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).

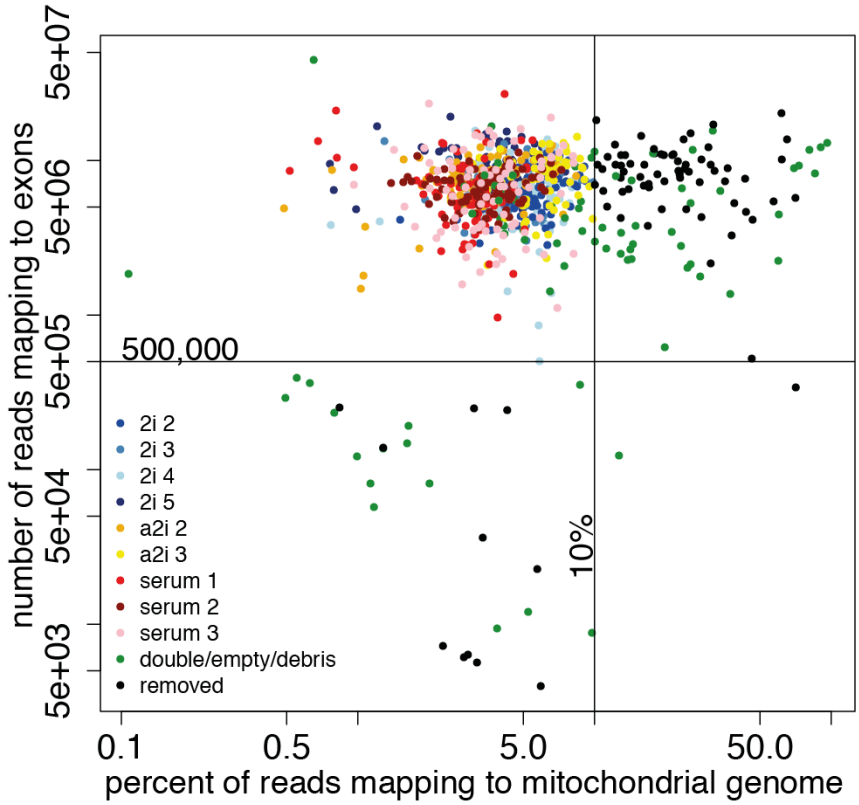

Figure 3.2 Quality control of cells
Quality control metrics were the number of reads mapping to exons (y axis), and the proportion of reads mapped to mitochondrial genes (x axis). Lines represent the thresholds used. Green points represent cells excluded upon microscopic examination of the C1 chip and 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.

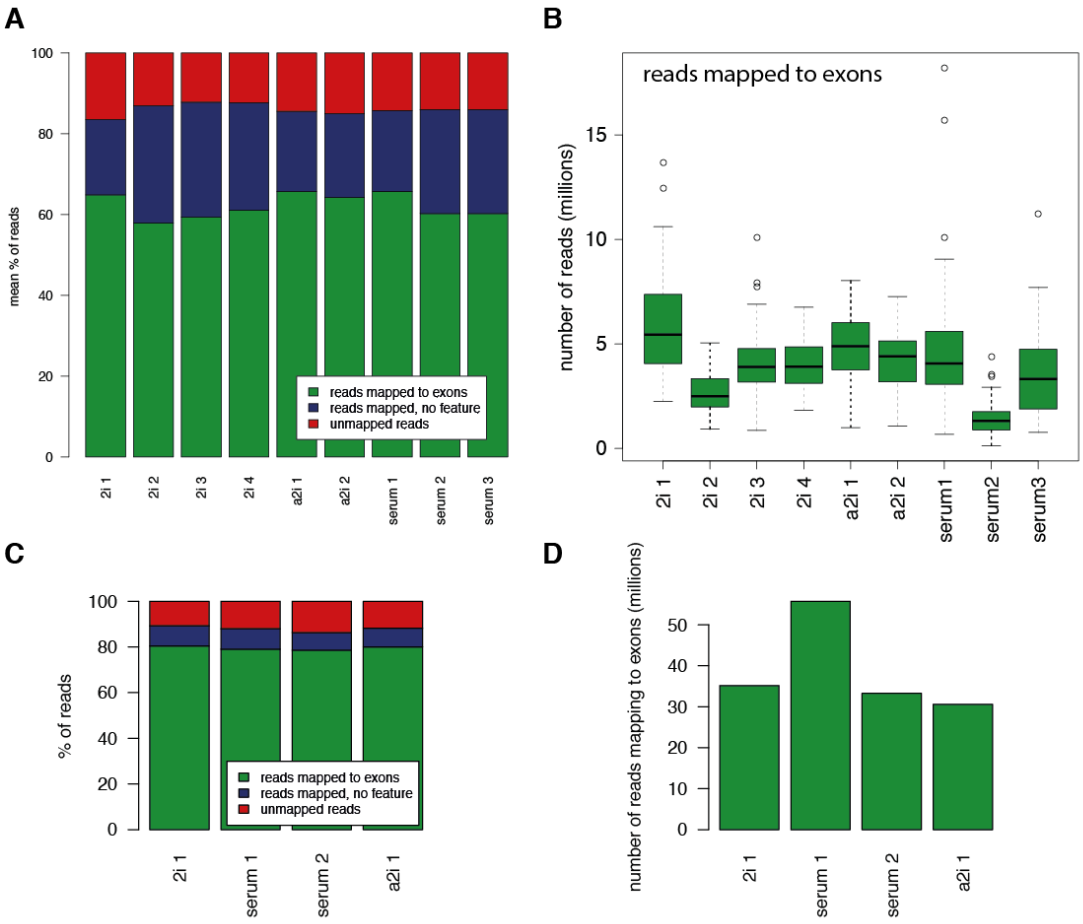

Figure 3.3 Mapping statistics for each cell for each sample.
The mean percentage of reads mapping to Ensembl exons (green), to the genome outside Ensembl annotated regions (blue) and unmapped reads (red) for each of nine experiments. (A) and (B) show results for single cell experiments while (C) and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

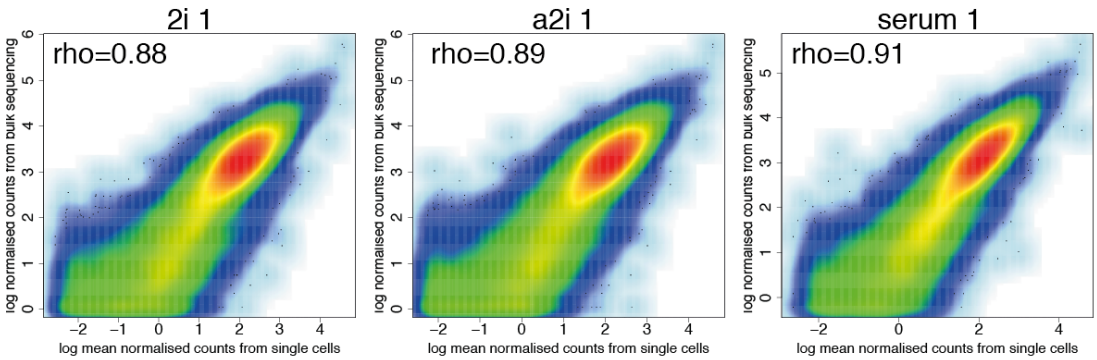

Figure 3.4 Comparison of gene expression levels between bulk and single cells.
2D kernel density estimation of scatter plot between expression level in bulk experiment and mean of gene expression from single cells in each condition. Value of Spearman rank correlation coefficient (rho) between bulk and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

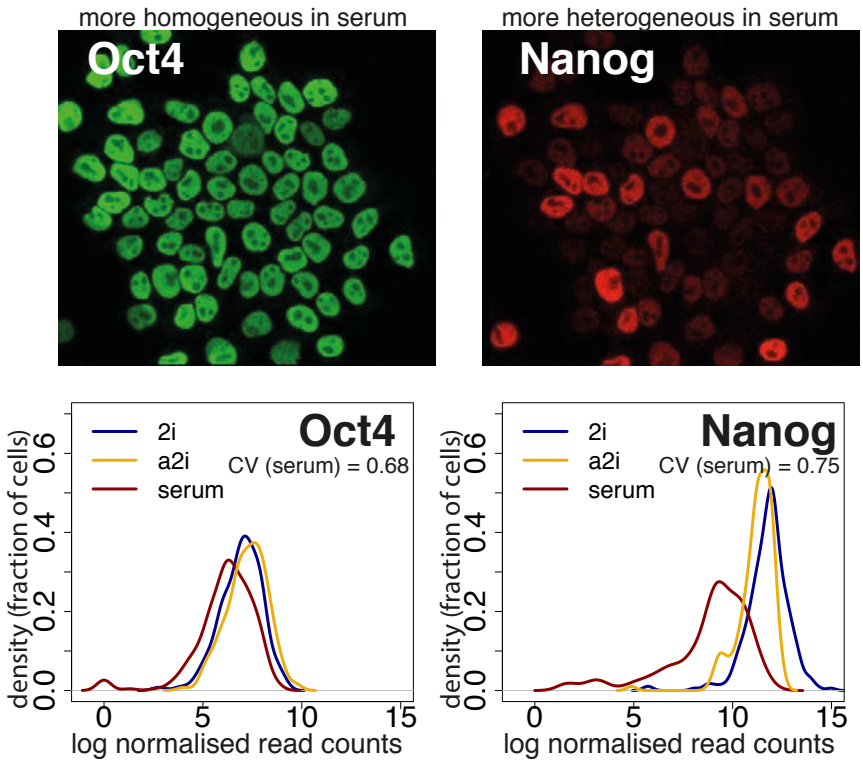

Figure 3.5 Variability of expression of Oct4 and Nanog.
Microscopy pictures showing fluorescently labelled Oct4 and Nanog are from Roeder and Radtke, 2009, and plots below show expression heterogeneity of Oct4 and Nanog in three culture conditions plotted using single cell mRNAseq data.

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.

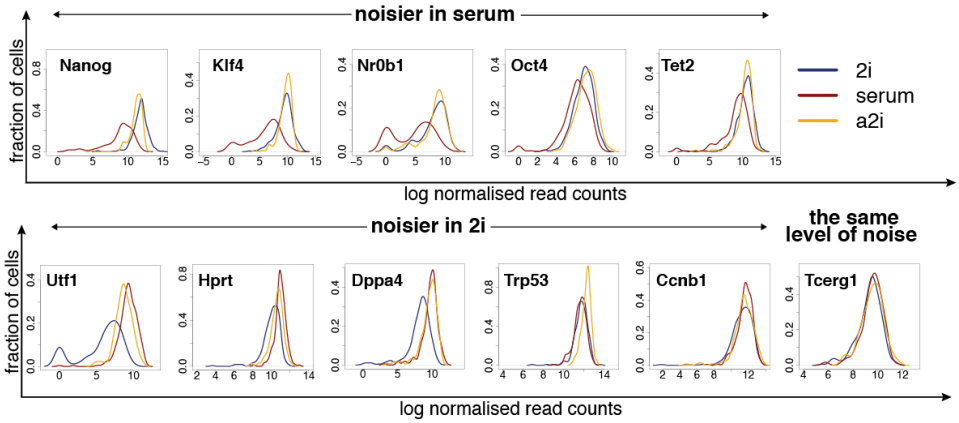

Figure 3.6 Gene expression distributions across cells

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).

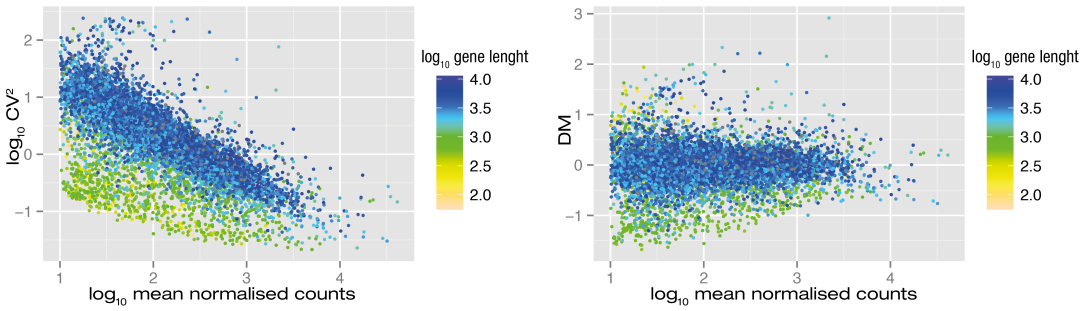

Figure 3.7 Gene expression variability measured with coefficient of variation (CV) and distance to the median (DM)

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.

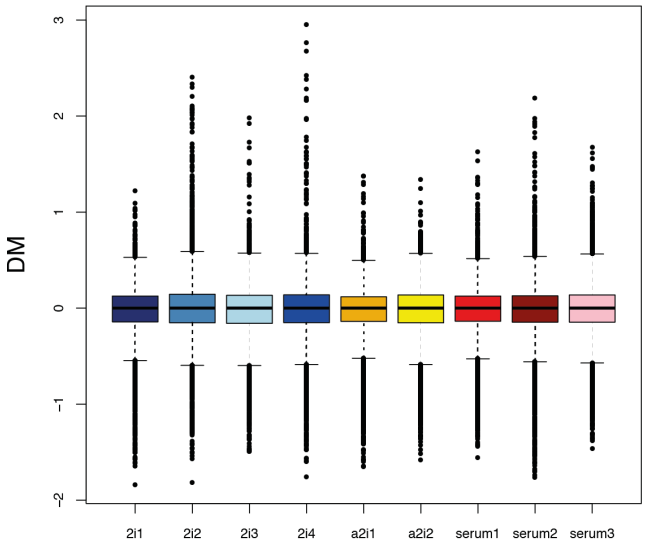

Figure 3.8 Gene expression variability across cells in different conditions measured with DM

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).

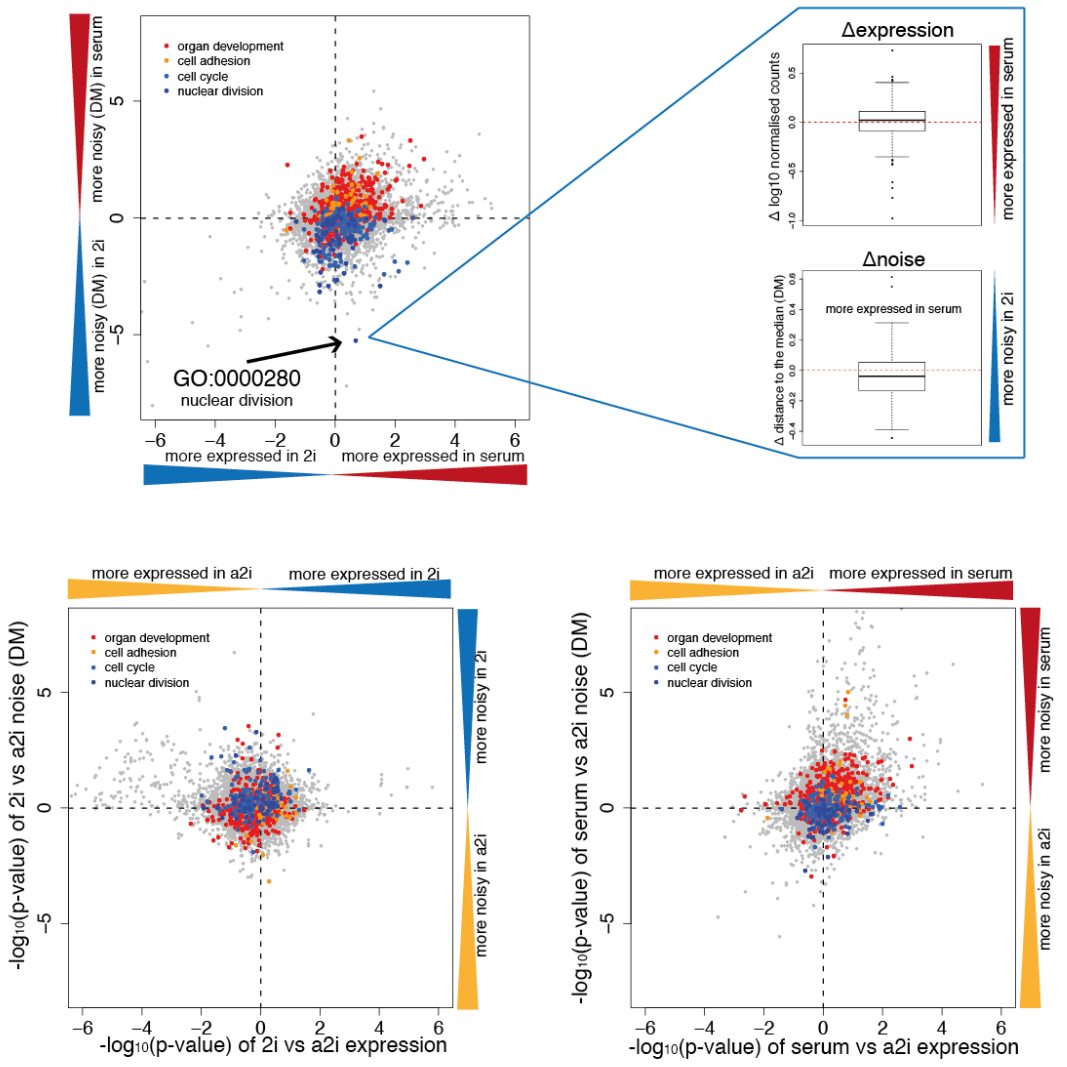

Figure 3.9 Gene expression heterogeneity of functional categories of genes
Comparison of the levels of gene expression and noise for gene ontology (GO) categories between the culture conditions (excluding 2i replicates containing 2C-like cells). The logarithm (log10) of P-values from two-sided paired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

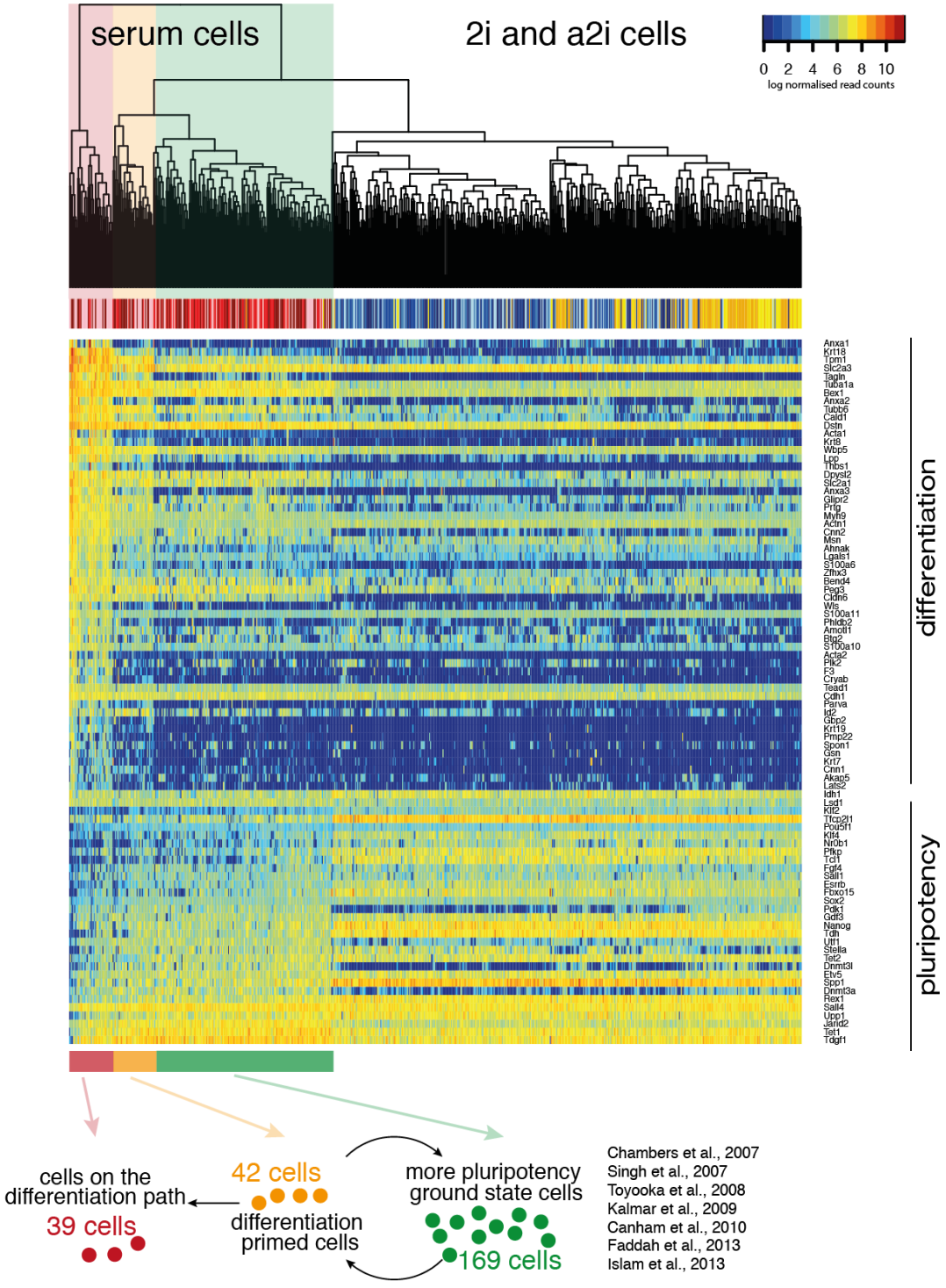

Figure 3.10 Subpopulation structure of cells cultured in serum

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).

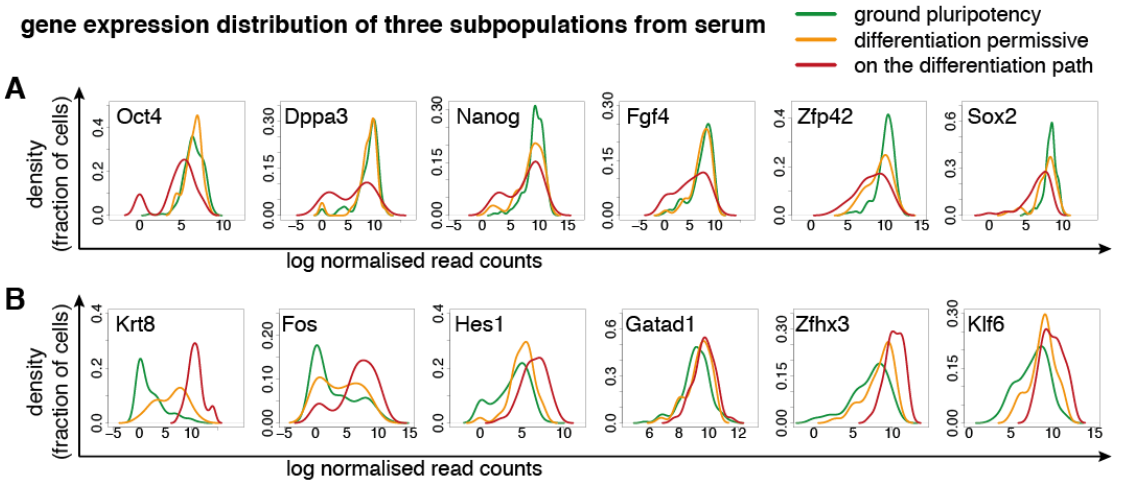

Figure 3.11 Gene expression differences between three clusters of cells in serum
Gene expression distributions of genes that become downregulated (A) and upregulated (B) upon differentiation. Expression is shown as $\log_2$ size factor normalized counts. Oct4 expression is similar in cells closer to 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.

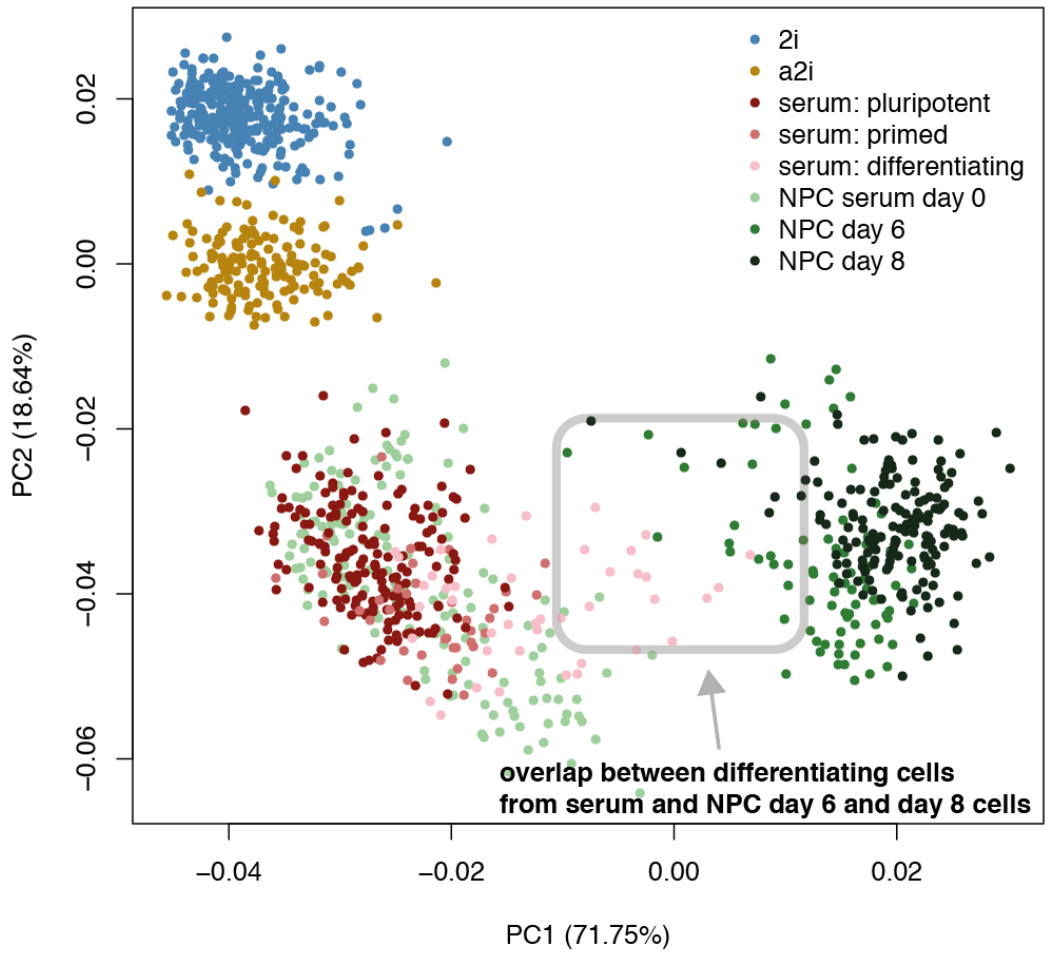

Figure 3.12 Principal component analysis of expression data from serum and cells progressing towards NPC fate.

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.

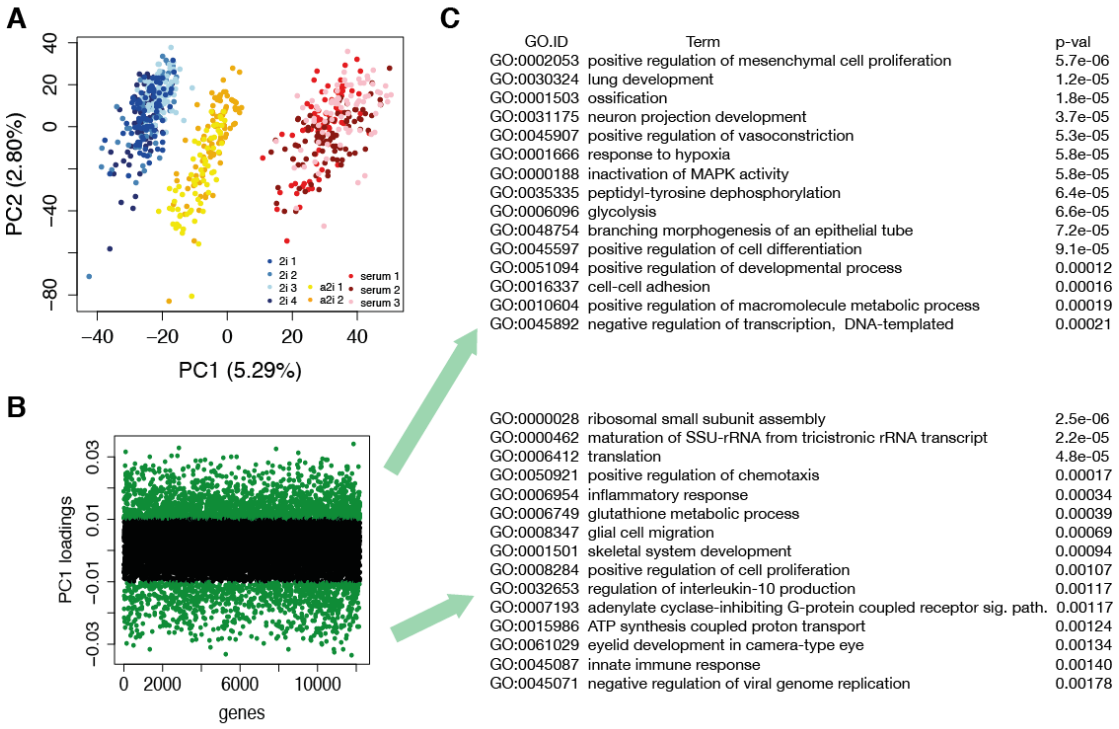

Figure 3.13 Clustering of mESCs grown in serum, 2i and a2i media
All cells (n=704) grown in the three different culture conditions are projected onto the first two principal components. All genes with mean normalized read counts larger than 10 were considered and principal component analysis (PCA) was 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).

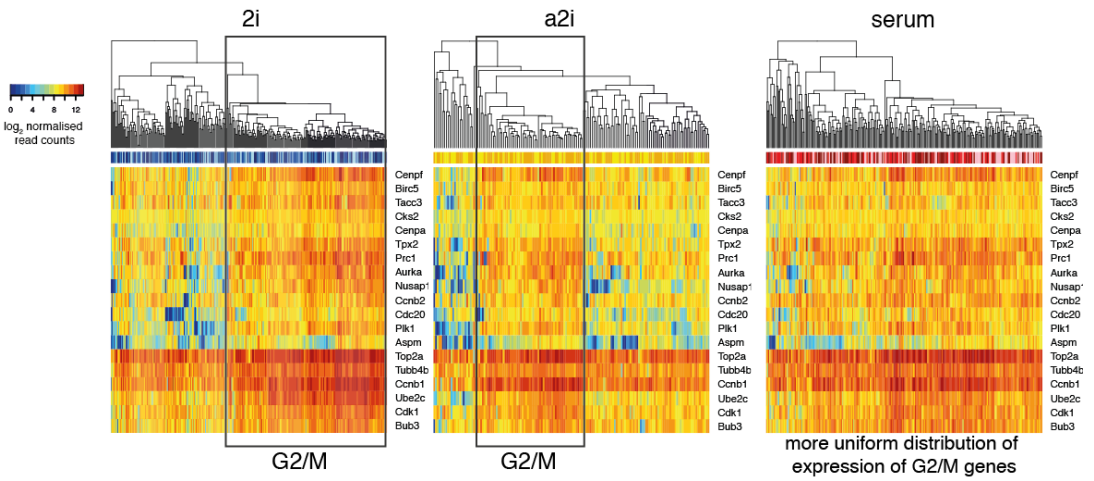

Figure 3.14 Cell cycle gene heterogeneity and cell cycle phase assignment
Heat maps showing the expression of cell cycle related genes in serum, 2i and a2i, with a distinct separation into G1/S versus G2/M cells in 2i and a2i, with less distinction between individual cells in serum.

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).

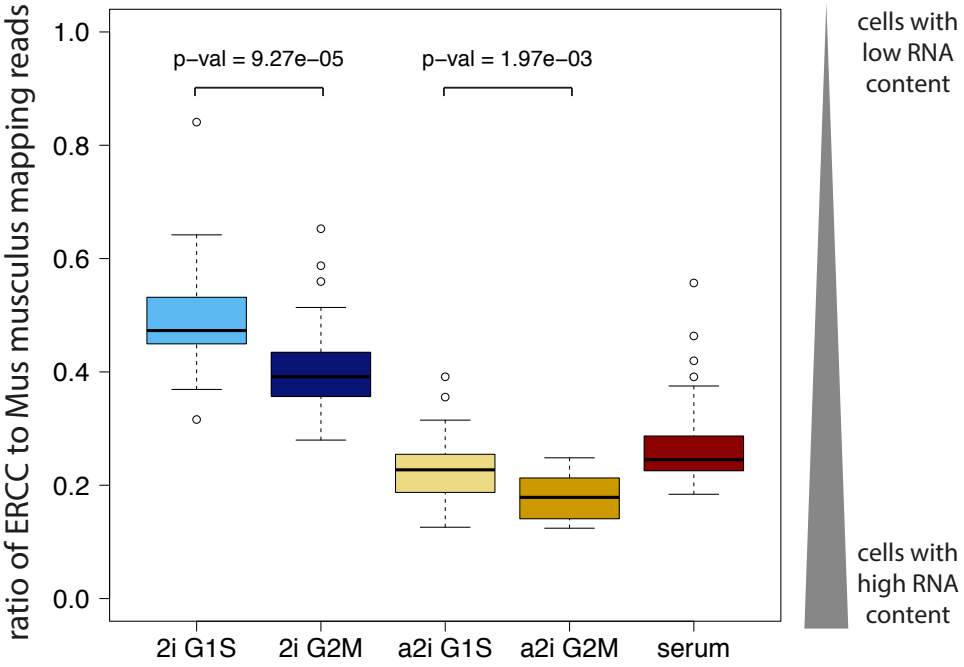

Figure 3.15 mRNA content in cells at different cell cycle stages
Comparison of mRNA content in cells using ratio of reads mapping to ERCCs (constant number of molecules spiked in in three conditions) to all exon mapped reads.

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).

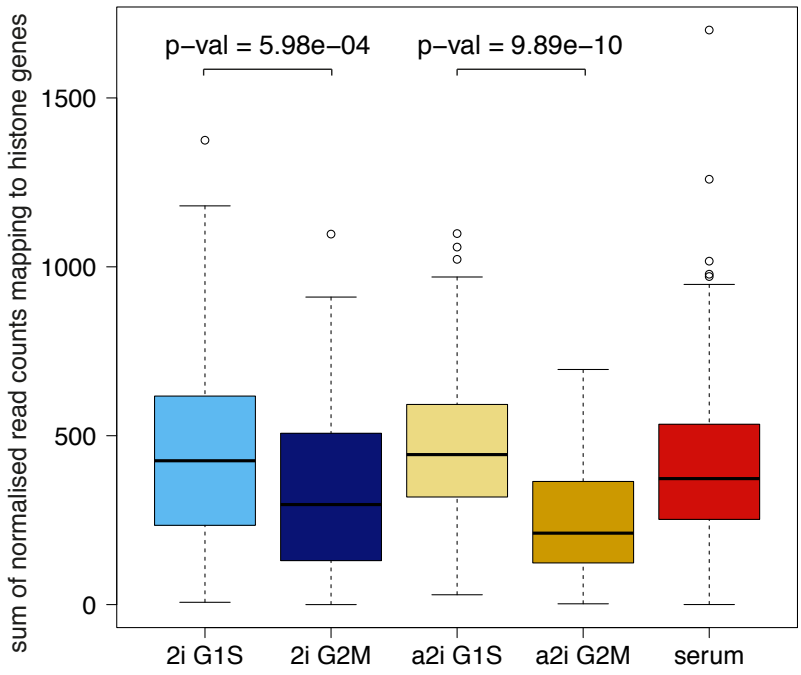

Figure 3.16 Histone mRNA expression in cells at different cell cycle stages
Comparison of histone mRNA content in cells from different cell cycle stages across culture conditions.

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.

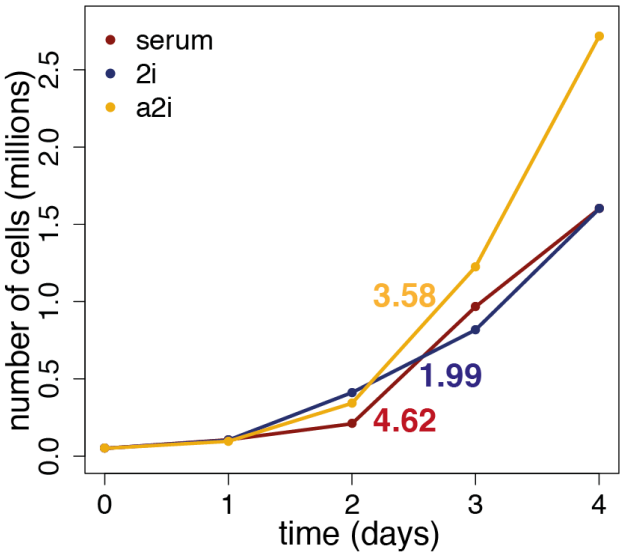

Figure 3.17 Growth kinetics of cells in three culture conditions.
Numbers shown are how many times cells grew between second and third day of culture, i.e. when cells were harvested for scRNA-seq experiment. At this point in culture cells cultured in serum grew slowest.

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.

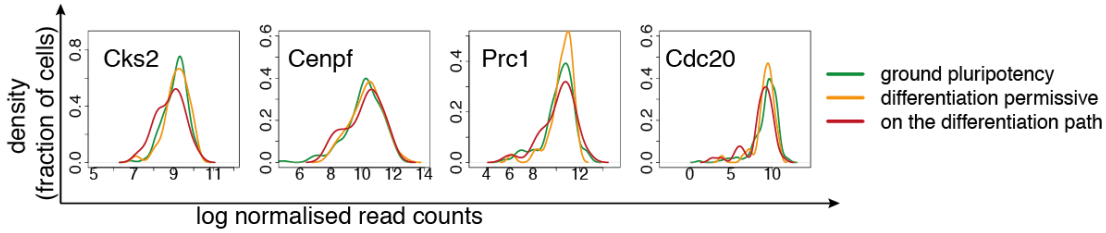

Figure 3.18 Gene expression distributions of cell cycle genes in subpopulation of cells cultured in serum.
Plots show distribution of cell cycle gene expression in cells from three subpopulations from serum. Cells that are on the differentiation path (red) are more heterogeneous than cells that are in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

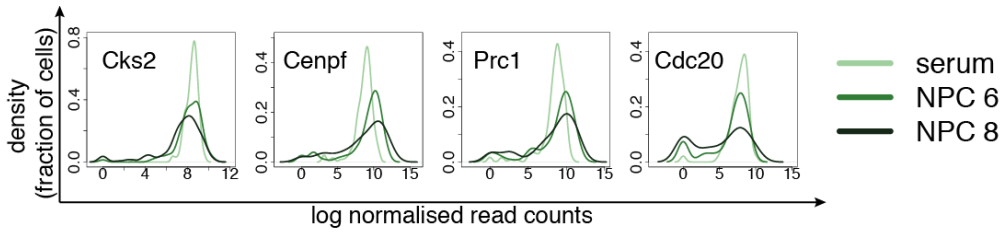

Figure 3.19 Gene expression distributions of cell cycle genes in cells from NPC differentiation time course

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).

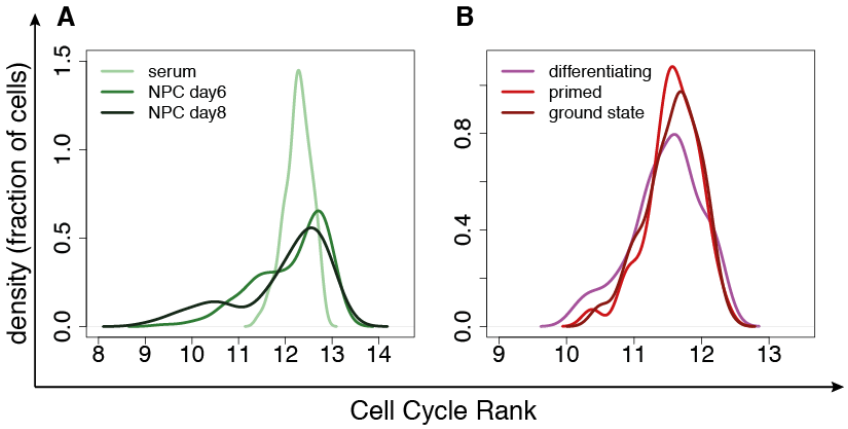

Figure 3.20 Cell Cycle Rank distribution
Distribution of Cell Cycle Ranks for (A) cells from NPC differentiation time course and (B) subpopulation of cells cultured in serum.

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.