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.

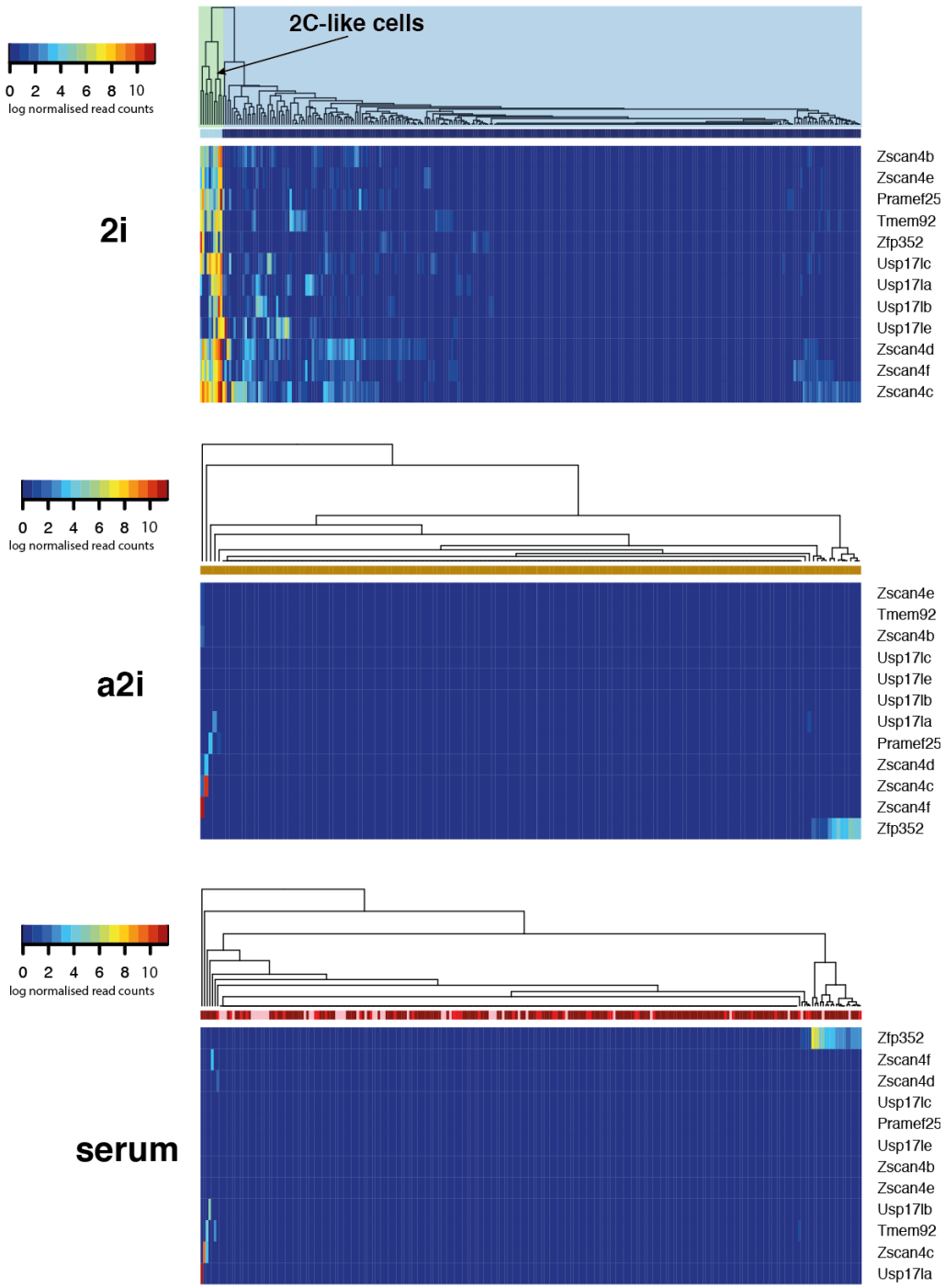

Figure 4.2 Identification of 2C-like cells

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).

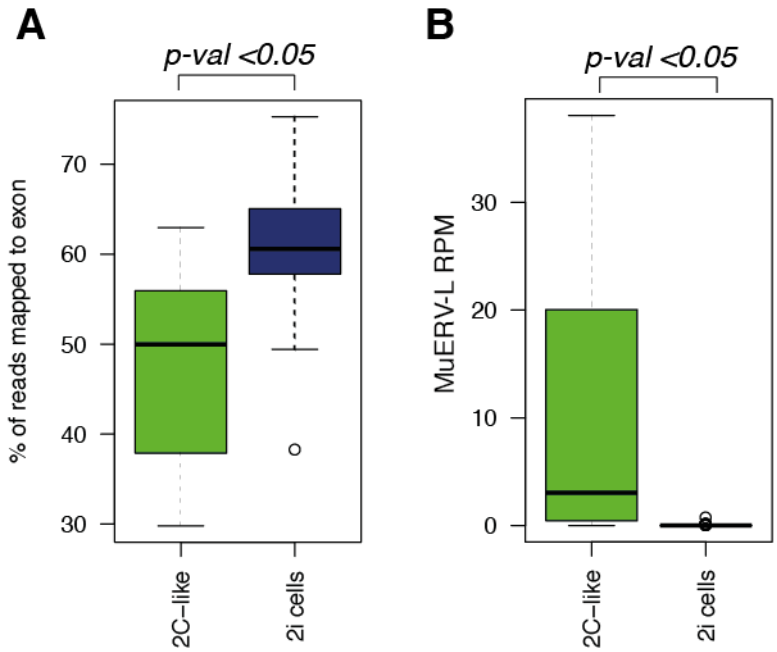

Figure 4.3 Exon mapping reads and endogenous retrovirus expression
(A) Boxplot showing % of reads mapping to the exons in both subpopulations of cells in 2i. P-value was calculated using Wilcoxon test. (B) Boxplot showing RPM (reads per million) mapping to the MuERV-L retrovirus in both subpopulations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

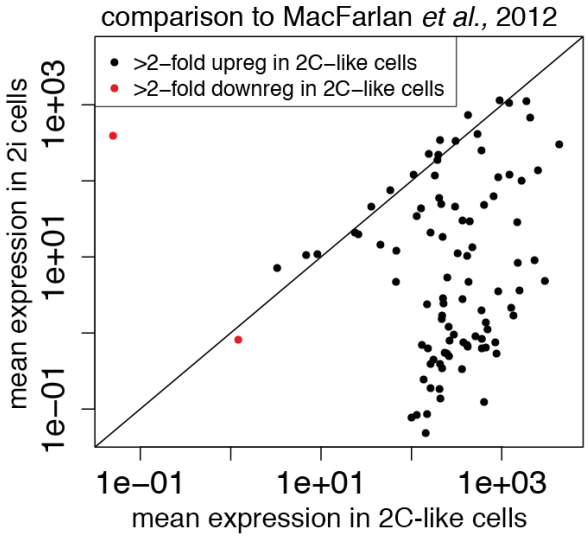

Figure 4.4 Comparison of differential expression results to MacFarlan
Mean expression of genes reported to be at least 2-fold upregulated or downregulated in 2C-like cells (Macfarlan et al., 2012) in cells that I identified as 2C-like cells and in the remaining 2i cells.

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).

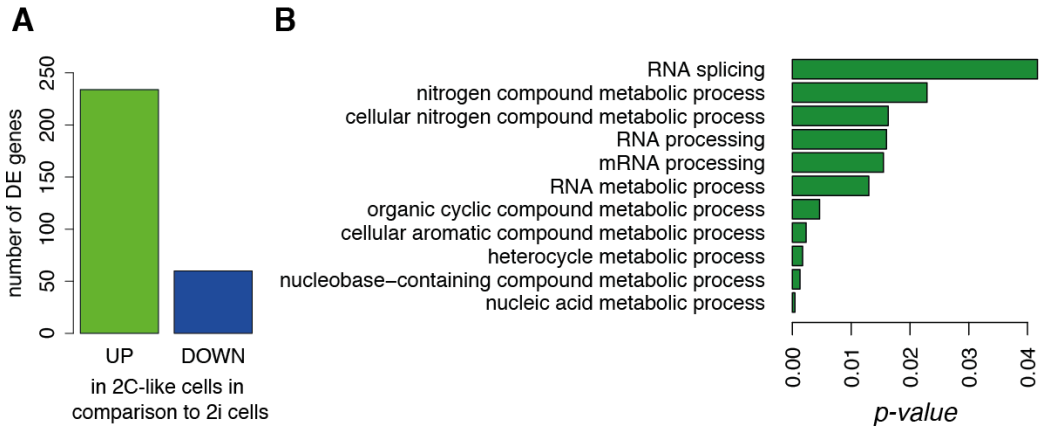

Figure 4.5 Differential expression analysis between 2C-like cells and 2i cells
(A) Bar plot showing the number of significantly (DESeq, adjusted *p*-val < 0.05) upregulated and downregulated genes in 2C-like cells. (B) Plot shows most significantly enriched gene ontology terms. *p*-value is corrected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

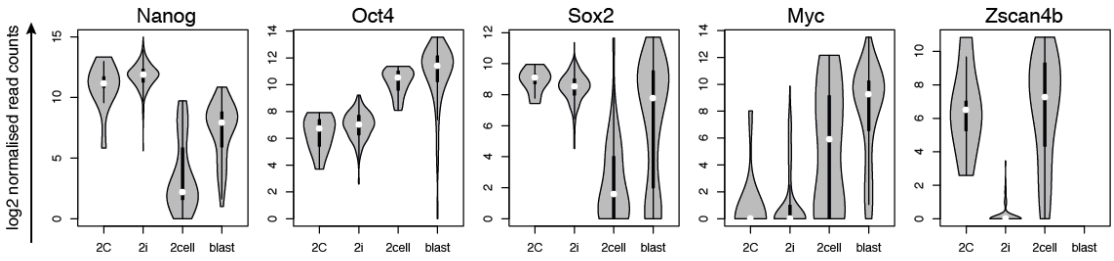

Figure 4.6 Key pluripotency genes in 2C-like cells, 2i cultured cells and in cells from *in vivo* embryo

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.

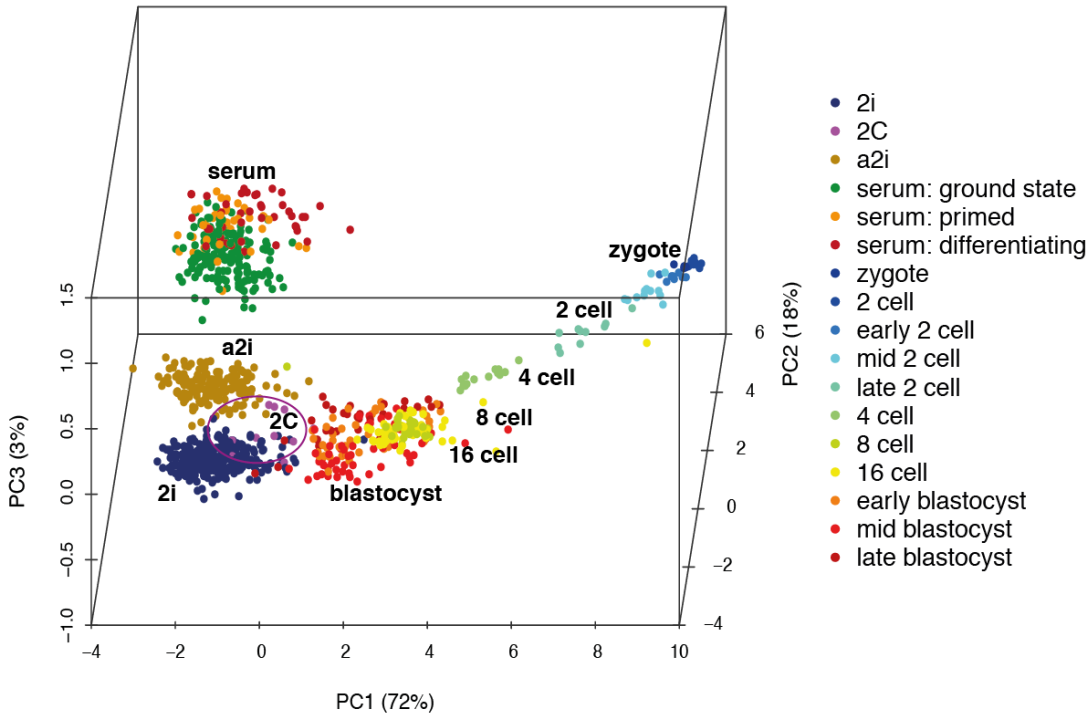

Figure 4.7 Comparison of mESC to cells from early embryo development
PCA loading plot of the Spearman's rank correlation coefficients from mESCs and single cells of mouse preimplantation embryos (Deng et al., 2014), showing the mapping of mESCs in mouse development stages. The cells are visualized by 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.