processing and metabolism as GO analysis suggested. Interestingly there are also several genes that function in the ubiquitin–proteasome pathway (*Fbxo15*, *Arih2*, *Cand1*, *Rbbp6*, *Cul5*, *Cbl*, *Ube2t*, *Usp17la*, *Usp17lb*, *Usp17lc*, *Usp17ld*) and Ca2+ uptake and binding related genes (*Calhm3*, *Micu1*, *Guca1a*, *Cldn12*, *Cab39*, *Cacna1s*). There are DNA binding genes, including the *Zscan4* family and many zinc-finger proteins of unknown function.

Interestingly there are several genes that function in DNA repair (*C1d*, *Ccnf*, *Ercc4*, *Rad51b*, *Rif1*) and genes that are related to viruses and retrotransposition (*Trim28*, *Zfp809*). In more detail, *C1d* and *Rif1* were shown to be associated with non-homologous end joining mechanism of DNA repair (Chapman et al.,

2013; Erdemir et al., 2002; Escribano-Diaz et al., 2013; Yavuzer et al., 1998; Zimmermann et al., 2013). Cyclin F (*Ccnf*) in addition to its role in regulation of cell cycle, functions in regulation of the DNA damage stress response (D'Angiolella et al., 2012). *Ercc4* encodes DNA repair endonuclease XPF, which functions in nucleotide excision repair and DNA double-strand break repair (Ahmad et al., 2008; Al-Minawi et al., 2008; Niedernhofer et al., 2001). On the other hand, *Rad51b* promotes homologous recombinational DNA repair (Sigurdsson et al., 2001; Takata et al., 2000; Yokoyama et al., 2003). *Trim28* and *Zfp809* regulate epigenetic silencing of retrotransposons and retrotransposition derived regulatory elements (Rowe et al., 2013; Turelli et al., 2014; Wolf and Goff, 2007, 2009; Wolf et al., 2015).

There are some genes that were shown to be important for pluripotency such as *Dppa2* (Du et al., 2010), *Mtf2* (Zhang et al., 2011), *Ncoa2* (Wu et al., 2012), *Ppp1r8* (Van Eynde et al., 2004), *Snw1* (Wu et al., 2011), *Trim43a* (Stanghellini et al., 2009), *Zfp217* (Aguilo et al., 2015). Furthermore, I checked if the expression levels of *Nanog*, *Oct4* and *Sox2* are indeed the same in 2i and 2C-like cells. There is no significant difference in expression between 2i and 2C-like cells (Wilcoxon test p-val > 0.05) for expression of these three markers (Figure 4.6).

4.4 Comparison to *in vivo* embryo cells

As the name suggests 2C-like cells were proposed to resemble the 2 cell stage of the embryo. This prompted us to investigate how similar 2C-like cells are to 2 cell stage embryos. To do this I used single cell mRNA-seq data from Deng and colleagues (Deng et al., 2014) who assayed cells from each stage of early embryo development. I first compared 2C-like cells to the rest of cells

from 2i culture, cells from *in vivo* blastocyst and 2 cell stage of the embryo. In terms of expression of key pluripotency genes, such as *Nanog*, *Oct4*, *Sox2* and *Myc*, 2C-like cells are most similar to 2i cells in comparison to the 2-cell and blastocyst stages of the embryo. On the other hand, as shown before *Zscan4* genes are exclusively expressed at 2 cell stage of the embryo and in 2C-like cells (Figure 4.6).

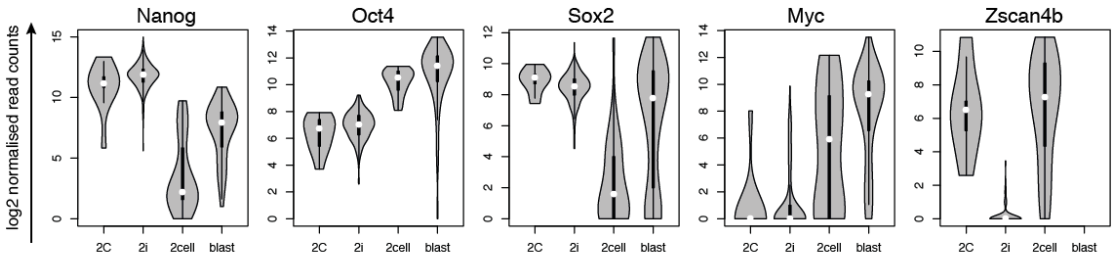


Figure 4.6 Key pluripotency genes in 2C-like cells, 2i cultured cells and in cells from *in vivo* embryo

Expression of key pluripotency genes in 2C-like cells (2C), and the rest of cells grown in 2i media (2i), cells from the 2-cell stage (2cell) and cells from the blastocyst stage (blast) of the embryo.

For global comparison of transcriptomes of *in vitro* cultured mESCs and cells for embryos I performed principle component analysis on the Spearman's rank correlation coefficient between our data and data from Deng et al. It showed that 2C-like cells are more similar to 2i cells and blastocyst than to cells from the 2-cell stage of the *in vivo* embryo (Figure 4.7). 2C-like cells cluster together with 2i cells, and there are only 294 differentially expressed genes between 2C-like cells and the remaining 2i cells. In comparison, I find 3056 differentially expressed genes between 2i and serum, 1700 genes between 2C-like cells and blastocyst and 1779 between 2C-like cells and 2-cell stage cells. This suggests that 2C-like cells share more characteristics of mESC cultured in 2i and blastocyst from which these cells are derived rather than cells from 2

cell stage embryos, although they express a few of the markers present at that stage.

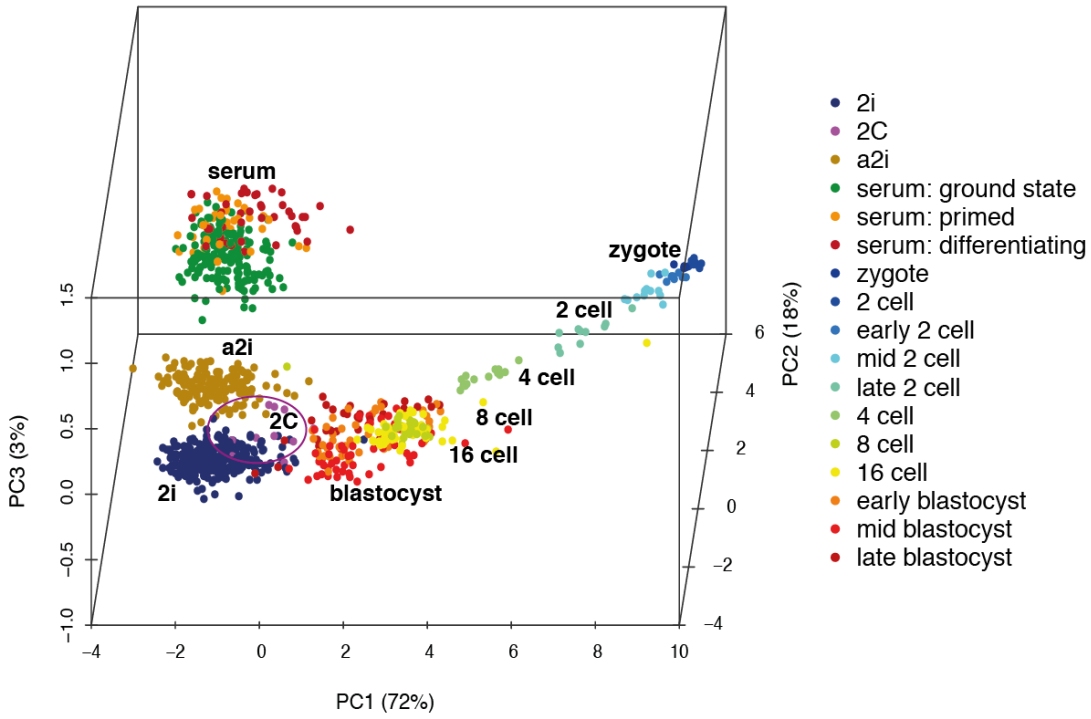


Figure 4.7 Comparison of mESC to cells from early embryo development
PCA loading plot of the Spearman's rank correlation coefficients from mESCs and single cells of mouse preimplantation embryos (Deng et al., 2014), showing the mapping of mESCs in mouse development stages. The cells are visualized by loadings of the first three principal components of the Spearman's rank correlation matrix between cells, where I used the same expression cut-off as that employed by Deng *et al*.

4.5 Conclusions

In 2i I observed a subpopulation, 2C-like cells, which also contribute to the noisiness of the 2i population. Notably, I could not identify 2C-like cells in serum and a2i, which is most likely because in these conditions they are present in frequencies significantly lower than 1% and were not sampled. As they are similar to the bulk of 2i cells and rare, their contribution to the global heterogeneity of 2i cells is much smaller than the three distinct subpopulations

in serum. My results show that, globally, 2C-like cells are not particularly similar to cells at the 2-cell stage of the embryo, as was suggested previously. Nevertheless, MacFarlan and colleagues showed that 2C-like cells when transferred into embryos contribute to both embryo and extraembryonic tissues, which means that they have more potency (Macfarlan et al., 2012).

2C-like cells found in 2i, in addition to standard gene expression pattern of 2i cells, express genes that are related to endogenous retrovirus MuERV-L expression that are expressed also at 2-cell stage of the embryo. It was suggested that repression of LTR acting as promoters is regulated by epigenetic silencing involving KDM1A, G9A, KAP1 and HDACs (Macfarlan et al., 2012). Emergence of 2C-like cells can contribute to the fact that cells' epigenetic states fluctuate in the artificial environment of cell culture.

4.6 Further Research

The biological significance of 2C-like cells is debatable and it is not obvious whether these cells do indeed have an ability to produce all extraembryonic tissues being derived from the inner cell mass of the blastocyst.

I think that they are not a good model for 2 cell stage of the embryo, but they can be used for studying function and mechanism of endogenous retrovirus, MuERV-L. During early embryo development, endogenous retroviruses are transiently derepressed and they insert into new positions in the genome, increasing genomic variability (Maksakova et al., 2006; Moyes et al., 2007; Wang et al., 2010). The reasons are that obtaining big numbers of 2C-like cells is quite easy in comparison to obtaining the same number of 2-cell embryos. Additionally studying 2-cell stage embryos is difficult, because it is a very dynamic stage of development. Deconvolution of different

developmental processes and simultaneously happening processes involved in endogenous retrovirus expression and functions would be very challenging. In 2i cultured mouse embryonic stem cells, which are homogeneous in expression of pluripotency genes it would be easier to focus solely on the expression changes that accompany activation and deactivation of MuERV-L.

Chapter 5

Transcriptomic gene regulatory network of pluripotency

5.1 Introduction

In chapters 3 and 4, I mined a set of high-throughput single cell RNA-sequencing data to explore correlations between cells, but these data also provide a rich resource for analysing correlations in gene expression. Gene-gene correlations can imply common regulatory mechanisms and functions of genes. I aimed to use this to develop new hypotheses about the transcriptional regulatory network that regulates pluripotency in mESCs, which is known to be highly interconnected and complex (Boyer et al., 2005; Kim et al., 2008; Loh et al., 2006).

Genes and their products that regulate cellular functions are organized in gene regulatory networks (Hasty et al., 2001; Hecker et al., 2009; Karlebach and Shamir, 2008). Members of the network interact with each other to fulfil particular functions, and these networks are particularly important in the response to external stimuli and during processes such as development and

differentiation. If one gene product positively regulates other genes in a network, then an increase in the number of molecules of this product will cause an increase in expression of its target genes (Bowsher and Swain, 2012). I can observe such relationships by measuring the correlation of expression between two genes. In this case I assume that the level of mRNA and the level of protein for which it codes, correlate in a cell (Liu et al., 2016). This is true for most cases, however for data interpretation it is important to keep in mind that the presence of mRNA does not imply it being translated (Peshkin et al., 2015). Correlated expression implies that two genes are within the same regulatory module, but it does not elucidate the relationship between these genes. A gene pair with a high correlation coefficient may encode a transcription factor and its target, but directionality of this interaction cannot be inferred solely from these data. It is also not possible to infer whether interactions reflect direct causation or where two genes with correlated expression are two downstream targets regulated by the same factor.

The pluripotency regulatory network has been intensively studied since the development of mouse embryonic stem cell cultures over 30 years ago, but our understanding of it remains incomplete (Boyer et al., 2005). External signals, such as LIF, activate STAT3, and BMP4, which in turn activate expression of *Id* (inhibition of differentiation) genes to promote pluripotency (Cartwright et al., 2005; Hall et al., 2009; Matsuda et al., 1999; Ying et al., 2003a). Several key transcription factors were also identified, most well described are OCT4, NANOG and SOX2 (Avilion et al., 2003; Chew et al., 2005; Orkin et al., 2008; Rodda et al., 2005; Sharov et al., 2008). ChIP-chip and ChIP-seq data showed, that these and other key pluripotency genes co-occupy promoters of many genes, making it difficult to disentangle the wiring of the network (Adachi et

al., 2013; Loh et al., 2006). Key pluripotency genes are also found at the promoters of each other suggesting that there is a complex network rather than a simple hierarchical structure (Kim et al., 2008; Ng and Surani, 2011; Xu et al., 2014).

In this chapter I aim to use single cell mRNA sequencing to investigate the gene regulatory networks involved in pluripotency and to potentially identify new factors that play a role in pluripotency maintenance.

5.2 Pluripotency gene regulatory network

To investigate gene regulatory networks I decided to look at the transcription factors, which regulate gene expression, and hence are key genes in shaping the gene expression network.

Focusing on transcription factors made this analysis more tractable, since such analysis for 48,034 genes (ENSEMBL annotation GRCm38.p4) is computationally intensive and requires additional filtering of pseudogenes and genes that arose from duplication and to which sequencing reads map ambiguously. Furthermore, transcription factors are the key genes that orchestrate the transcriptional response and changes in their expression are crucial in transcriptional control. To obtain a comprehensive list of transcription factors and chromatin modifiers I took genes from the gene ontology category 'DNA binding' from the GO database embedded at Ensembl Biomart (http://www.ensembl.org/biomart) and calculated the Spearman rank correlation coefficients for all gene-to-gene comparisons using data from serum cultured cells. To perform such gene network analyses one needs to have a perturbed system, meaning the population of cells cannot be homogeneous. Cells have to undergo an unsynchronized response to a

stimulus or traverse between developmental stages. This is the case in serum cultures, which I showed in Chapter 3 to be more heterogeneous.

Lowly expressed genes and genes which have stable expression do not correlate with genes that change expression as a response to external stimulus and so are not informative for the construction of gene regulatory networks. I aimed to select genes that correlate with other genes at least to some level. I tested different levels of Spearman Rank Correlation Coefficient thresholds and empirically found that for this case a threshold of at least below -0.35 or above 0.35 is sufficient to filter non-correlated genes and leave enough genes for further analysis (Figure 5.1).

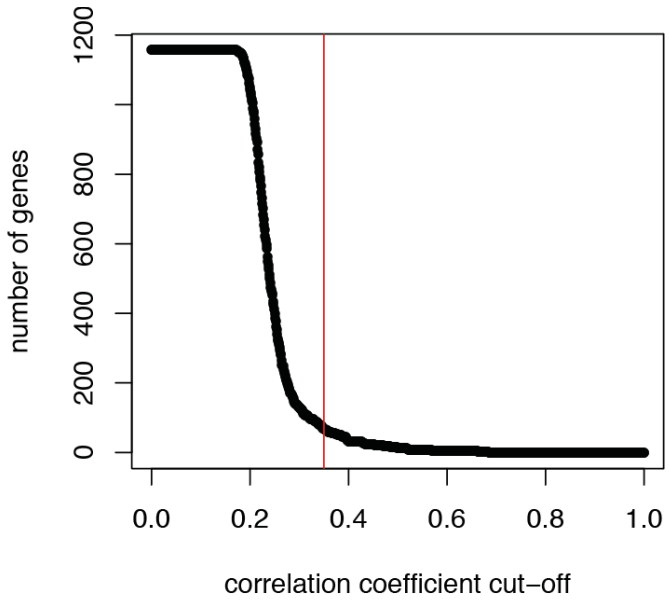


Figure 5.1 Correlation coefficient cut-off.
Plot shows the number of genes that correlate with at least one other gene above Spearman rank correlation coefficient value.

Finally, I plotted the correlations between the remaining genes as a heatmap, which revealed two clusters (Figure 5.2).

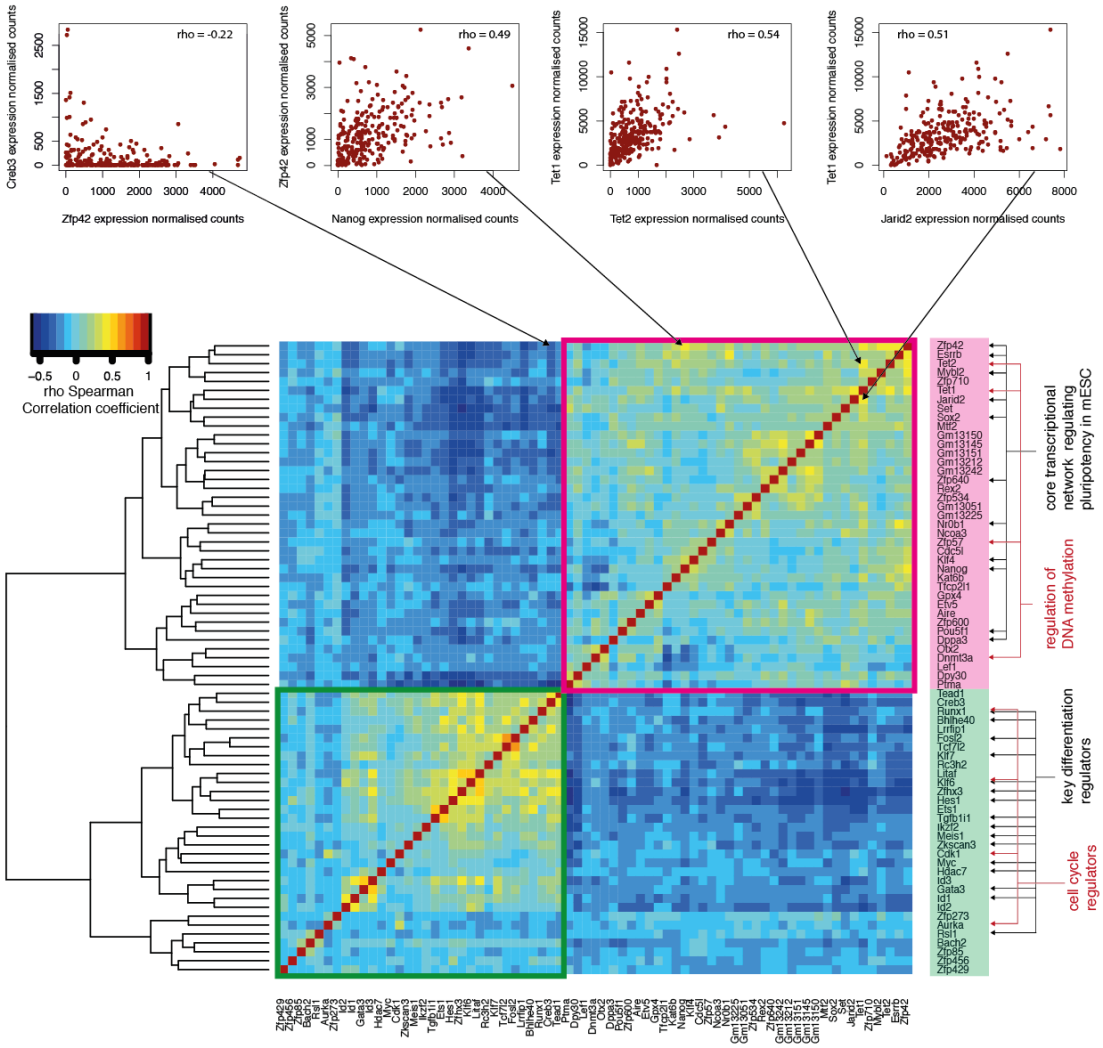


Figure 5.2 Spearman correlation matrix of transcription factors and key pluripotency genes.

The heatmap shows the correlation coefficients between a set of transcription factors and other key genes involved in pluripotency. Above are examples of genes with expression patterns that correlate positively and negatively (from the left *Zfp42* and *Creb3*, *Zfp42* and *Nanog*, *Tet1* and *Tet2*, *Tet1* and *Jarid2*).

I found that in serum cultured cells, *Nanog* expression correlates with other pluripotency factors and key regulatory genes. The Nanog-correlated genes include transcription factors (*Esrrb*, *Klf4*, *Oct4*/*Pou5f1*, *Sox2* and *Zfp42*), genes involved in DNA methylation (*Dnmt3a*, *Tet1*, *Tet2*), and other genes such as nuclear receptor *Nr0b1* and histone lysine acetyltransferase *Kat6b*.

Interestingly, Nanog expression is negatively correlated with differentiation regulators including transcription factors *Gata3* and *Klf7*. These findings agree with known interactions in the pluripotency regulatory network, where *Nanog* regulates *Esrrb* (Boyer et al., 2005), *Zfp42* (Shi et al., 2006), and *Klf4* (Zhang et al., 2010).

Beyond confirming known interacting genes, I identified correlations between characterized pluripotency genes and candidate new components of the pluripotency transcriptional regulatory network.

I found that genes such as *Ptma*, which was previously implicated in immune response modulation (Pineiro et al., 2000), oncogene *Set*, which regulates the cell cycle and is involved in chromatin remodelling (Seo et al., 2001), prostate cancer associated gene Etv5 (Helgeson et al., 2008) several zinc finger proteins of unknown functions: ZFP534, ZFP600, ZFP640, ZFP710 and other unknown genes, such as *Gm13145*, *Gm13150*, *Gm131451*, *Gm13212*, *Gm13242*, *Gm13051*, *Gm13225*. Interestingly genes from the last group and *Zfp600* are clustered in the genome on chromosome 4 within one roughly 1.9 Mb region. In this region there are predicted lncRNAs on the reverse strand (*Gm26573*, *Gm26624*, *C230088H06Rik*) spanning several genes. Single cell mRNA sequencing does not provide strand data information and it is possible that the correlation between these genes is because I detect lncRNAs from the opposite strand and the correlation is simply because it is one molecule.

5.3 Validation of putative pluripotency genes using CRISPRi transcriptional silencing

Of the novel genes that displayed highly correlated expression profiles with known pluripotency factors I selected 7 genes for validation: *Ptma*, *Zfp640*,

Zfp710, *Dpy30*, *Set*, *Etv5*, *Kat6b*. First, I mined ChIP-seq and Chip-chip data from the ESCAPE database (Xu et al., 2013) to check if there are potential interactions between these genes and the pluripotency network. This database provides a list of interactions between promoters and transcription factors and I found that the promoters of 6 out of the 7 candidate genes are bound by at least one of the core pluripotency genes (Figure 5.3).

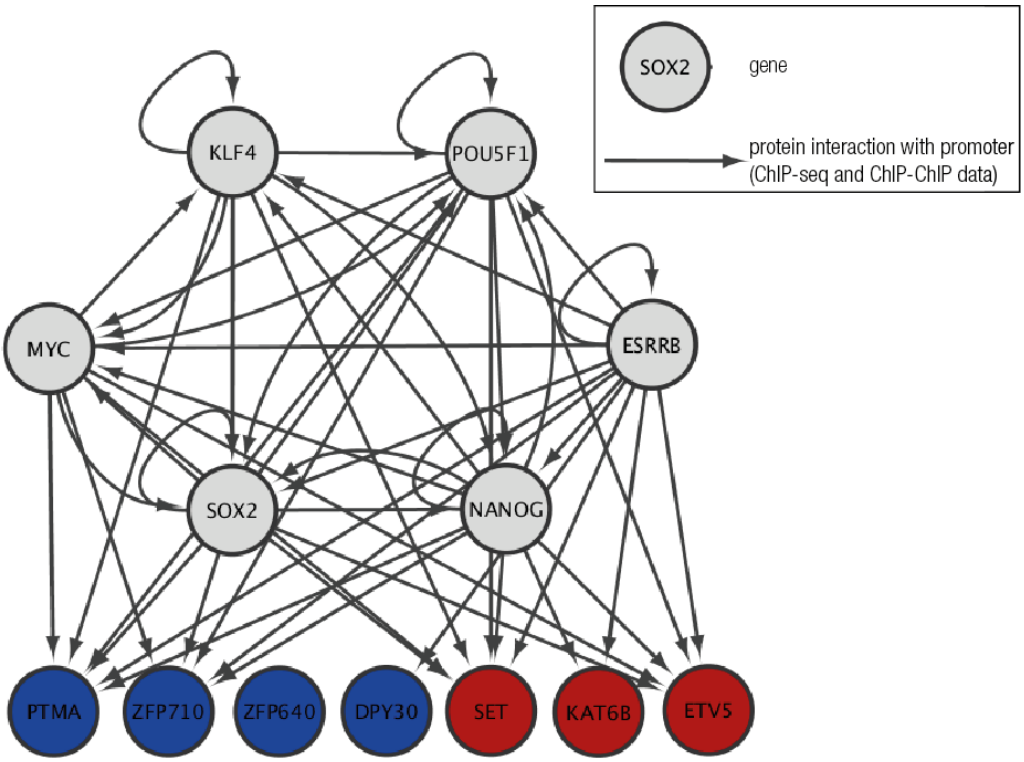


Figure 5.3 Pluripotency network
Network showing known interactions of core pluripotency factors with the novel candidate genes. Data obtained from ChIP-seq and ChIP-ChIP experiments from ESCAPE database.

To provide insight into the functional role of these genes, I attempted to downregulate their expression using CRISPR/dCas9 repressor targeting of their promoters (Gao et al., 2014).

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) is a prokaryotic immune system that was very successfully applied in eukaryotic cells to knock out genes (Doudna and Charpentier, 2014; Jinek et al., 2012). It uses guide RNA (gRNA), which consists of a short RNA matching the sequence of the gene of interest and a tracer, which binds to the Cas9 endonuclease that subsequently cleaves the DNA. Importantly this way one can target any 20nt long sequence provided its 3' end has a so called Protospacer Adjacent Motif (PAM) sequence, which is TGG for Cas9. Cleaved target DNA is then efficiently repaired by the Non-Homologous End Joining pathway, which is very error prone and introduces insertions and deletions that can cause frameshifts. In some cases the repair can also go through the Homology Directed Repair pathway, which is high fidelity and does not result in sequence mutations (Cong et al. 2013; Makarova et al., 2011).

Based on this system, CRISPR interference was established (Larson et al., 2013). The endonuclease Cas9 was mutated at the active site of its nuclease domain to remove its ability to cut DNA. Additionally, the catalytically inactive Cas9 was fused to the transcriptional repressor, Krüppel associated box (KRAB) domain. In this approach one uses gRNA to target dCas9-KRAB to the promoter or enhancer of a gene of interest and the interaction of the KRAB domain with the DNA causes a decrease in the level of transcription of this gene (Gao et al., 2014; Gilbert et al., 2014; Gilbert et al., 2013).

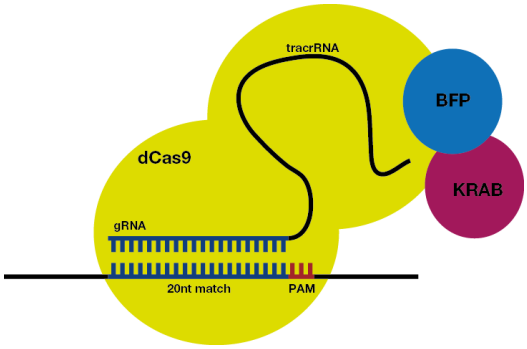


Figure 5.4 Schematic of CRISPRi

I cloned gRNA targeting promoters of 7 selected genes (for more details please refer to chapter 2). Subsequently, Dr Xuefei Gao co-transfected mESCs with gRNA-mCherry and dCas9-BFP plasmids and double positive cells were purified by flow cytometry in the facility at the Sanger Institute. For each downregulated gene three biological replicates were made. Subsequently, I examined the transcriptomes of populations of transfected cells by bulk mRNA sequencing. On average I sequenced over 10 million reads per sample and 48% of reads maps to exons (Figure 5.5). In standard bulk RNA sequencing of mESCs I observed that about 80% of reads map to the exons (Figure 3.3). Lower than usual percentage of reads mapping to exons is a result of the fact that libraries for these samples were prepared from only 10,000 cells each using SmartSeq2 protocol, which involves a cDNA amplification step.

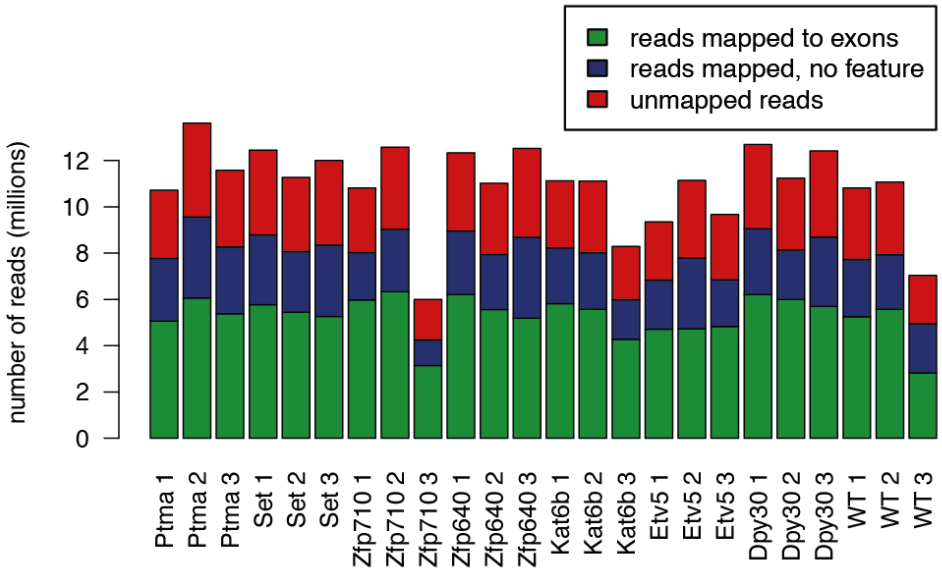


Figure 5.5 Mapping statistics
Barplot shows how many reads map to exons, mouse genome and how many do not map for all samples in three replicates.

For four out of the seven samples there was significant repression of the targeted gene, and I narrowed down our focus to these four genes (Figure 5.6). To achieve successful downregulation of gene expression it is important to target the right position of the promoter, but unfortunately this position cannot be predicted in advance. It is particularly difficult to target genes that have multiple alternative transcription start sites, as inhibiting one may lead to more expression from the alternative. Additionally, CRISPR technology limited me to positions that have PAM sequences immediately upstream. In cases where repression gives only subtle results it may not be significant due to the fact that I only have three samples per condition, so statistical tests have low power.

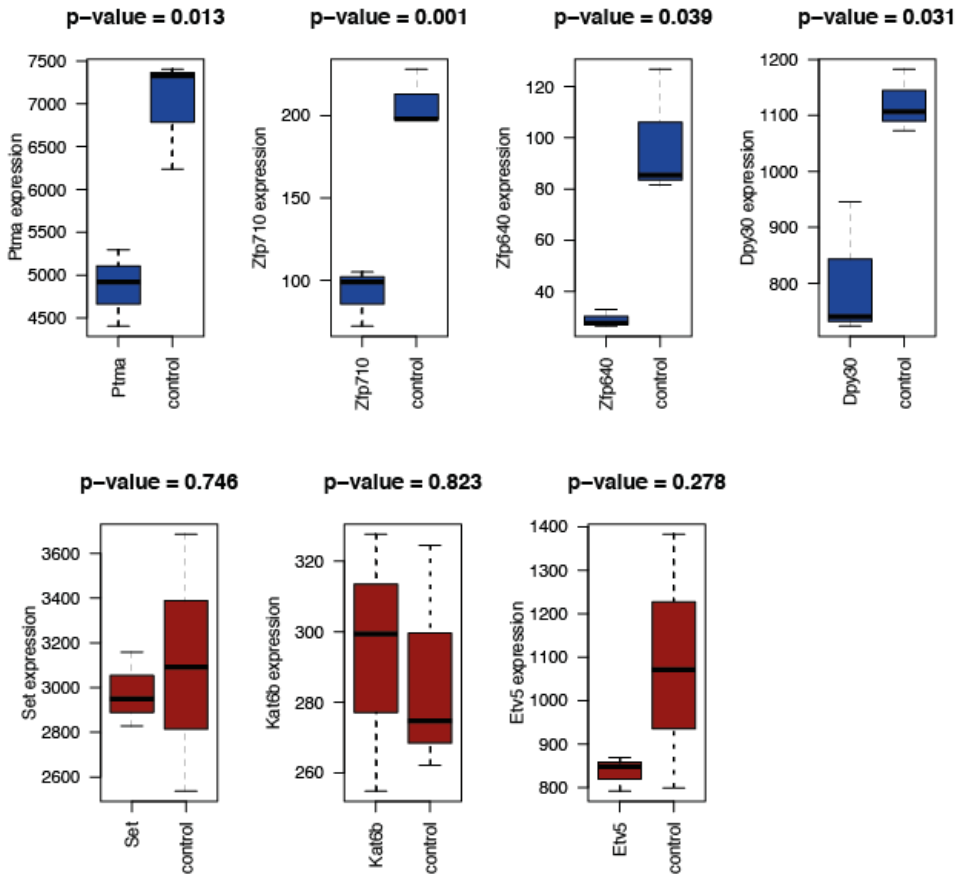


Figure 5.6 CRISPRi results

Boxplots show the expression level of repressed genes in samples and control. Targets for which we achieved significant repression are in blue. Gene expression levels are shown as DESeq size factor normalise counts.

I performed differential expression analysis between samples transfected with a control gRNA that does not have a target mouse genome, but instead targets the human Rosa26 locus and the gRNA targeting the gene of interest using DESeq. After multiple hypothesis testing correction I found significantly differentially expressed (p-value <0.05) genes in two cases: *Ptma* and *Zfp640* (Figure 5.7). There were 16 differentially expressed genes in the *Ptma* knock-down and 7 in the *Zfp640* knock-down.

Three significantly upregulated genes in the *Ptma* knock-down are all involved in pluripotency and early embryonic development. Extracellular

matrix proteoglycan versican (VCAN) is an important mediator of endothelial-mesenchymal transition (EMT) during embryoid body differentiation from mESCs (Shukla et al., 2010; Wight, 2002). Adhesion junction plaque protein dystonin (DST) was shown to be transiently upregulated upon LIF withdrawal (Trouillas et al., 2009) and retrotransposon-derived protein PEG10 is essential for early embryonic development (Ono et al., 2006).

Among the downregulated genes most interestingly I found a key pluripotency regulator *Fgf4* (Kunath et al., 2007; Tanaka et al., 1998). Additionally downregulated genes included poly (ADP-ribose) polymerase 12 (*Parp12*) implicated in protein translation control and NF-κB signalling (Welsby et al., 2014); early growth response protein 1 (*Egr1*), a zinc-finger transcription factor that regulates cell apoptosis *via* the p53 pathway (Baron et al., 2006; Thiel and Cibelli, 2002); NAD(P)H dehydrogenase 1 (*Nqo1*), whose main metabolic function is reduction of quinones to hydroquinones, and also regulates the ubiquitin-independent p53 degradation pathway (Asher et al., 2001; Ross and Siegel, 2004); and secreted frizzled related protein 1 (*Sfrp1*) a key player in the WNT pathway and a positive regulator of differentiation to the neuronal lineage in human mESCs (Schwartz et al., 2012). Several cancer-related genes were also downregulated. Those include cleft lip and palate transmembrane protein 1-like protein (*Clptm1l*), which is overexpressed in lung cancer and has antiapoptotic activity mediated *via* PI3K/Akt survival signalling (James et al., 2014). Additional cancer-related genes were protein tyrosine phosphatase type IVA 3 (*Ptp4a3*) and proteasome activator complex subunit 3 (*Psme3*) associated with melanoma and colon cancer respectively (Laurent et al., 2011; Roessler et al., 2006). Finally, uracil-DNA glycosylase (*Ung*) that acts to prevent mutagenesis by base-excision repair (BER) pathway,

but was also shown to promote DNA demethylation (Savva et al., 1995; Xue et al., 2016); mitochondrial import receptor subunit TOM70 (*Tomm70a*); translation initiation factor eIF-2B subunit epsilon (EIF2B5) and 14–3-3 protein, YWHAH coding genes were also downregulated when *Ptma* was downregulated.

Downregulation of *Zfp640* similarly to downregulation of Ptma caused upregulation of *Peg10* and downregulation of *Egr1*. In addition I also observed upregulation of pluripotency associated gene *Dppa3* (Bowles et al., 2003; Waghray et al., 2015) and downregulation of three genes of unknown function: *Gm17404*, *Gm10323*, *2410141K09Rik*.

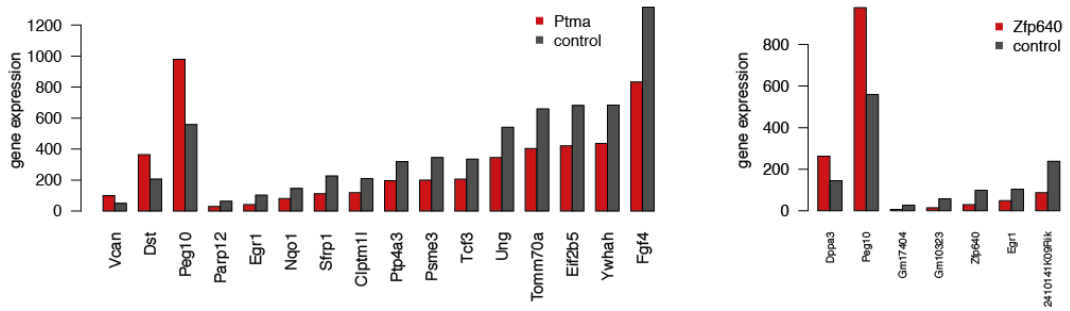
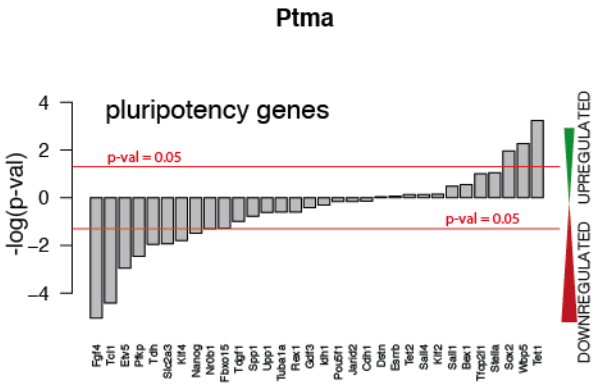


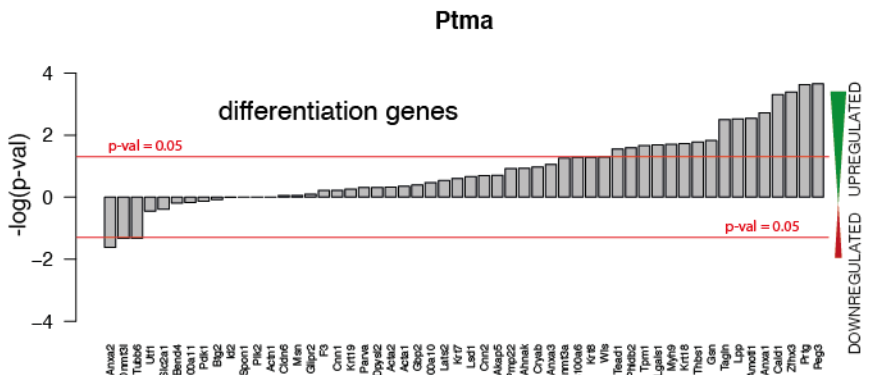
Figure 5.7 Differentially expressed genes in *Ptma* and *Zfp640* downregulated samples

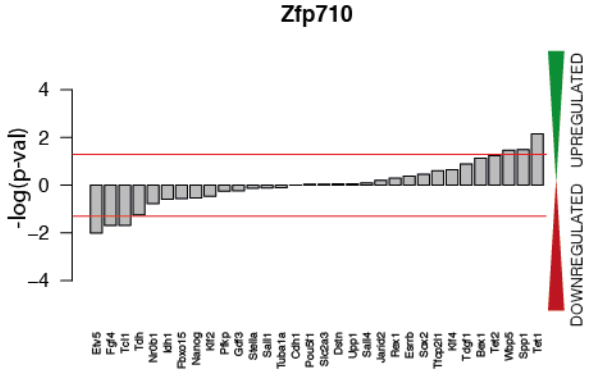
Barplot of gene expression levels of significantly differentially expressed genes in *Ptma* and *Zfp640* repressed samples (DESeq, multiple hypotheses testing adjusted p-value < 0.05).

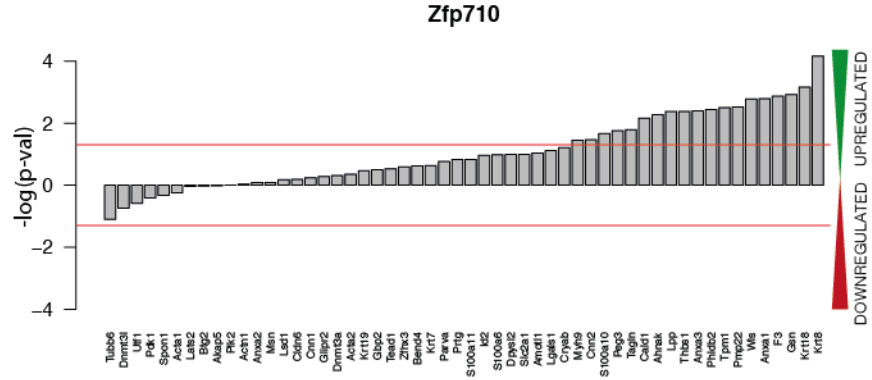
Due to having only three replicates per condition and the relatively low quality of sequencing data I was able to detect only a few significantly differentially expressed genes. To observe if there is a trend for change in expression of major pluripotency and differentiation factors I plotted *p*-values obtained for comparison of the expression of this gene in the knockdown and

control using DESeq (Figure 5.8). In the samples with repressed *Ptma*, I observed a trend of decreased expression of pluripotency genes, and increased expression of genes associated with differentiation (pluripotency and differentiation genes are as in Figure 3.9). *Zfp710* and *Zfp640* show a similar but milder phenotype; while for *Dpy30* there is no clear change in the expression of pluripotency genes. The lack of effect of *Dpy30* downregulation on the pluripotency gene expression is consistent with a previous report (Jiang et al., 2011). Overall, these results suggest that *Ptma* and *Zfp640*, and potentially also *Zfp710*, are new candidate genes involved in regulating the exit from pluripotency.


Ptma
pluripotency genes
-log(p-val)
p-val = 0.05
p-val = 0.05
UPREGULATED
DOWNREGULATED


Ptma
differentiation genes
-log(p-val)
p-val = 0.05
p-val = 0.05
UPREGULATED
DOWNREGULATED


Zfp710
-log(p-val)
UPREGULATED
DOWNREGULATED


Zfp710
-log(p-val)
UPREGULATED
DOWNREGULATED

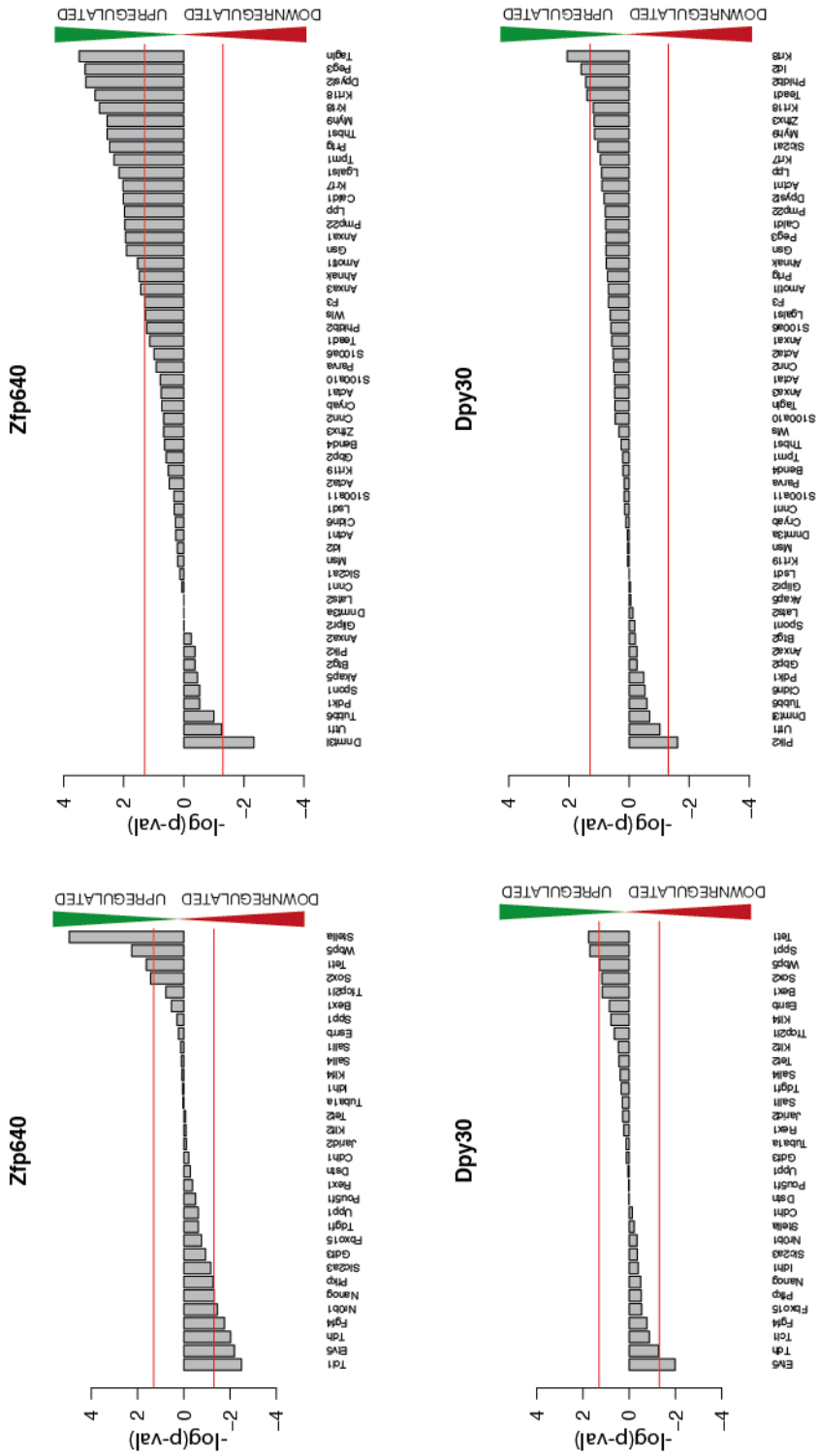


Figure 5.8 Significance of pluripotency and differentiation genes expression changes in knock down samples.

Barplots showing the logarithm of p-values for differential expression from DESeq of pluripotency (left) and differentiation (right) genes in the knock down samples. For genes that are downregulated, the numbers are negative, and positive for upregulated genes. The red line indicates p-value threshold of 0.05.

5.4 Conclusions

My data and methodology allowed me to find new genes involved in the pluripotency network, which I validated using CRISPR repression (Gilbert et al., 2014). I found that downregulating *Zfp640*, *Zfp710* and *Ptma* affected the expression of both pluripotency and differentiation genes. *Ptma* repression resulted in the strongest deviation from control samples, and I infer that these cells deviate from pluripotency towards a differentiated state.

Interestingly, *Ptma* is a well-known gene encoding prothymosin alpha, precursor of thymosin alpha. It is mostly described in the context of immunology, as thymosin alpha protein was first extracted from thymus and were subsequently shown to modulate the immune response. It is used as a drug (Thymalfasin) in treatment of chronic hepatitis B and C and is used as an adjuvant in therapy for some types of cancer (Ciancio and Rizzetto, 2010; Garaci et al., 2012; Ioannou et al., 2012). Biochemically prothymosin alpha is unique, as it is extremely basic especially the fragment that is cleaved off to form thymosin alpha. This suggests it is not binding DNA directly. The mode of action of *Ptma* has been studied in cancer and immune cells, and it has been shown to play a role in proliferation through mechanisms involving chromatin remodelling and interaction with numerous pathways associated with pluripotency maintenance such as the JAK-STAT pathway, the PI3K/AKT pathway, and the NF-κB pathway, but its exact molecular mechanism is unknown (George and Brown, 2010; Guo et al., 2015; Romani et al., 2012; Yang et al., 2004). Functions of *Zfp640* and *Zfp710* are not described in the literature.

5.5 Future research

Further experiments should be performed to understand the function of *Ptma*, *Zfp640* and *Zfp710* in the pluripotency network. Understanding how mechanistically these genes are involved in pluripotency maintenance would provide additional strong evidence for involvement of these genes in the process and would shed new light on how pluripotency and exit to differentiation are regulated. Unfortunately, that was not possible within this project timeline.

For finding downstream targets, ChIP-seq would elucidate which promoters are bound by ZFP640 and ZFP710. There is an antibody for ZFP710 available to purchase, but antibodies for ZFP640 would have to be generated and both have to be tested.

It is unclear how PTMA interacts with DNA. It is highly acidic and thus if it binds to the DNA it is likely to be *via* interaction with other more basic proteins. ChIP-seq of PTMA and comparison to known data in addition to finding downstream targets may reveal which proteins it often co-localizes with, suggesting potential interactions.

Previously pull-down experiments were performed using PTMA which identified histones as its interacting partners (Díaz-Jullien et al., 1996). It is possible however, that this is an artefact, because positively charged and abundant histones may associate non-specifically with PTMA when cells are lysed and chromatin is disrupted. Another paper suggested interaction of PTMA with oestrogen receptor (Garnier et al., 1997, Martini et al., 2000). It is important to perform pull-down experiments without disrupting chromatin to avoid potential sticking of histones to the protein.

Furthermore, single cell mRNA sequencing of cells with different levels of *Ptma*, *Zfp640* and *Zfp710* downregulation is likely to yield further information about the transcriptional network of these target genes pointing to their function within these cells.

Chapter 6

Concluding Remarks

It is remarkable that a whole complex organism with myriad different cell types and tissues develops from one single zygote. Embryonic stem cells are derived from pluripotent cells within the inner cell mass of the embryo and they have the capacity to differentiate into all tissue types of the organism as well as being able to contribute to chimeric embryos. This creates a promising avenue for the field of regenerative medicine. Understanding the molecular mechanisms of pluripotency and the exit into differentiation is key for designing protocols to grow tissues in culture.

Depending on cell culture condition, mouse embryonic stem cells have different transcriptomes and the population has a different structure. In my thesis, using single cell mRNA sequencing I dissected the heterogeneity of the population of mouse embryonic stem cells cultured in three cell culture conditions. Comparison with previous studies allowed me to generate a comprehensive picture of the gene expression variability. I confirmed that

genes previously suggested to be heterogeneous or fluctuating are indeed doing so. In serum, pluripotency and differentiation genes fluctuate as two modules. In cells with low expression of the pluripotency module, the differentiation module is high and *vice versa*. This corresponds to functional differences between cells, where some of them are more pluripotent and some already express the differentiation programme.

Cells cultured in 2i medium that mediates a ground state of pluripotency are homogeneous for expression of the pluripotency module and do not express markers of differentiation. On the other hand, cell cycle gene expression is heterogeneous in 2i. I was able to use this heterogeneity to assign cells to cell cycle stages: G1/S or G2/M. Using the data presented in this dissertation and other previously published data (Tsang et al., 2015) I observed that there is a relationship between cell cycle heterogeneity and the length of the cell cycle. Cells that cycle quickly have homogeneous expression of cell cycle genes. In cells that cycle with moderate speed, such as those cultured in 2i one can discriminate G1/S form G2/M cells. In slowly cycling cells, such as HSC all phases of the cell cycle can be identified and even G1 can be divided into early and late (Tsang et al., 2015).

I speculate that there are two reasons for low cell cycle noise in fast cycling cells. Firstly, in very quickly cycling cells G1 phase is virtually non-existent causing lower heterogeneity. Secondly, the degradation half-lives of cell cycle related genes are 6-8 hours (Sharova et al., 2009). If the cell cycle is very quick there is not enough time for mRNAs from one phase to degrade when the cell enters the next phase of the cell cycle. This leads to mRNAs from one phase to "bleed" into the following phase. Biologically, this does not necessarily have much effect on cell cycle regulation, because this is achieved at the level of

protein signalling, mostly post-translational phosphorylation by CDKs and protein degradation, mostly of cyclins.

This relationship can be exploited to estimate the speed of the cell cycle in heterogeneous populations. For example, cells from complex tumour tissues can be profiled using single cell mRNA sequencing and subpopulations can be identified. Subsequently, the relative proliferation rates of the subpopulations can be measured using heterogeneity of expression of cell cycle genes in each of the subpopulations. It is quite remarkable that a dynamic feature of a system can be measured from snapshot data such as single cell mRNA sequencing of one time point.

Cells cultured in alternative 2i are similar transcriptomically to cells cultured in 2i, especially for expression of pluripotency genes, suggesting that inhibition of SRC gives rise to a similar phenotype as inhibition of MEK1/2.

Furthermore, I identified a population of previously-reported so-called "2C-like cells" (Macfarlan et al., 2012) in 2i medium and looked at their transcriptomes in relation to transcriptomes of cells from subsequent stages of embryo development. These cells are substantially more similar to cells from the blastocyst than cells from the embryo at the 2 cell stage. The transcriptomes of 2C-like cells are similar to those of the other cells in 2i culture, but in addition to the transcriptomic profile of 2i cells they express some additional genes. 2C-like cells arise probably due to chromatin changes that are forced by signals from the media the cells were cultured in. Derepression of endogenous retroviral elements causes expression of genes that are regulated by MuERV-L in addition to the transcriptomic profile of 2i cells. This is a useful observation, as it allows decoupling regulation of gene

expression by MuERV-L from changes that occur in 2 cell stage embryos, and the study of this process in stead-state culture.

Finally, I discovered several potential regulators of pluripotency and validated that three genes, namely *Ptma*, *Zpf640* and *Zfp710* are regulators of pluripotency. The approach I used can be used for any biological system, for understanding genes that change in transitions or as a result of response to stimulus.

In my work, in addition to gaining biological and mechanistic insights into the pluripotency of mouse embryonic stem cells, I have shown how and what information can be harvested from the single cell transcriptomic data. I measured and understood sources of heterogeneity, found and characterized a rare cell population, assigned cell cycle stage to cells and identified new players important for gene expression networks. These approaches will prove useful for analysis of any type of data in the future.

Abbreviations

AID	activation-induced cytidine deaminase pathway
ANOVA	analysis of variance
ATP	adenosine-5'-triphosphate
BER	base-excision repair
CAP	circular a posteriori projection
cDNA	complementary DNA
CEL-Seq	single-cell RNA-Seq by multiplexed linear amplification
CHD	chromodomain-helicase-DNA-binding protein
ChIP	chromatin immunoprecipitation
circRNA	circular RNA
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPRi	CRISPR interference
CV	coefficient of variation
DM	distance to the median
DMEM	Dulbecco modified Eagle's minimal essential medium
EC	embryonic carcinoma
EDTA	ethylenediaminetetraacetic acid
EMT	endothelial to mesenchymal transition
EpiSC	epiblast stem cells
ERCC	external RNA controls consortium
ESCAPE	embryonic stem cell atlas from pluripotency evidence
FACS	fluorescence-activated cell sorting
FDR	false discovery rate

FISH	fluorescence in situ hybridization
FISSEQ	fluorescent in situ sequencing
FWER	family-wise error rate
GDP	guanosine-5'-diphosphate
GFP	green fluorescent protein
GO	gene ontology
gRNA	guide RNA
GSNAP	genomic short-read nucleotide alignment program
GTF	general transfer format
GTP	guanosine-5'-triphosphate
HDAC	histone deacetylase
hESC	human embryonic stem cells
HSC	hematopoietic stem cell
ICA	independent component analysis
IFC	integrated fluidic circuit
iPSC	induced-pluripotent stem cells
IVT	in vitro transcription
JAK	janus-associated kinase
KRAB	Krüppel associated box
KS test	Kolmogorov–Smirnov test
LB	Luria-Bertani broth
LCM	laser capture microdissection
LIF	leukaemia inhibitory factor
lncRNA	long non-coding RNA
LTR	long terminal repeat
MAPK	mitogen-activated protein kinase
MAP2K	mitogen-activated protein kinase kinase
MAP3K	mitogen-activated protein kinase kinase kinase
MARS-Seq	massively parallel single-cell RNA-sequencing
MEF	mouse embryonic fibroblast
mESCs	mouse embryonic stem cells
mRNA	messenger RNA
MST	minimal spanning tree
MuERV-L	murine endogenous retrovirus L
NFAT	nuclear factor of activated T-cells

NPC	neuronal progenitor cell
NURD	nucleosome remodelling and histone deacetylase complex
NURF	nucleosome remodelling factor
PAM	protospacer Adjacent Motif
PC	principal component
PCA	principal component analysis
PLA	proximity ligation assay
PI3K	phosphoinositide 3-kinase
PRC1	polycomb-group repressive complex 1
PRC2	polycomb-group repressive complex 2
qPCR	quantitative real-time polymerase chain reaction
RNA	ribonucleic acid
RPM	reads per million
rRNA	ribosomal RNA
SC3-seq	single-cell mRNA 3-prime end sequencing
scRNA-seq	single cell RNA sequencing
SCUBA	single-cell clustering using bifurcation analysis
SH2	src homology 2 domain
SOM	self-organizing map
SNN	shared nearest neighbour
SRF	serum response factor
SRY-box	sex determining region Y box
STAT	signal transducers and activators of transcription
STO	Sandos Inbred Mice Thioguanine/Ouabain-resistant mouse fibroblast cell line
STRT-Seq	single-cell tagged reverse transcription sequencing
TIVA	transcriptome in vivo analysis
tSNE	t-distributed stochastic neighbour embedding
UMI	unique molecular identifier
WGCNA	weighted gene co-expression network analysis
ZGA	zygote genome activation
ZIFA	zero inflated factor analysis

Bibliography

Adachi, K., Nikaido, I., Ohta, H., Ohtsuka, S., Ura, H., Kadota, M., Wakayama, T., Ueda, H.R., and Niwa, H. (2013). Context-dependent wiring of Sox2 regulatory networks for self-renewal of embryonic and trophoblast stem cells. Molecular Cell *52*, 380-392.

Adamo, A., Sese, B., Boue, S., Castano, J., Paramonov, I., Barrero, M.J., and Izpisua Belmonte, J.C. (2011). LSD1 regulates the balance between self-renewal and differentiation in human embryonic stem cells. Nature Cell Biology *13*, 652-659.

Aguilo, F., Zhang, F., Sancho, A., Fidalgo, M., Di Cecilia, S., Vashisht, A., Lee, D.F., Chen, C.H., Rengasamy, M., Andino, B., *et al.* (2015). Coordination of m(6)A mRNA Methylation and Gene Transcription by ZFP217 Regulates Pluripotency and Reprogramming. Cell Stem Cell *17*, 689-704.

Ahmad, A., Robinson, A.R., Duensing, A., van Drunen, E., Beverloo, H.B., Weisberg, D.B., Hasty, P., Hoeijmakers, J.H., and Niedernhofer, L.J. (2008). ERCC1-XPF endonuclease facilitates DNA double-strand break repair. Molecular and Cellular Biology *28*, 5082-5092.

Al-Minawi, A.Z., Saleh-Gohari, N., and Helleday, T. (2008). The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells. Nucleic Acids Research *36*, 1-9.

Amir, e.A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology *31*, 545-552.

Amit, M., Carpenter, M.K., Inokuma, M.S., Chiu, C.P., Harris, C.P., Waknitz, M.A., Itskovitz-Eldor, J., and Thomson, J.A. (2000). Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Developmental Biology *227*, 271-278.

Andang, M., Moliner, A., Doege, C.A., Ibanez, C.F., and Ernfors, P. (2008). Optimized mouse ES cell culture system by suspension growth in a fully defined medium. Nature Protocols *3*, 1013-1017.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq – A Python framework to work with high-throughput sequencing data. bioRxiv preprint.

Ang, Y.S., Tsai, S.Y., Lee, D.F., Monk, J., Su, J., Ratnakumar, K., Ding, J., Ge, Y., Darr, H., Chang, B., *et al.* (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell *145*, 183-197.

Angerer, P., Haghverdi, L., Buttner, M., Theis, F.J., Marr, C., and Buettner, F. (2015). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32, 1241-1243.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S.A., Ponting, C.P., Voet, T., *et al.* (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature Methods *13*, 229-232.

Anneren, C., Cowan, C.A., and Melton, D.A. (2004). The Src family of tyrosine kinases is important for embryonic stem cell self-renewal. The Journal of Biological Chemistry *279*, 31590-31598.

Antebi, Y.E., Reich-Zeliger, S., Hart, Y., Mayo, A., Eizenberg, I., Rimer, J., Putheti, P., Pe'er, D., and Friedman, N. (2013). Mapping differentiation under mixed culture conditions reveals a tunable continuum of T cell fates. PLoS Biology *11*, e1001616.

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M., *et al.* (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nature Methods *6*, 647-649.

Asher, G., Lotem, J., Cohen, B., Sachs, L., and Shaul, Y. (2001). Regulation of p53 stability and p53-dependent apoptosis by NADH quinone oxidoreductase 1. PNAS *98*, 1188-1193.

Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes & Development *17*, 126-140.

Balazsi, G., van Oudenaarden, A., and Collins, J.J. (2011). Cellular decision making and biological noise: from microbes to mammals. Cell *144*, 910-925.

Baron, V., Adamson, E.D., Calogero, A., Ragona, G., and Mercola, D. (2006). The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFbeta1, PTEN, p53, and fibronectin. Cancer Gene Therapy *13*, 115-124.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, e.-A.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., *et al.* (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science *332*, 687-696.

Bengtsson, M., Hemberg, M., Rorsman, P., and Stahlberg, A. (2008). Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. BMC Molecular Biology *9*, 63.

Bénit, L., De Parseval, N., Casella, J.F., Callebaut, I., Cordonnier, A., and Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. Journal of Virology *71*, 5652-5657.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bhargava, V., Ko, P., Willems, E., Mercola, M., and Subramaniam, S. (2013). Quantitative transcriptomics using designed primer-based amplification. Scientific Reports *3*, 1740.

Bibel, M., Richter, J., Lacroix, E., and Barde, Y.A. (2007). Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. Nature Protocols *2*, 1034-1043.

Blackledge N.P., Farcas A.M., Kondo T., King H.W., McGouran J.F., Hanssen L.L., Ito S., Cooper S., Kondo K., Koseki Y., Ishikura T., Long H.K., Sheahan T.W., Brockdorff N., Kessler B.M., Koseki H., and Klose R.J. (2014) Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445-1459

Bose, S., Wan, Z., Carr, A., Rizvi, A.H., Vieira, G., Pe'er, D., and Sims, P.A. (2015). Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biology *16*, 120.

Bowles, J., Teasdale, R.P., James, K., and Koopman, P. (2003). Dppa3 is a marker of pluripotency and has a human homologue that is expressed in germ cell tumours. Cytogenetic and Genome Research *101*, 261-265.

Bowsher, C.G., and Swain, P.S. (2012). Identifying sources of variation and the flow of information in biochemical networks. PNAS *109*, 1320-1328.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Bradley, A., Evans, M., Kaufman, M.H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature *309*, 255–256.

Bradley, A., Hasty, P., Davis, A., and Ramirez-Solis, R. (1992). Modifying the mouse: design and desire. Biotechnology *10*, 534-539.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., *et al.* (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods *10*, 1093-1095.

Brockdorff, N. (2013) Noncoding RNA and Polycomb recruitment. RNA *19*, 429–442.

Brons, I.G., Smithers, L.E., Trotter, M.W., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A., *et al.* (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191-195.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., *et al.* (2012). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157-170.

Buchwald, G., van der Stoop, P., Weichenrieder, O., Perrakis, A., van Lohuizen, M., and Sixma, T.K. (2000). Structure and E3-ligase activity of the Ring–Ring complex of Polycomb proteins Bmi1 and Ring1b. The EMBO Journal *25*, 2465-2474.

Buehr, M., Meek, S., Blair, K., Yang, J., Ure, J., Silva, J., McLay, R., Hall, J., Ying, Q.L., and Smith, A. (2008). Capture of authentic embryonic stem cells from rat blastocysts. Cell *135*, 1287-1298.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression

analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209-1222.

Bultman, S., Gebuhr, T., Yee, D., La Mantia, C., Nicholson, J., Gilliam, A., Randazzo, F., Metzger, D., Chambon, P., Crabtree, G., *et al.* (2009). A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes. Molecular Cell *6*, 1287-1295.

Burdon, T., Smith, A., and Savatier, P. (2002). Signalling, cell cycle and pluripotency in embryonic stem cells. Trends in Cellular Biology *12*, 432-438.

Burdon, T., Stracey, C., Chambers, I., Nichols, J., and Smith, A. (1999). Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. Developmental Biology *210*, 30–43.

Canham, M.A., Sharov, A.A., Ko, M.S., and Brickman, J.M. (2010). Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. PLoS Biology *8*, e1000379.

Cao Q., Wang X., Zhao M., Yang R., Malik R., Qiao Y., Poliakov A., Yocum A.K., Li Y., Chen W., Cao X., Jiang X., Dahiya A., Harris C., Feng F.Y., Kalantry S., Qin Z.S., Dhanasekaran S.M., Chinnaiyan A.M. (2014). The central role of EED in the orchestration of polycomb group complexes. Nature Communications *5*, 3127.

Capecchi, M.R. (2005). Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. Nature Reviews. Genetics *6*, 507-512.

Cartwright, P., McLean, C., Sheppard, A., Rivett, D., Jones, K., and Dalton, S. (2005). LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. Development *132*, 885-896.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230-1234.

Chambers, I., and Tomlinson, S.R. (2009). The transcriptional foundation of pluripotency. Development *136*, 2311-2322.

Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544-547.

Chapman, J.R., Barral, P., Vannier, J.B., Borel, V., Steger, M., Tomas-Loba, A., Sartori, A.A., Adams, I.R., Batista, F.D., and Boulton, S.J. (2013). RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. Molecular Cell *49*, 858-871.

Chason, R.J., Csokmay, J., Segars, J.H., DeCherney, A.H., and Armant, D.R. (2011). Environmental and epigenetic effects upon preimplantation embryo metabolism and development. Trends in endocrinology and metabolism: TEM *22*, 412-420.

Chattopadhyay, P.K., Price, D.A., Harper, T.F., Betts, M.R., Yu, J., Gostick, E., Perfetto, S.P., Goepfert, P., Koup, R.A., De Rosa, S.C., *et al.* (2006). Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. Nature Medicine *12*, 972-977.

Chew, J.L., Loh, Y.H., Zhang, W., Chen, X., Tam, W.L., Yeap, L.S., Li, P., Ang, Y.S., Lim, B., Robson, P., *et al.* (2005). Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Molecular and Cellular Biology *25*, 6031-6046.

Chickarmane, V., Troein, C., Nuber, U.A., Sauro, H.M., and Peterson, C. (2006). Transcriptional dynamics of the embryonic stem cell switch. PLoS Computational Biology 2, e123.

Ciancio, A., and Rizzetto, M. (2010). Thymalfasin in the treatment of hepatitis B and C. Annals of the New York Academy of Sciences *1194*, 141-146.

Clapier, C.R., and Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annual Review of Biochemistry *78*, 273-304.

Cole, M.F., Johnstone, S.E., Newman, J.J., Kagey, M.H., and Young, R.A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes & Development *22*, 746-755.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Consortium, E.R.C. (2005). Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics *6*, 150.

Cooper S., Dienstbier M., Hassan R., Schermelleh L., Sharif J., Blackledge N.P., De Marco V., Elderkin S., Koseki H., Klose R., Heger A., Brockdorff N. (2014) Targeting polycomb to pericentric heterochromatin in embryonic stem cells reveals a role for H2AK119u1 in PRC2 recruitment. Cell Reports *7*, 1456-1470

D'Angiolella, V., Donato, V., Forrester, F.M., Jeong, Y.T., Pellacani, C., Kudo, Y., Saraf, A., Florens, L., Washburn, M.P., and Pagano, M. (2012). Cyclin F-mediated degradation of ribonucleotide reductase M2 controls genome integrity and DNA repair. Cell *149*, 1023-1034.

Dahéron, L., Opitz, S.L., Zaehres, H., Lensch, M.W., Andrews, P.W., Itskovitz-Eldor, J., and Daley, G.Q. (2004). LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells. Stem Cells *22*, 770-778.

Davidson, K.C., Mason, E.A., and Pera, M.F. (2015). The pluripotent state in mouse and human. Development *142*, 3090-3099.

de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M., *et al.* (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Developmental Cell *7*, 663-676.

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

Díaz-Jullien, C., Pérez-Estévez, A., Covelo, G., and Freire, M. (1996). Prothymosin alpha binds histones in vitro and shows activity in nucleosome assembly assay. Biochimica et Biophysica Acta *1296*, 219-227.

Doetschman, T., Gregg, R.G., Maeda, N., Hooper, M.L., Melton, D.W., Thompson, S., and Smithies, O. (1987). Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. Nature *330*, 576-578.

Doetschman, T.C., Eistetter, H., Katz, M., Schmidt, W., and Kemler, R. (1985). The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. Journal of Embryology and Experimental Morphology *87*, 27-45.

Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

Du, J., Chen, T., Zou, X., Xiong, B., and Lu, G. (2010). Dppa2 knockdown-induced differentiation and repressed proliferation of mouse embryonic stem cells. Journal of Biochemistry *147*, 265-271.

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. PNAS *89*, 3010-3014.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D., Buetow, K.H., *et al.* (2008). Global transcription in pluripotent embryonic stem cells. Cell Stem Cell *2*, 437-447.

Efroni, S., Melcer, S., Nissim-Rafinia, M., and Meshorer, E. (2009). Stem cells do play with dice: a statistical physics view of transcription. Cell cycle *8*, 43-48.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic Gene Expression in a Single Cell. Science *297*, 1183-1186.

Erdemir, T., Bilican, B., Cagatay, T., Goding, C.R., and Yavuzer, U. (2002). Saccharomyces cerevisiae C1D is implicated in both non-homologous DNA end joining and homologous recombination. Molecular Microbiology *46*, 947-957.

Ernst, M., Oates, A., and Dunn, A.R. (1996). Gp130-mediated signal transduction in embryonic stem cells involves activation of Jak and Ras/mitogen-activated protein kinase pathways. Journal of Biological Chemistry *271*, 30136-30143.

Escribano-Diaz, C., Orthwein, A., Fradet-Turcotte, A., Xing, M., Young, J.T., Tkac, J., Cook, M.A., Rosebrock, A.P., Munro, M., Canny, M.D., *et al.* (2013). A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. Molecular Cell *49*, 872-883.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154-156.

Faddah, D.A., Wang, H., Cheng, A.W., Katz, Y., Buganim, Y., and Jaenisch, R. (2013). Single-cell analysis reveals that expression of nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. Cell Stem Cell *13*, 23-29.

Fan, H.C., Fu, G.K., and Fodor, S.P. (2015a). Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science *347*, 1258367.

Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., and Huang, Y. (2015b). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biology *16*, 148.

Fang, M., Xie, H., Dougan, S.K., Ploegh, H., and van Oudenaarden, A. (2013). Stochastic cytokine expression induces mixed T helper cell States. PLoS Biology *11*, e1001618.

Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schonegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Reports *10*, 1386-1397.

Fazzio, T.G., Huff, J.T., and Panning, B. (2008). An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. Cell *134*, 162-174.

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of Single RNA Transcripts in Situ. Science *280*, 585-590.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398-402.

Ficz, G., Hore, T.A., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.L., Walter, J., *et al.* (2013). FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell *13*, 351-359.

Filipczyk, A., Gkatzis, K., Fu, J., Hoppe, P.S., Lickert, H., Anastassiadis, K., and Schroeder, T. (2013). Biallelic expression of nanog protein in mouse embryonic stem cells. Cell Stem Cell *13*, 12-13.

Fok, E.Y., and Zandstra, P.W. (2005). Shear-controlled single-step mouse embryonic stem cell expansion and embryoid body-based differentiation. Stem Cells *23*, 1333-1342.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., *et al.* (2005). Epigenetic differences arise during the lifetime of monozygotic twins. PNAS *102*, 10604-10609.

Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science *306*, 1574-1577.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Harmelin, A., Rechavi, G., and Shapiro, E. (2008). Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. BMC Biotechnology *8*, 17.

Fu, G.K., Hu, J., Wang, P.H., and Fodor, S.P. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. PNAS *108*, 9026-9031.

Galupa, R., and Heard, E. (2015). X-chromosome inactivation: new insights into cis and trans regulation. Current Opinion in Genetics & Development *31*, 57-66.

Gao, X., Tsang, J.C., Gaba, F., Wu, D., Lu, L., and Liu, P. (2014). Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. Nucleic Acids Research *42*, e155.

Garaci, E., Pica, F., Serafino, A., Balestrieri, E., Matteucci, C., Moroni, G., Sorrentino, R., Zonfrillo, M., Pierimarchi, P., and Sinibaldi-Vallebona, P. (2012). Thymosin alpha1 and cancer: action on immune effector and tumor target cells. Annals of the New York Academy of Sciences *1269*, 26-33.

Garnier M., Di Lorenzo D., Albertini A., and Maggi A. (1997). Identification of estrogen-responsive genes in neuroblastoma SK-ER3 cells. Journal of Neuroscience 17, 4591-4599

Gawad C., Koh W., and Quake S.R. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics *17*, 175-88.

Gearing, D.P., Thut, C.J., VandeBos, T., Gimpel, S.D., Delaney, P.B., King, J., Price, V., Cosman, D., and Beckmann, M.P. (1991). Leukemia inhibitory factor receptor is structurally related to the IL-6 signal transducer, gp130. The EMBO Journal *10*, 2839-2848.

George, E.M., and Brown, D.T. (2010). Prothymosin alpha is a component of a linker histone chaperone. FEBS Letters *584*, 2833-2836.

George, S.H., Gertsenstein, M., Vintersten, K., Korets-Smith, E., Murphy, J., Stevens, M.E., Haigh, J.J., and Nagy, A. (2007). Developmental and adult phenotyping directly from mutant embryonic stem cells. PNAS *104*, 4455-4460.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., *et al.* (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647-661.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., *et al.* (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442-451.

Ginis, I., Luo, Y., Miura, T., Thies, S., Brandenberger, R., Gerecht-Nir, S., Amit, M., Hoke, A., Carpenter, M.K., Itskovitz-Eldor, J., *et al.* (2004). Differences between human and mouse embryonic stem cells. Developmental Biology *269*, 360-380.

Griffiths, D.S., Li, J., Dawson, M.A., Trotter, M.W., Cheng, Y.H., Smith, A.M., Mansfield, W., Liu, P., Kouzarides, T., Nichols, J., *et al.* (2011). LIF-independent JAK signalling to chromatin in embryonic stem cells uncovered from an adult stem cell disease. Nature Cell Biology *13*, 13-21.

Grimbergen, A.J., Siebring, J., Solopova, A., and Kuipers, O.P. (2015). Microbial bet-hedging: the power of being different. Current opinion in microbiology *25*, 67-72.

Grun, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat Methods *11*, 637-640.

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251-255.

Grun, D., and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell *163*, 799-810.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Research *23*, 2126-2135.

Guo, Y., Chang, H., Li, J., Xu, X.Y., Shen, L., Yu, Z.B., and Liu, W.C. (2015). Thymosin alpha 1 suppresses proliferation and induces apoptosis in breast cancer cells through PTEN-mediated inhibition of PI3K/Akt/mTOR signaling pathway. Apoptosis: an international journal on programmed cell death *20*, 1109-1121.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., *et al.* (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360-369.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989-2998.

Hall, J., Guo, G., Wray, J., Eyres, I., Nichols, J., Grotewold, L., Morfopoulou, S., Humphreys, P., Mansfield, W., Walker, R., *et al.* (2009). Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal. Cell Stem Cell *5*, 597-609.

Hanna, J., Cheng, A.W., Saha, K., Kim, J., Lengner, C.J., Soldner, F., Cassady, J.P., Muffat, J., Carey, B.W., and Jaenisch, R. (2010). Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. PNAS *107*, 9222-9227.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports *2*, 666-673.

Hasty, J., McMillen, D., Isaacs, F., and J., C.J. (2001). Computational studies of gene regulatory networks: in numero molecular biology. Nature Reviews. Genetics *2*, 268-279.

Hayashi, K., Lopes, S.M., Tang, F., and Surani, M.A. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell Stem Cell *3*, 391-401.

Hayashi, T., Shibata, N., Okumura, R., Kudome, T., Nishimura, O., Tarui, H., and Agata, K. (2010). Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. Development, Growth & Differentiation *52*, 131-144.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. Bio Systems *96*, 86-103.

Helgeson, B.E., Tomlins, S.A., Shah, N., Laxman, B., Cao, Q., Prensner, J.R., Cao, X., Singla, N., Montie, J.E., Varambally, S., *et al.* (2008). Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. Cancer Research *68*, 73-80.

Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474-484.

Ho, L., Jothi, R., Ronan, J.L., Cui, K., Zhao, K., and Crabtree, G.R. (2009). An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. PNAS *106*, 5187-5191.

Hondo, E., and Stewart, C.L. (2004). Profiling gene expression in growth-arrested mouse embryos in diapause. Genome Biology *6*, 202.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. Nature *326*, 292-295.

Huang, Y., Osorno, R., Tsakiridis, A., and Wilson, V. (2012). In Vivo differentiation potential of epiblast stem cells revealed by chimeric embryo formation. Cell Reports *2*, 1571-1578.

Huh, D., and Paulsson, J. (2011). Random partitioning of molecules at cell division. PNAS *108*, 15004-15009.

Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., and Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biology *17*, 29.

Ioannou, K., Samara, P., Livaniou, E., Derhovanessian, E., and Tsitsilonis, O.E. (2012). Prothymosin alpha: a ubiquitous polypeptide with potential use in cancer diagnosis and therapy. Cancer immunology, immunotherapy : CII *61*, 599-614.

Ishihara, S., Yasuda, M., Harada, I., Mizutani, T., Kawabata, K., and Haga, H. (2013). Substrate stiffness regulates temporary NF-kappaB activation via actomyosin contractions. Experimental Cell Research *319*, 2916-2927.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research *21*, 1160-1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell rna-seq with unique molecular identifiers. Nature Methods *11*, 163.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129-1133.

Itskovitz-Eldor, J., Schuldiner, M., Karsenti , D., Eden, A., Yanuka, O., Amit, M., Soreq, H., and Benvenisty, N. (2000). Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. Molecular Medicine *6*, 88-95.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533-538.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776-779.

James, M.A., Vikis, H.G., Tate, E., Rymaszewski, A.L., and You, M. (2014). CRR9/CLPTM1L regulates cell survival signaling and is required for Ras transformation and lung tumorigenesis. Cancer Research *74*, 1116-1127.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jia J., Zheng X., Hu G., Cui K., Zhang J., Zhang A., Jiang H., Lu B., Yates J. 3rd, Liu C., Zhao K., and Zheng Y. (2012). Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. Cell *151*, 576-589.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Research *21*, 1543-1551.

Jinek, M., Chylinski, K.F., I., and Hauer, M.D., J. A. Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816-821.

Johnson, G.L., and Lapadat, R. (2002). Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. Science *298*, 1911-1912.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Johnston, I.G., Gaal, B., Neves, R.P., Enver, T., Iborra, F.J., and Jones, N.S. (2012). Mitochondrial variability as a source of extrinsic cellular noise. PLoS Computational Biology *8*, e1002416.

Julia, M., Telenti, A., and Rausell, A. (2015). Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics *31*, 3380-3382.

Kaji, K., Caballero, I.M., MacLeod, R., Nichols, J., Wilson, V.A., and Hendrich, B. (2006). The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. Nature Cell Biology *8*, 285-292.

Kaji, K., Nichols, J., and Hendrich, B. (2007). Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. Development *134*, 1123-1132.

Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biology *7*, e1000149.

Kang, Y., Norris, M.H., Zarzycki-Siek, J., Nierman, W.C., Donachie, S.P., and Hoang, T.T. (2011). Transcript amplification from single bacterium for transcriptome analysis. Genome Research *21*, 925-935.

Kantlehner, M., Kirchner, R., Hartmann, P., Ellwart, J.W., Alunni-Fabbroni, M., and Schumacher, A. (2011). A high-throughput DNA methylation analysis of a single cell. Nucleic Acids Research *39*, e44.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. Nature Reviews. Molecular Cell Biology *9*, 770-780.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database *2011*.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wahlby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nature Methods *10*, 857-860.

Keays, K.M., Owens, G.P., Ritchie, A.M., Gilden, D.H., and Burgoon, M.P. (2005). Laser capture microdissection and single-cell RT-PCR without RNA purification. Journal of Immunological Methods *302*, 90-98.

Keller, G.M. (1995). In vitro differentiation of embryonic stem cells. Current Opinion in Cell Biology *7*, 862-869.

Kidder, B.L., Palmer, S., and Knott, J.G. (2009). SWI/SNF-Brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. Stem Cells *27*, 317-328.

Kigami, D., Minami, N., Takayama, H., and Imai, H. (2002). MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos. Biology of Reproduction *68*, 651-654.

Kim, D.H., Marinov, G.K., Pepke, S., Singer, Z.S., He, P., Williams, B., Schroth, G.P., Elowitz, M.B., and Wold, B.J. (2015a). Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. Cell Stem Cell *16*, 88-101.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049-1061.

Kim, J.K., Kolodziejczyk, A.A., Illicic, T., Teichmann, S.A., and Marioni, J.C. (2015b). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nature Communications *6*, 8687.

Kim, Y.J., Cecchini, K.R., and Kim, T.H. (2011). Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. PNAS *108*, 7391-7396.

Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods *9*, 72-74.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187-1201.

Kleinsmith, L.J., and Pierce, G.B.J. (1964). Multipotentiality of single embryonal carcinoma cells. Cancer Research, 1544-1551.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G., *et al.* (2011). Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell *8*, 200-213.

Kolch, W. (2000). Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. Biochemistry Journal *351*, 289-305.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015a). The technology and biology of single-cell RNA sequencing. Molecular Cell *58*, 610-620.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Buhler, M., Liu, P., *et al.* (2015b). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell *17*, 471-485.

Kuehn, M.R., Bradley, A., Robertson, E.J., and Evans, M.J. (1987). A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. Nature *326*, 295-298.

Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., DaleyKeyser, A.J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., *et al.* (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature *516*, 56-61.

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of

pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895-2902.

Kushwaha, R., Jagadish, N., Kustagi, M., Tomishima, M.J., Mendiratta, G., Bansal, M., Kim, H.R., Sumazin, P., Alvarez, M.J., Lefebvre, C., *et al.* (2015). Interrogation of a context-specific transcription factor network identifies novel regulators of pluripotency. Stem Cells *33*, 367-377.

Landry, J., Sharov, A.A., Piao, Y., Sharova, L.V., Xiao, H., Southon, E., Matta, J., Tessarollo, L., Zhang, Y.E., Ko, M.S., *et al.* (2008). Essential role of chromatin remodeling protein Bptf in early mouse embryos and embryonic stem cells. PLoS Genet *4*, e1000241.

Larson, M.H., Gilbert, L.A., Wang, X., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nature Protocols *8*, 2180-2196.

Laurent, C., Valet, F., Planque, N., Silveri, L., Maacha, S., Anezo, O., Hupe, P., Plancher, C., Reyes, C., Albaud, B., *et al.* (2011). High PTP4A3 phosphatase expression correlates with metastatic risk in uveal melanoma patients. Cancer Research *71*, 666-674.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S., Li, C., Amamoto, R., *et al.* (2014a). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360-1363.

Lee, M., Collins, J.W., Aubrecht, D.M., Sperling, R.A., Solomon, L., Ha, J.W., Yi, G.R., Weitz, D.A., and Manoharan, V.N. (2014b). Synchronized reinjection and coalescence of droplets in microfluidics. Lab on a Chip *14*, 509.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., *et al.* (2013). Naive pluripotency is associated with global DNA hypomethylation. Nature Structural & Molecular Biology *20*, 311-316.

Leng, N., Chu, L.F., Barry, C., Li, Y., Choi, J., Li, X., Jiang, P., Stewart, R.M., Thomson, J.A., and Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nature Methods *12*, 947-950.

Li, M., Liu, G.H., and Izpisua Belmonte, J.C. (2012). Navigating the epigenetic landscape of pluripotent stem cells. Nature Reviews. Molecular Cell Biology *13*, 524-535.

Li, X., Zhu, L., Yang, A., Lin, J., Tang, F., Jin, S., Wei, Z., Li, J., and Jin, Y. (2011). Calcineurin-NFAT signaling critically regulates early lineage specification in mouse embryonic stem cells and embryos. Cell Stem Cell *8*, 46-58.

Lim, S., and Kaldis, P. (2013). Cdks, cyclins and CKIs: roles beyond cell cycle regulation. Development *140*, 3079-3093.

Lin, C.Y., Loven, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56-67.

Liu Y., Beyer A., and Aebersold R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535-550.

Loh, K.M., and Lim, B. (2011). A precarious balance: pluripotency factors as lineage specifiers. Cell stem cell *8*, 363-369.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.* (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genetics *38*, 431-440.

Losick, R., and Desplan, C. (2008). Stochasticity and Cell Fate. Science *320*, 65-68.

Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M., Wolf, J.A., Grady, M.S., *et al.* (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nature Methods *11*, 190-196.

Lun, A.T.L., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology *17*, 75.

MacArthur, B.D., Sevilla, A., Lenz, M., Muller, F.J., Schuldt, B.M., Schuppert, A.A., Ridden, S.J., Stumpf, P.S., Fidalgo, M., Ma'ayan, A., *et al.* (2012). Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. Nature Cell Biology *14*, 1139-1147.

Macaulay, I.C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S.A., and Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Reports *14*, 966-977.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., *et al.* (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes & Development *25*, 594-607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57-63.

Machin, G.A. (1996). Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. American Journal of Medical Genetics *61*, 216-228.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., *et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Mahata, B., Zhang, X., Kolodziejczyk, A.A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A.E., Hebenstreit, D., Dingler, F.A., Moignard, V., Gottgens, B., *et al.* (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Reports *7*, 1130-1142.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., *et al.* (2011). Evolution and classification of the CRISPR-Cas systems. Nature Reviews. Microbiology *9*, 467-477.

Maksakova I.A., Romanish M.T., Gagnier L., Dunn C.A., van de Lagemaat L.N., and Mager D.L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genetics 2:e2

Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J., Singh, P.B., Karimi, M.M., and Lorincz, M.C. (2013). Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics Chromatin. *6*, 15.

Mantalenakis, S.J., and Ketchel, M.M. (1966). Frequency and extent of delayed implantation in lactating rats and mice. J Reprod Fertil *12*, 391-394.

Marco, E., Karp, R.L., Guo, G., Robson, P., Hart, A.H., Trippa, L., and Yuan, G.C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proceedings of the National Academy of Sciences of the United States of America *111*, E5643-5650.

Marcy Y., Ouverney C., Bik E.M., Lösekann T., Ivanova N., Martin H.G., Szeto E., Platt D., Hugenholtz P., Relman D.A., and Quake S.R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes

from the human mouth. Proceedings of the National Academy of Sciences of the United States of America *104*, 11889-94

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343-349.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Research.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., *et al.* (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590-604.

Martello, G., and Smith, A. (2014). The nature of embryonic stem cells. Annual review of cell and developmental biology *30*, 647-675.

Martin, D.M. (2010). Chromatin remodeling in development and disease: focus on CHD7. PLoS Genet. *6*, e1001010.

Martin, G.R. (1975). Teratocarcinomas as a model system for the study of embryogenesis and neoplasia. Cell *5*, 229-243.

Martin, G.R. (1980). Teratocarcinomas and mammalian embryogenesis. Science *209*, 768-776.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. PNAS, 7634–7638.

Martin, G.R., and Evans, M.J. (1974). The morphology and growth of a pluripotent teratocarcinoma cell line and its derivatives in tissue culture. Cell *2*, 163-172.

Martini P.G., Delage-Mourroux R., Kraichely D.M., Katzenellenbogen B.S. (2000) Prothymosin alpha selectively enhances estrogen receptor transcriptional activity by interacting with a repressor of estrogen receptor activity. Molecular and Cellular Biology *20*, 6224-6232.

Martinez Arias, A., and Brickman, J.M. (2011). Gene expression heterogeneities in embryonic stem cell populations: origin and function. Current opinion in Cell Biology *23*, 650-656.

Matsuda, T., Nakamura, T., Nakao, K., Arai, T., Katsuki, M., Heike, T., and Yokota, T. (1999). STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. The EMBO journal *18*, 4261-4269.

Matsui, H., Harada, I., and Sawada, Y. (2012). Src, p130Cas, and Mechanotransduction in Cancer Cells. Genes & cancer *3*, 394-401.

Mazutis, L., Gilbert, J., Ung, W.L., Weitz, D.A., Griffiths, A.D., and Heyman, J.A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. Nature Protocols *8*, 870-891.

Merkle, F.T., and Eggan, K. (2013). Modeling human disease with pluripotent stem cells: from genome association to function. Cell Stem Cell *12*, 656-668.

Meshorer, E., and Misteli, T. (2006). Chromatin in pluripotent embryonic

stem cells and differentiation. Nature Reviews. Molecular Cell Biology *7*, 450-456.

Meyn, M.A.r., and Smithgall, T.E. (2009). Chemical genetics identifies c-Src as an activator of primitive ectoderm formation in murine embryonic stem cells. Sci Signal *2*, ra64.

Minina, Y.M., Zhdanova, N.S., Shilov, A.G., Tolkunova, E.N., Liskovykh, M.A., and Tomilin, A.N. (2010). Chromosomal instability of mouse pluripotent cells cultured in vitro. Cell and Tissue Biology *4*, 223-227.

Mintz, B., and Illmensee, K. (1975). Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proceedings of the National Academy of Sciences of the United States of America *72*, 3585–3589.

Mishra, P., and Chan, D.C. (2014). Mitochondrial dynamics and inheritance during cell division, development and disease. Nature Reviews. Molecular Cell Biology *15*, 634-646.

Mitra, R.D., Shendure, J., Olejnik, J., E., K.-O., and Church, G.M. (2003). Fluorescent in situ sequencing on polymerase colonies. Analytical Biochemistry *320*, 55-65.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. Cell *113*, 631-642.

Miyanari, Y., and Torres-Padilla, M.E. (2012). Control of ground-state pluripotency by allelic regulation of Nanog. Nature *483*, 470-473.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., *et al.* (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nature Biotechnology *33*, 269-276.

Morey, L., Pascual, G., Cozzuto, L., Roma, G., Wutz, A., Benitah, S.A., and Di Croce, L. (2012). Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. Cell Stem Cell *10*, 47-62.

Morgan, H.D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. Human Molecular Genetics *14 Spec No 1*, R47-58.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., *et al.* (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics *44*, 981-990.

Morrison, S.J., and Kimble, J. (2006). Asymmetric and symmetric stem-cell divisions in development and cancer. Nature *441*, 1068-1074.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., and Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods *7*, 621-628.

Moyes D., Griffiths D.J., and Venables P.J. (2007). Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends in Genetics *23*, 326-333.

Mozzetta C., Pontis J., Fritsch L., Robin P., Portoso M., Proux C., Margueron R., Ait-Si-Ali S. (2014). The histone H3 lysine 9 methyltransferases G9a and GLP regulate polycomb repressive complex 2-mediated gene silencing. Molecular Cell *53*, 277-289

Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W., and Roder, J.C. (1993). Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. PNAS *90*, 8424-8428.

Nakamura, T., Yabuta, Y., Okamoto, I., Aramaki, S., Yokobayashi, S., Kurimoto, K., Sekiguchi, K., Nakagawa, M., Yamamoto, T., and Saitou, M. (2015). SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. Nucleic Acids Research 43, e60.

Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. Cell *154*, 490-503.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Ng, H.H., and Surani, M.A. (2011). The transcriptional and signalling networks of pluripotency. Nature Cell Biology *13*, 490-496.

Nichols, J., Chambers, I., Taga, T., and Smith, A. (2001). Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. Development *128*, 2333-2339.

Nichols, J., Jones, K., Phillips, J.M., Newland, S.A., Roode, M., Mansfield, W., Smith, A., and Cooke, A. (2009). Validated germline-competent embryonic stem cell lines from nonobese diabetic mice. Nature Medicine *15*, 814-818.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. Cell *95*, 379-391.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., *et al.* (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell *151*, 68-79.

Niedernhofer, L.J., Essers, J., Weeda, G., Beverloo, B., de Wit, J., Muijtjens, M., Odijk, H., Hoeijmakers, J.H., and Kanaar, R. (2001). The structure-specific endonuclease Ercc1-Xpf is required for targeted gene replacement in embryonic stem cells. The EMBO Journal *20*, 6540-6549.

Nimmo, R.A., May, G.E., and Enver, T. (2015). Primed and ready: understanding lineage commitment through single cell analysis. Trends Cellular Biology *25*, 459-467.

Niwa, H., Miyazaki, J., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature Genetics *24*, 372-376.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118-122.

Nurse, P. (2000). A long twentieth century of the cell cycle and beyond. Cell *100*, 71-78.

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Molecular and Cellular Biology *21*, 4330-4336.

O'Loghlen, A., Munoz-Cabello, A.M., Gaspar-Maia, A., Wu, H.A., Banito, A., Kunowska, N., Racek, T., Pemberton, H.N., Beolchi, P., Lavial, F., *et al.* (2012). MicroRNA regulation of Cbx7 mediates a switch of Polycomb orthologs during ESC differentiation. Cell Stem Cell *10*, 33-46.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Onishi, K., and Zandstra, P.W. (2015). LIF signaling in stem cells and development. Development *142*, 2230-2236.

Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., Hino, T., Suzuki-Migishima, R., Ogonuki, N., Miki, H., *et al.* (2006). Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nature Genetics *38*, 101-106.

Ooi, S.K., and Bestor, T.H. (2008). The colorful history of active DNA demethylation. Cell *133*, 1145-1148.

Orford, K.W., and Scadden, D.T. (2008). Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. Nature Reviews. Genetics *9*, 115-128.

Orkin, S.H., Wang, J., Kim, J., Chu, J., Rao, S., Theunissen, T.W., Shen, X., and Levasseur, D.N. (2008). The transcriptional network controlling pluripotency in ES cells. Cold Spring Harbor Symposia on Quantitative Biology, 195-202.

Palani, S., and Sarkar, C.A. (2012). Transient noise amplification and gene expression synchronization in a bistable mammalian cell-fate switch. Cell Reports *1*, 215-224.

Pan, G., and Thomson, J.A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. Cell Research *17*, 42-49.

Pan, G.J., Chang, Z.Y., Schöler, H.R., and Pei, D. (2002). Stem cell pluripotency and transcription factor Oct4. Cell Research *12*, 321-329.

Papatsenko, D., Darr, H., Kulakovskiy, I.V., Waghray, A., Makeev, V.J., MacArthur, B.D., and Lemischka, I.R. (2015). Single-Cell Analyses of ESCs Reveal Alternative Pluripotent Cell States and Molecular Mechanisms that Control Self-Renewal. Stem Cell Reports *5*, 207-220.

Pardo, M., Lang, B., Yu, L., Prosser, H., Bradley, A., Babu, M.M., and Choudhary, J. (2010). An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. Cell Stem Cell *6*, 382-395.

Pawlak, M., and Jaenisch, R. (2011). De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state. Genes & Development *25*, 1035-1040.

Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Developmental Cell *7*, 597-606.

Peshkin, L., Wuhr, M., Pearl, E., Haas, W., Freeman, R.M., Jr., Gerhart, J.C., Klein, A.M., Horb, M., Gygi, S.P., and Kirschner, M.W. (2015). On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. Developmental Cell *35*, 383-394.

Petropoulos S., Edsgärd D., Reinius B., Deng Q., Panula S.P., Codeluppi S., Plaza Reyes A., Linnarsson S., Sandberg R., and Lanner F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012-1026

Picelli, S., Bjorklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods *10*, 1096-1098.

Picelli, S., Faridani, O.R., Bjorklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols *9*, 171-181.

Pierce, G.B. (1967). Chapter 8: Teratocarcinoma: Model for A Developmental Concept of Cancer. Current Topics in Developmental Biology *2*, 223-246.

Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biology *16*, 241.

Pineiro, A., Cordero, O.J., and Nogueira, M. (2000). Fifteen years of prothymosin alpha: contradictory past and new horizons. Peptides *21*, 1433-1446.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P.*, et al.* (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology *32*, 1053-1058.

Pollex, T., and Heard, E. (2012). Recent advances in X-chromosome inactivation research. Current Opinion in Cell Biology *24*, 825-832.

Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates, J.R.r., and Grant, P.A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature *433*, 434-438.

Proserpio V., and Lönnberg T. (2016). Single-cell technologies are revolutionizing the approach to rare cells. Immunology and Cell Biology *94*, 225-9.

Pruitt, K., and Der, C.J. (2001). Ras and Rho regulation of the cell cycle and oncogenesis. Cancer Letters *171*, 1-10.

Qiu, L., Liu, M., and Pan, K. (2013). A triple staining method for accurate cell cycle analysis using multiparameter flow cytometry. Molecules *18*, 15412-15421.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biology *4*, e309.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods *5*, 877-879.

Ramskold, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., *et al.* (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology *30*, 777-782.

Raser, J.M., and O'Shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Renfree, M.B., and Shaw, G. (2000). Diapause. Annual Review of Physiology *62*, 353-375.

Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A., and Bongso, A. (2000). Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nature Biotechnology *18*, 399-404.

Reynolds, N., Latos, P., Hynes-Allen, A., Loos, R., Leaford, D., O'Shaughnessy, A., Mosaku, O., Signolet, J., Brennecke, P., Kalkan, T., *et al.* (2012). NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. Cell Stem Cell *10*, 583-594.

Ringrose, L., Ehret, H., and Paro, R. (2004). Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. Molecular Cell *16*, 641-653.

Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. (2005). Transcriptional regulation of nanog by OCT4 and SOX2. The Journal of Biological Chemistry *280*, 24731-24737.

Roeder, I., and Radtke, F. (2009). Stem cell biology meets systems biology. Development *136*, 3525-3530.

Roessler, M., Rollinger, W., Mantovani-Endl, L., Hagmann, M.L., Palme, S., Berndt, P., Engel, A.M., Pfeffer, M., Karl, J., Bodenmüller, H., *et al.* (2006). Identification of PSME3 as a novel serum tumor marker for colorectal cancer by combining two-dimensional polyacrylamide gel electrophoresis with a strictly mass spectrometry-based approach for data analysis. Mol Cell Proteomics *5*, 2092-2101.

Romani, L., Moretti, S., Fallarino, F., Bozza, S., Ruggeri, L., Casagrande, A., Aversa, F., Bistoni, F., Velardi, A., and Garaci, E. (2012). Jack of all trades: thymosin alpha1 and its pleiotropy. Annals of the New York Academy of Sciences *1269*, 1-6.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene regulation at the single-cell level. Science *307*, 1962-1965.

Rosenthal, M.D., Wishnow, R.M., and Sato, G.H. (1970). In vitro growth and differetiation of clonal populations of multipotential mouse cells derived from a

transplantable testicular teratocarcinoma. The Journal of the National Cancer Institute *44*, 1001-1014.

Ross, D., and Siegel, D. (2004). NAD(P)H:Quinone Oxidoreductase 1 (NQO1,DT-Diaphorase), Functions and Pharmacogenetics. Methods in Enzymology 382, 114-144.

Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S.L., and Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Research *23*, 452-461.

Saha, A., Wittmeyer, J., and Cairns, B.R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. Nature Reviews. Molecular Cell Biology *7*, 437-447.

Saiz, N., and Plusa, B. (2013). Early cell fate decisions in the mouse embryo. Reproduction *145*, R65-80.

Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. Science *342*, 1188-1193.

Santos, A., Wernersson, R., and Jensen, L.J. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Research *43*, D1140-1144.

Santos-Rosa, H., Schneider, R., Bernstein, B.E., Karabetsou, N., Morillon, A., Weise, C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. Molecular Cell *12*, 1325-1332.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., and Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biology *14*, R31.

Saunders, A., Faiola, F., and Wang, J. (2013). Concise review: pursuing self-renewal and pluripotency with the stem cell factor Nanog. Stem Cells *31*, 1227-1236.

Savva, R., McAuley-Hecht, K., Brown, T., and Pearl, L. (1995). The structural basis of specific base-excision repair by uracil-DNA glycosylase. Nature *373*, 487-493.

Schlesinger, S., and Goff, S.P. (2015). Retroviral transcriptional regulation and embryonic stem cells: war and peace. Molecular and Cellular Biology *35*, 770-777.

Schwartz, C.M., Tavakoli, T., Jamias, C., Park, S.S., Maudsley, S., Martin, B., Phillips, T.M., Yao, P.J., Itoh, K., Ma, W., *et al.* (2012). Stromal factors SDF1alpha, sFRP1, and VEGFD induce dopaminergic neuron differentiation of human pluripotent stem cells. Journal of Neuroscience Research *90*, 1367-1381.

Scialdone, A., Achim, K., and Marioni, J.C. (2014). Single Cell Genomics meeting in Stockholm: from single cells to cell types. Genome Biology *15*, 496.

Scognamiglio, R., Cabezas-Wallscheid, N., Thier, M.C., Altamura, S., Reyes, A., Prendergast, A.M., Baumgartner, D., Carnevalli, L.S., Atzberger, A., Haas, S., *et al.* (2016). Myc Depletion Induces a Pluripotent Dormant State Mimicking Diapause. Cell *164*, 668-680.

Seo, S.B.M., P. Heo, S., Turner, A., and Lane, W.S.C., D. (2001). Regulation of histone acetylation and transcription by INHAT, a human cellular complex containing the set oncoprotein. Cell *104*, 119-130.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., *et al.* (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363-369.

Sharov, A.A., Masui, S., Sharova, L.V., Piao, Y., Aiba, K., Matoba, R., Xin, L., Niwa, H., and Ko, M.S. (2008). Identification of Pou5f1, Sox2, and Nanog downstream target

genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics *9*, 269.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S. (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA research *16*, 45-58.

Shi, W., Wang, H., Pan, G., Geng, Y., Guo, Y., and Pei, D. (2006). Regulation of the pluripotency marker Rex-1 by Nanog and Sox2. The Journal of Biological Chemistry *281*, 23319-23325.

Shimizu, T., Ueda, J., Ho, J.C., Iwasaki, K., Poellinger, L., Harada, I., and Sawada, Y. (2012). Dual inhibition of Src and GSK3 maintains mouse embryonic stem cells, whose differentiation is mechanically regulated by Src signaling. Stem Cells *30*, 1394-1404.

Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G.L., *et al.* (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell *17*, 360-372.

Shiraishi, T., Matsuyama, S., and Kitano, H. (2010). Large-scale analysis of network bistability for human cancers. PLoS Comput Biology *6*, e1000851.

Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proceedings of the National Academy of Sciences of the United States of America *109*, 1347-1352.

Shukla, S., Nair, R., Rolle, M.W., Braun, K.R., Chan, C.K., Johnson, P.Y., Wight, T.N., and McDevitt, T.C. (2010). Synthesis and organization of hyaluronan and versican by embryonic stem cells undergoing embryoid body differentiation. The Journal of Histochemistry and Cytochemistry *58*, 345-358.

Sigurdsson, S., Van Komen, S., Bussen, W., Schild, D., Albala, J.S., and Sung, P. (2001). Mediator function of the human Rad51B–Rad51C complex in Rad51/RPA-catalyzed DNA strand exchange. Genes & Development *15*, 3308-3318.

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T.W., and Smith, A. (2008). Promotion of reprogramming to ground state pluripotency by signal inhibition. PLoS Biology *6*, e253.

Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell *138*, 722-737.

Singh, A.M., and Dalton, S. (2009). The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. Cell Stem Cell *5*, 141-149.

Singh, A.M., Hamazaki, T., Hankowski, K.E., and Terada, N. (2007). A heterogeneous expression pattern for Nanog in embryonic stem cells. Stem Cells *25*, 2534-2542.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods *11*, 817-820.

Smith, A.G. (2001). Embryo-derived stem cells: of mice and men. Annual Review of Cell and Developmental Biology *17*, 435-462.

Smith, A.G., Heath, J.K., Donaldson, D.D., Wong, G.G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature *336*, 688-690.

Smith, A.G., and Hooper, M.L. (1987). Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. Developmental Biology *121*, 1-9.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339-344.

Soderberg, O., Gullberg, M., Jarvius, M., Ridderstrale, K., Leuchowius, K.J., Jarvius, J., Wester, K., Hydbring, P., Bahram, F., Larsson, L.G., *et al.* (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nature Methods *3*, 995-1000.

Sokolik, C., Liu, Y., Bauer, D., McPherson, J., Broeker, M., Heimberg, G., Qi, L.S., Sivak, D.A., and Thomson, M. (2015). Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. Cell Systems *1*, 117-129.

Stanghellini, I., Falco, G., Lee, S.L., Monti, M., and Ko, M.S. (2009). Trim43a, Trim43b, and Trim43c: Novel mouse genes expressed specifically in mouse preimplantation embryos. Gene Expression Patterns *9*, 595-602.

Stavridis, M.P., Lunn, J.S., Collins, B.J., and Storey, K.G. (2007). A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification. Development *134*, 2889-2894.

Stevens, L.C. (1970). The development of transplantable teratocarcinomas from intratesticular grafts of pre- and postimplantation mouse embryos. Developmental Biology *21*, 364-382.

Stevens, L.C., and Little, C.C. (1954). Spontaneous Testicular Teratomas in an Inbred Strain of Mice. PNAS *40*, 1080-1087.

Steward, F.C., Mapes, M.O., and Mears, K. (1958). Growth and organized development of cultured cells. II. Organization in cultures grown from freely suspended cells. American Journal of Botany *45*, 705-708.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Suda, Y., Suzuki, M., Ikawa, Y., and Aizawa, S. (1987). Mouse embryonic stem cells exhibit indefinite proliferative potential. Journal of Cell Physiology *133*, 197-201.

Suzuki, O., Matsuda, J., Takano, K., Yamamoto, Y., Asano, T., Naiki, M., and Kusanagi, M. (1999). Effect of genetic background on establishment of mouse embryonic stem cells. Experimental Animals *48*, 213-216.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. PNAS *99*, 12795-12800.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Takata, M., Sasaki, M.S., Sonoda, E., Fukushima, T., Morrison, C., Albala, J.S., Swagemakers, S.M., Kanaar, R., Thompson, L.H., and Takeda, S. (2000). The Rad51 paralog Rad51B promotes homologous recombinational repair. Molecular and Cellular Biology *20*, 6476-6482.

Tam, P.P., and Behringer, R.R. (1997). Mouse gastrulation: the formation of a mammalian body plan. Mechanisms of Development *68*, 3-25.

Tam, P.P., and Loebel, D.A. (2007). Gene function in mouse embryogenesis: get set for gastrulation. Nature Reviews. Genetics *8*, 368-381.

Tanaka, S., Kunath, T., Hadjantonakis, A.K., Nagy, A., and Rossant, J. (1998). Promotion of trophoblast stem cell proliferation by FGF4. Science *282*, 2072-2075.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell *6*, 468-478.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., *et al.* (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods *6*, 377-382.

Taniguchi, K., Kajiyama, T., and Kambara, H. (2009). Quantitative analysis of gene expression in a single cell by qPCR. Nature Methods *6*, 503-506.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196-199.

Thiel, G., and Cibelli, G. (2002). Regulation of life and death by the zinc finger transcription factor Egr-1. Journal of Cellular Physiology *193*, 287-292.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science *282*, 1145-1147.

Tokuzawa, Y., Kaiho, E., Maruyama, M., Takahashi, K., Mitsui, K., Maeda, M., Niwa, H., and Yamanaka, S. (2003). Fbx15 Is a Novel Target of Oct3/4 but Is Dispensable for Embryonic Stem Cell Self-Renewal and Mouse Development. Molecular and Cellular Biology *23*, 2699-2708.

Torres-Padilla, M.E., and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Development *141*, 2173-2181.

Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K., and Niwa, H. (2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. Development *135*, 909-918.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology *32*, 381-386.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371-375.

Trouillas, M., Saucourt, C., Guillotin, B., Gauthereau, X., Ding, L., Buchholz, F., Doss, M.X., Sachinidis, A., Hescheler, J., Hummel, O., *et al.* (2009). Three LIF-dependent signatures and gene clusters with atypical expression profiles, identified by transcriptome studies in mouse ES cells and early derivatives. BMC Genomics *10*, 73.

Tsang, J.C.H., Yu, Y., Burke, S., Buettner, F., Wang, C., Kolodziejczyk, A.A., Teichmann, S.A., Lu, L., and Liu, P. (2015). Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biology *16*, 178.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., *et al.* (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes to Cells *11*, 805-814.

Turelli, P., Castro-Diaz, N., Marzetta, F., Kapopoulou, A., Raclot, C., Duc, J., Tieng, V., Quenneville, S., and Trono, D. (2014). Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. Genome Research *24*, 1260-1270.

Tyagi, S., and Kramer, F.R. (1996). Molecular beacons: probes that fluoresce upon hybridization. Nature Biotechnology *14*, 303-308.

van den Berg, D.L., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell *6*, 369-381.

van den Berg, D.L., Zhang, W., Yates, A., Engelen, E., Takacs, K., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2008). Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. Molecular and Cellular Biology *28*, 5986-5995.

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 2579-2605.

Van Eynde, A., Nuytten, M., Dewerchin, M., Schoonjans, L., Keppens, S., Beullens, M., Moons, L., Carmeliet, P., Stalmans, W., and Bollen, M. (2004). The nuclear scaffold protein NIPP1 is essential for early embryonic development and cell proliferation. Molecular and Cellular Biology *24*, 5863-5874.

Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Research *41*, 4378-4391.

Veening, J.W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. Annual Review of Microbiology *62*, 193-210.

Vermeulen, K., Van Bockstaele, D.R., and Berneman, Z.N. (2003). The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. Cell Proliferation *36*, 131-149.

Waghray, A., Saiz, N., Jayaprakash, A.D., Freire, A.G., Papatsenko, D., Pereira, C.F., Lee, D.F., Brosh, R., Chang, B., Darr, H., *et al.* (2015). Tbx3 Controls Dppa3 Levels and Exit from Pluripotency toward Mesoderm. Stem Cell Reports *5*, 97-110.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873-878.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature *444*, 364-368.

Wang Y., Liska F., Gosele C., Sedová L., Kren V., Krenová D., Ivics Z., Hubner N., and Izsvák Z. (2010) A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. Genome Research *20*, 19-27.

Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. PNAS *103*, 17807-17812.

Watanabe, D., Suetake, I., Tada, T., and Tajima, S. (2002). Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. Mechanisms of Development *118*, 187-190.

Welsby, I., Hutin, D., Gueydan, C., Kruys, V., Rongvaux, A., and Leo, O. (2014). PARP12, an interferon-stimulated gene involved in the control of protein translation and inflammation. The Journal of Biological Chemistry *289*, 26642-26657.

Wennekamp, S., Mesecke, S., Nedelec, F., and Hiiragi, T. (2013). A self-organization framework for symmetry breaking in the mammalian embryo. Nature Reviews. Molecular Cell Biology *14*, 452-459.

Whyte, W.A., Bilodeau, S., Orlando, D.A., Hoke, H.A., Frampton, G.M., Foster, C.T., Cowley, S.M., and Young, R.A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature *482*, 221-225.

Wight, T.N. (2002). Versican: a versatile extracellular matrix proteoglycan in cell biology. Current Opinion in Cell Biology *14*, 617-623.

Williams, R.L., Hilton, D.J., Pease, S., Willson, T.A., Stewart, C.L., Gearing, D.P., Wagner, E.F., Metcalf, D., Nicola, N.A., and Gough, N.M. (1988). Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. Nature *336*, 684-687.

Wolf, D., and Goff, S.P. (2007). TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell *131*, 46-57.

Wolf, D., and Goff, S.P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature *458*, 1201-1204.

Wolf, G., Yang, P., Fuchtbauer, A.C., Fuchtbauer, E.M., Silva, A.M., Park, C., Wu, W., Nielsen, A.L., Pedersen, F.S., and Macfarlan, T.S. (2015). The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes & Development *29*, 538-554.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., *et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics *46*, 1173-1186.

Wu, M.Y., Ramel, M.C., Howell, M., and Hill, C.S. (2011). SNW1 is a critical regulator of spatial BMP activity, neural plate border formation, and neural crest specification in vertebrate embryos. PLoS Biology *9*, e1000593.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Wu, Z., Yang, M., Liu, H., Guo, H., Wang, Y., Cheng, H., and Chen, L. (2012). Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. The Journal of Biological Chemistry *287*, 38295-38304.

Wysocka, J., Swigut, T., Milne, T.A., Dou, Y., Zhang, X., Burlingame, A.L., Roeder, R.G., Brivanlou, A.H., and Allis, C.D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. Cell *121*, 859-872.

Xu, C., Inokuma, M.S., Denham, J., Golds, K., Kundu, P., Gold, J.D., and Carpenter, M.K. (2001). Feeder-free growth of undifferentiated human embryonic stem cells. Nature Biotechnology *19*, 971-974.

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics *31*, 1974-80.

Xu, H., Ang, Y.S., Sevilla, A., Lemischka, I.R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. PLoS computational biology *10*, e1003777.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database *2013*, bat045.

Xue, J.H., Xu, G.F., Gu, T.P., Chen, G.D., Han, B.B., Xu, Z.M., Bjoras, M., Krokan, H.E., Xu, G.L., and Du, Y.R. (2016). Uracil-DNA Glycosylase UNG Promotes Tet-mediated DNA Demethylation. The Journal of Biological Chemistry *291*, 731-738.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., *et al.* (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593-597.

Yan L., Yang M., Guo H., Yang L., Wu J., Li R., Liu P., Lian Y., Zheng X., Yan J., Huang J., Li M., Wu X., Wen L., Lao K., Li R., Qiao J., and Tang F. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature Structural & Molecular Biology *20*, 1131-1139.

Yan, Z., Wang, Z., Sharova, L., Sharov, A.A., Ling, C., Piao, Y., Aiba, K., Matoba, R., Wang, W., and Ko, M.S. (2008). BAF250B-associated SWI/SNF chromatin-remodeling complex is required to maintain undifferentiated mouse embryonic stem cells. Stem Cells *26*, 1155-1165.

Yang, C.H., Murti, A., Baker, S.J., Frangou-Lazaridis, M., Vartapetian, A.B., Murti, K.G., and Pfeffer, L.M. (2004). Interferon induces the interaction of prothymosin-alpha with STAT3 and results in the nuclear translocation of the complex. Experimental Cell Research *298*, 197-206.

Yavuzer, U., Smith, G.C., Bliss, T., Werner, D., and Jackson, S.P. (1998). DNA end-independent activation of DNA-PK mediated via association with the DNA-binding protein C1D. Genes & Development *12*, 2188-2199.

Ying, Q.L., Nichols, J., Chambers, I., and Smith, A. (2003a). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell *115*, 281-292.

Ying, Q.L., Stavridis, M., Griffiths, D., Li, M., and Smith, A. (2003b). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nature Biotechnology *21*, 183-186.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.

Yokoyama, H., Kurumizaka, H., Ikawa, S., Yokoyama, S., and Shibata, T. (2003). Holliday junction binding activity of the human Rad51B protein. The Journal of Biological Chemistry *278*, 2767-2772.

Young, R.A. (2011). Control of the embryonic stem cell state. Cell *144*, 940-954.

Yu M., Mazor T., Huang H., Huang H.T., Kathrein K.L., Woo A.J., Chouinard C.R., Labadorf A., Akie T.E., Moran T.B., Xie H., Zacharek S., Taniuchi I., Roeder R.G., Kim C.F., Zon L.I., Fraenkel E., Cantor A.B. (2012) Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. Molecular Cell *45*, 330-343

Yusa, K., Zhou, L., Li, M.A., Bradley, A., and Craig, N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. PNAS *108*, 1531-1536.

Zalzman, M., Falco, G., Sharova, L.V., Nishiyama, A., Thomas, M., Lee, S.L., Stagg, C.A., Hoang, H.G., Yang, H.T., Indig, F.E., *et al.* (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature *464*, 858-863.

Zeileis, A., and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software *14*.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138-1142.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology *4*, Article17.

Zhang, P., Andrianakos, R., Yang, Y., Liu, C., and Lu, W. (2010). Kruppel-like factor 4 (Klf4) prevents embryonic stem (ES) cell differentiation by regulating Nanog gene expression. The Journal of Biological Chemistry *285*, 9180-9189.

Zhang, Q., Dan, J., Wang, H., Guo, R., Mao, J., Fu, H., Wei, X., and Liu, L. (2016). Tcstv1 and Tcstv3 elongate telomeres of mouse ES cells. Scientific Reports *6*, 19852.

Zhang, W., and Liu, H.T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Research *12*, 9-18.

Zhang, Z., Jones, A., Sun, C.W., Li, C., Chang, C.W., Joo, H.Y., Dai, Q., Mysliwiec, M.R., Wu, L.C., Guo, Y., *et al.* (2011). PRC2 complexes with JARID2, MTF2, and esPRC2p48 in ES cells to modulate ES cell pluripotency and somatic cell reprogramming. Stem Cells *29*, 229-240.

Zimmermann, M., Lottersberger, F., Buonomo, S.B., Sfeir, A., and de Lange, T. (2013). 53BP1 regulates DSB repair using Rif1 to control 5' end resection. Science *229*, 700-704.

zur Nieden, N.I., Cormier, J.T., Rancourt, D.E., and Kallos, M.S. (2007). Embryonic stem cells remain highly pluripotent following long term expansion as aggregates in suspension bioreactors. Journal of Biotechnology *129*, 421-432.

Zvetkova, I., Apedaile, A., Ramsahoye, B., Mermoud, J.E., Crompton, L.A., John, R., Feil, R., and Brockdorff, N. (2005). Global hypomethylation of the genome in XX embryonic stem cells. Nature Genetics *37*, 1274-1279.

Appendix 1

Differentially expressed genes between 2C-like cells and 2i cells

Gene Name	level in 2i cells	level in 2C-like	log2FoldChange	adj p-val
Prss23	0.00	6128.94	Inf	7.14E-05
Hs6st2	0.03	6241.06	17.84	5.64E-04
Gm8723	0.02	127.88	12.95	2.14E-06
Zfp352	0.12	604.95	12.36	2.91E-03
Trim75	0.04	123.58	11.71	4.31E-08
AC133095.2	0.05	141.36	11.53	3.94E-04
Pramef25	0.50	852.11	10.73	1.75E-03
Zscan4e	0.12	202.23	10.68	4.99E-07
AC168977.1	0.09	141.18	10.67	5.44E-06
Trim43a	7.82	12486.60	10.64	3.77E-02
Gm11487	0.08	110.78	10.45	3.29E-05
Drr1	0.05	73.34	10.42	2.79E-04
Gm8711	0.18	246.11	10.42	4.11E-06
Olfr881	0.10	126.97	10.38	5.18E-03
Fosl2	0.18	239.84	10.35	4.38E-02
B020004J07Rik	0.08	96.94	10.30	1.85E-05
Gm6803	0.69	808.99	10.20	3.72E-04
Zscan4b	0.31	352.71	10.16	3.21E-05
BC147527	0.17	194.73	10.15	4.60E-03
Usp17lc	0.62	636.01	10.01	3.90E-17
AA623943	0.22	212.87	9.95	1.03E-10
Usp17lb	0.59	578.63	9.94	1.12E-02
Cacna1s	0.34	322.82	9.89	4.78E-03
Gm13078	0.16	151.30	9.85	1.35E-07
Usp17ld	0.17	156.01	9.81	9.27E-08
Gm11544	0.16	138.47	9.73	6.65E-07
Gm2016	1.60	1289.77	9.66	5.67E-09
Zscan4f	0.78	580.98	9.55	8.67E-08
Gm11543	0.33	233.34	9.48	1.02E-10
Gm6489	0.61	406.24	9.38	6.98E-06
Gm20767	0.32	208.93	9.37	1.73E-09
Gm5698	1.02	671.58	9.36	7.46E-05
Gm5662	4.52	2858.87	9.31	8.43E-09
Gm21936	0.66	396.86	9.24	1.37E-11
Gm8300	2.00	1202.68	9.23	1.30E-10
BC080695	0.84	493.98	9.20	3.91E-04
Gm2075	0.23	130.08	9.14	6.20E-06
Tmem92	0.37	197.25	9.06	1.70E-08
Gm21319	0.48	250.83	9.04	9.81E-08
Abcb5	0.70	356.64	8.99	2.27E-03

Gm2022	1.29	643.76	8.97	5.37E-06
Gm21761	0.83	406.64	8.93	1.55E-04
Gm11546	0.42	199.64	8.91	2.39E-08
Usp17la	0.51	238.08	8.87	7.71E-06
Zscan4d	3.36	1526.44	8.83	3.43E-04
Gm2035	0.51	223.04	8.76	4.39E-09
Gm12794	0.36	155.01	8.75	1.70E-06
Gm2056	0.42	167.67	8.64	8.75E-07
Gm4027	0.76	252.95	8.38	9.38E-07
AF067061	0.89	281.40	8.31	3.86E-04
Gm8332	1.90	573.06	8.24	1.92E-06
Gm8994	3.30	868.95	8.04	4.81E-05
Zscan4c	8.59	2227.95	8.02	6.86E-05
BB287469	0.58	145.47	7.98	2.76E-05
Phf11a	1.10	246.67	7.81	3.05E-08
Gm5039	7.71	1446.34	7.55	8.38E-06
Calhm3	0.67	124.60	7.55	3.13E-04
Chit1	1.43	204.75	7.17	4.11E-06
AU019990	0.75	102.87	7.09	3.43E-03
B020031M17Rik	1.55	208.45	7.07	2.75E-03
Guca1a	2.66	349.99	7.04	8.32E-06
Gm5117	7.62	911.40	6.90	2.13E-02
Mdga2	0.53	61.89	6.86	4.43E-02
Gm16239	1.01	111.03	6.78	1.21E-02
Arg2	2.35	214.18	6.51	1.23E-03
Pdgfrl	5.11	414.93	6.34	1.74E-03
Scd3	7.00	518.04	6.21	6.29E-08
Gm16892	5.43	395.12	6.19	2.52E-02
Gm10800	259.06	16352.73	5.98	8.87E-03
Pdlim3	2.27	138.71	5.93	5.59E-03
Aqp9	2.75	158.22	5.85	1.38E-03
Limch1	6.50	350.07	5.75	8.10E-03
Tmem132c	7.29	383.23	5.72	1.38E-04
Ccser1	5.90	308.61	5.71	2.07E-06
Zfp560	8.21	426.55	5.70	4.47E-05
Arhgef26	27.17	1407.23	5.69	9.65E-06
Gm21738	28.03	1392.64	5.63	9.25E-07
Ctf2	4.97	242.04	5.61	1.49E-02
Antxr1	4.93	238.82	5.60	3.94E-04
Sh3kbp1	6.22	285.56	5.52	1.20E-04
Gm10717	8.16	354.69	5.44	8.16E-06
Gm26870	9.89	426.04	5.43	1.74E-06
Neto2	9.63	383.95	5.32	3.41E-03
Gm11168	5.76	228.28	5.31	1.95E-04
P4ha2	12.86	447.34	5.12	1.68E-05
Gm10722	6.76	224.68	5.06	3.94E-04

Gm5435	9.34	290.93	4.96	6.03E-03
Slc35e3	18.07	466.91	4.69	1.68E-05
Scamp1	46.94	1210.95	4.69	1.32E-05
Gm10801	10.75	272.71	4.66	1.82E-04
ENSMUSG00000060393	40.54	961.84	4.57	6.87E-04
Uap1	87.02	2044.20	4.55	2.42E-09
Gm13228	51.98	1180.16	4.50	8.95E-06
Gm12183	125.60	2688.09	4.42	7.46E-05
Mcm9	28.57	601.03	4.39	1.97E-05
Dennd4c	75.34	1559.44	4.37	2.50E-07
Gm13226	8.36	165.86	4.31	1.31E-02
Dhtkd1	105.29	1980.85	4.23	2.71E-13
Ercc4	163.70	3021.31	4.21	4.08E-12
Fbxo15	1568.23	28808.76	4.20	9.66E-04
2010315B03Rik	7.80	129.72	4.06	4.75E-02
Prex2	79.38	1240.22	3.97	4.11E-06
Lgals4	18.93	291.06	3.94	5.59E-03
Ric3	27.27	417.63	3.94	3.90E-02
Cwc22	279.65	4100.02	3.87	1.39E-06
Nelfa	270.81	3944.80	3.86	2.52E-08
Tanc2	17.52	249.84	3.83	2.60E-03
Arsk	45.17	606.24	3.75	3.47E-04
Gm13622	56.13	685.20	3.61	7.30E-08
Fam234b	19.02	215.15	3.50	2.23E-02
Pemt	15.81	167.33	3.40	4.38E-02
Tbc1d23	126.18	1331.81	3.40	3.60E-07
Dnajb14	70.39	728.76	3.37	2.68E-06
Dcbld1	113.27	1159.11	3.36	1.27E-07
Glrx2	95.46	973.31	3.35	2.96E-07
1700025G04Rik	22.56	216.97	3.27	3.43E-02
Fundc1	121.03	1130.11	3.22	2.11E-04
Mbd5	20.93	187.38	3.16	4.14E-02
Ppm1a	295.98	2641.47	3.16	3.74E-06
Rxra	22.71	197.87	3.12	3.47E-02
Tsen2	155.44	1305.13	3.07	2.07E-05
Bola1	106.12	865.28	3.03	4.82E-08
Cab39	222.01	1792.52	3.01	9.25E-07
Aqr	389.52	3137.48	3.01	1.03E-10
ENSMUSG00000095908	25.76	207.48	3.01	3.60E-02
Zfp809	148.32	1177.65	2.99	6.46E-07
Sdhaf3	48.71	375.25	2.95	1.83E-03
Zfp386	108.69	818.30	2.91	1.40E-02
Bud13	162.42	1159.31	2.84	3.01E-03
Cbl	120.50	850.29	2.82	1.68E-04
Robo1	53.08	372.62	2.81	2.35E-03
Dst	954.49	6429.51	2.75	4.78E-11

Ddit4l	43.07	290.03	2.75	1.31E-02
Trak2	53.03	353.68	2.74	3.95E-03
Rad51b	39.28	258.42	2.72	2.49E-02
Ncoa2	169.25	1089.28	2.69	2.07E-05
Micu1	87.51	542.30	2.63	1.35E-03
Sgms1	209.59	1294.63	2.63	1.75E-05
Cldn12	38.47	236.50	2.62	4.43E-02
Pcnxl4	86.69	528.92	2.61	2.28E-04
C1d	401.64	2395.59	2.58	1.25E-15
Sord	70.63	411.05	2.54	4.38E-02
Coil	676.12	3894.99	2.53	6.49E-04
Lonp2	264.29	1449.56	2.46	4.99E-07
Cand1	482.56	2589.46	2.42	1.92E-06
Zfp119b	69.33	354.82	2.36	3.47E-02
Dock9	89.90	451.82	2.33	1.51E-02
2210409E12Rik	111.38	551.79	2.31	6.00E-04
Slc7a6os	66.22	309.70	2.23	3.18E-02
Neo1	204.73	948.54	2.21	1.47E-03
Gpbp1l1	263.67	1213.79	2.20	2.94E-07
Arid4a	335.22	1519.45	2.18	1.67E-03
Katnbl1	137.05	618.46	2.17	5.99E-04
Daam1	87.49	394.77	2.17	1.72E-02
Clp1	416.47	1832.57	2.14	4.26E-07
Pcca	88.10	376.89	2.10	2.77E-02
Zfp217	101.08	422.16	2.06	1.62E-02
Ticrr	563.52	2344.40	2.06	5.52E-03
Prkaa1	126.09	520.29	2.04	3.90E-02
Ppig	838.86	3376.17	2.01	3.91E-10
Bach1	118.92	476.26	2.00	2.10E-02
Rimklb	266.03	1060.79	2.00	2.47E-03
2810004N23Rik	278.95	1070.40	1.94	1.83E-05
Rbbp6	542.32	2067.94	1.93	6.81E-06
Cep57l1	240.55	912.16	1.92	8.91E-04
Akap13	151.29	570.55	1.92	1.88E-02
Diaph3	132.83	500.59	1.91	9.78E-03
Utp23	100.59	378.44	1.91	3.39E-02
Rbm25	3500.92	13085.13	1.90	8.77E-21
Zfp936	329.76	1212.27	1.88	1.24E-03
Zfp516	129.93	472.93	1.86	1.58E-02
Spdl1	184.49	661.98	1.84	2.58E-02
Snapc3	183.73	646.81	1.82	3.14E-03
Arnt	134.08	459.71	1.78	4.14E-02
Ctr9	297.36	994.12	1.74	5.99E-04
Iars	1470.43	4875.66	1.73	4.66E-06
Wdr70	128.85	424.10	1.72	4.73E-02
Ppp1r8	141.75	462.93	1.71	3.14E-02

Pnkd	150.27	482.14	1.68	3.23E-02
Rlf	186.31	587.43	1.66	4.22E-02
PlSD	297.47	928.34	1.64	6.03E-03
Gtf3c3	207.22	631.64	1.61	2.99E-02
Nup205	303.52	909.53	1.58	1.11E-02
Atf2	466.39	1396.44	1.58	1.49E-03
Mreg	442.39	1314.81	1.57	1.40E-03
Snw1	728.61	2143.80	1.56	9.68E-07
Dppa2	226.50	653.99	1.53	1.53E-02
Scfd1	445.51	1249.17	1.49	2.13E-02
Pnp	238.15	663.44	1.48	1.64E-02
Zcchc17	218.56	607.37	1.47	2.95E-02
Klf3	227.96	631.80	1.47	2.44E-02
Cstf2t	206.47	562.88	1.45	4.65E-02
Avpi1	386.02	1051.98	1.45	1.83E-03
D230025D16Rik	817.24	2225.67	1.45	7.30E-03
Papolg	293.78	794.44	1.44	3.07E-02
Gtf2b	330.34	888.38	1.43	1.10E-02
Ube2t	332.01	874.54	1.40	8.00E-03
Cenpe	755.01	1912.56	1.34	5.59E-03
Srsf5	1046.68	2626.62	1.33	4.64E-06
Haus3	298.99	746.25	1.32	2.52E-02
Mtf2	2331.48	5765.01	1.31	2.64E-05
Gm26917	1840.84	4499.22	1.29	1.74E-03
Fnip1	265.62	649.11	1.29	4.96E-02
Klf9	384.58	939.18	1.29	1.47E-02
Gtf2h2	346.67	817.05	1.24	4.60E-02
Arih2	559.83	1286.21	1.20	4.96E-02
Wtap	475.66	1090.61	1.20	1.21E-02
Utp3	498.47	1112.17	1.16	1.40E-02
Agpat5	374.00	828.45	1.15	4.89E-02
Ccnf	998.52	2192.06	1.13	3.17E-02
Map1b	1111.65	2405.68	1.11	2.13E-02
Hmmr	732.19	1569.15	1.10	4.43E-02
Luc7l3	661.04	1368.42	1.05	2.13E-02
Btg1	560.61	1152.38	1.04	2.99E-02
Arid1a	641.25	1315.52	1.04	1.79E-02
Triml2	780.84	1592.99	1.03	2.49E-02
Rsrc2	751.37	1507.71	1.00	1.49E-02
Tacc3	1049.95	2065.64	0.98	3.47E-02
Cirh1a	1481.08	2902.23	0.97	3.01E-03
Thoc2	1023.75	2002.21	0.97	1.72E-02
Rplp0	6898.18	13421.12	0.96	1.63E-06
Nmd3	1249.74	2373.43	0.93	4.56E-02
Gm9625	966.08	1795.59	0.89	2.52E-02
Rif1	3165.71	5854.94	0.89	2.33E-04

Trim28	1715.70	3149.49	0.88	3.76E-03
Cul5	1347.83	2468.83	0.87	3.80E-02
Gm8730	1718.89	3105.61	0.85	5.97E-03
Kif20b	1161.66	2092.99	0.85	2.39E-02
2810474019Rik	1902.65	3419.75	0.85	3.04E-02
Dnttip2	1088.83	1905.36	0.81	4.38E-02
Smc4	1211.35	2116.97	0.81	3.47E-02
mt-Rnr1	3179.73	5483.37	0.79	7.63E-03
Gnl3	1392.54	2380.12	0.77	3.96E-02
Eif4a2	4343.62	6859.50	0.66	2.44E-02
Sept2	8965.87	5590.07	-0.68	2.64E-02
Nedd4	3371.49	1973.56	-0.77	3.20E-02
Actr2	2977.84	1705.15	-0.80	3.29E-02
Paics	5388.05	3083.72	-0.81	9.27E-03
Srpk1	1870.29	1017.49	-0.88	4.38E-02
Myl12b	1241.76	614.64	-1.01	3.77E-02
Ywhag	895.49	410.67	-1.12	4.92E-02
Vdac1	2577.24	1113.31	-1.21	1.91E-04
Aamp	1004.15	417.62	-1.27	1.03E-02
Gdf3	1260.90	502.62	-1.33	1.36E-02
Stmn2	835.90	330.22	-1.34	1.94E-02
Acadm	1281.21	499.48	-1.36	1.49E-02
Apobec3	597.04	222.42	-1.42	4.52E-02
Chchd4	735.24	270.24	-1.44	1.50E-02
Gm11223	660.76	231.30	-1.51	1.10E-02
Idh2	1344.63	454.15	-1.57	1.11E-02
Tpd52	888.19	292.62	-1.60	1.02E-03
Hmces	520.32	170.75	-1.61	1.79E-02
Slc25a13	562.12	178.44	-1.66	1.21E-02
Dsg2	729.58	216.19	-1.75	2.62E-02
Lypla1	1653.09	468.62	-1.82	2.18E-05
Tmem245	479.67	132.27	-1.86	1.60E-02
Acaa2	342.44	90.92	-1.91	3.30E-02
Bcat2	382.21	94.23	-2.02	9.21E-03
Prpf6	535.01	131.76	-2.02	2.02E-02
Anxa4	472.15	107.85	-2.13	4.19E-03
H2-M6-ps	299.53	58.93	-2.35	8.45E-03
Rc3h1	199.94	35.67	-2.49	4.65E-02
Ccdc141	329.71	56.21	-2.55	4.51E-02
Bscl2	206.23	31.46	-2.71	4.48E-02
Laptm5	329.97	50.12	-2.72	1.36E-02
Zfp157	265.47	36.49	-2.86	2.49E-02
Plagl1	338.73	44.95	-2.91	3.23E-03
Prkd3	195.26	25.41	-2.94	1.17E-02
Armcx1	133.54	15.61	-3.10	4.96E-02
Lefty2	425.10	47.77	-3.15	4.38E-02

Zfp553	152.65	16.37	-3.22	2.13E-02
Ddx58	201.28	20.34	-3.31	2.44E-02
Ormdl1	172.51	16.43	-3.39	2.13E-02
Gatsl3	346.06	31.75	-3.45	2.82E-04
Sulf1	151.42	12.59	-3.59	2.13E-02
Crtap	110.35	8.96	-3.62	4.38E-02
Fetub	143.64	10.85	-3.73	8.11E-03
Rhpn2	127.34	8.05	-3.98	2.33E-02
Msantd3	123.04	7.19	-4.10	9.27E-03
Angptl4	134.41	3.24	-5.37	2.50E-02
Ccdc136	63.27	1.43	-5.47	4.65E-02
Fhl1	108.58	1.82	-5.90	1.21E-02
Kdm4a	57.22	0.94	-5.93	4.95E-02
Dtx1	77.76	0.83	-6.55	2.32E-02
Rnf41	72.57	0.75	-6.60	5.79E-03
Zfp36l2	78.67	0.74	-6.73	2.60E-03
Ackr3	83.98	0.65	-7.00	1.38E-02
Senp8	64.34	0.38	-7.40	6.08E-03
2900011O08Rik	128.86	0.38	-8.41	1.63E-06
Tmem14a	44.27	0.10	-8.84	2.94E-02
Slc52a3	84.05	0.00	-Inf	2.06E-04
Qpctl	46.74	0.00	-Inf	1.50E-02
Mylip	54.75	0.00	-Inf	2.38E-03
March2	43.88	0.00	-Inf	1.49E-02