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).

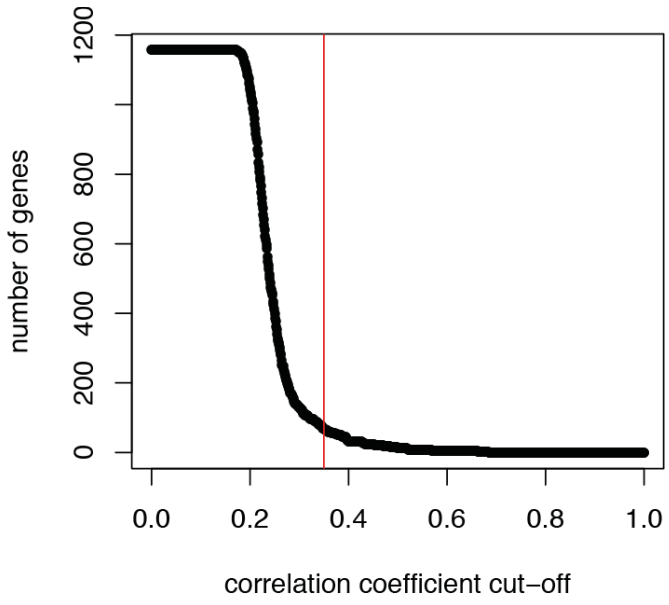


Figure 5.1 Correlation coefficient cut-off.
Plot shows the number of genes that correlate with at least one other gene above Spearman rank correlation coefficient value.

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

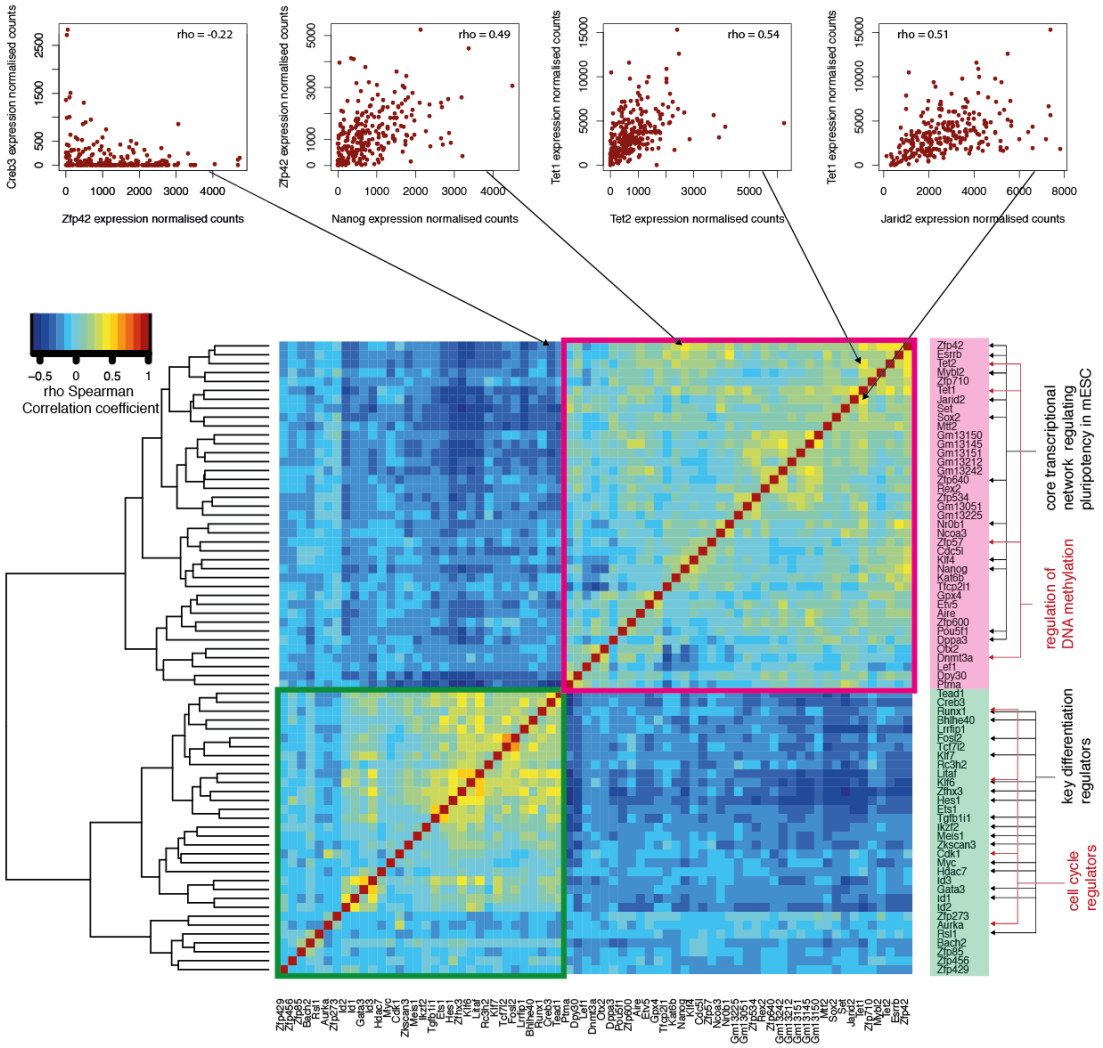


Figure 5.2 Spearman correlation matrix of transcription factors and key pluripotency genes.

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).

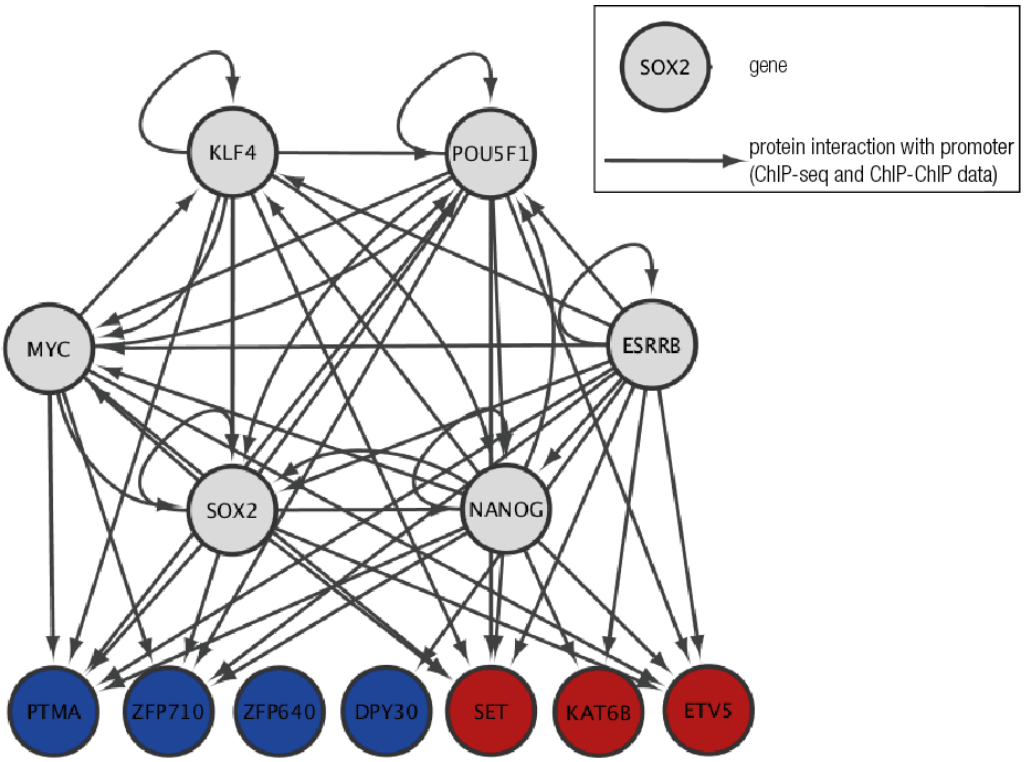


Figure 5.3 Pluripotency network
Network showing known interactions of core pluripotency factors with the novel candidate genes. Data obtained from ChIP-seq and ChIP-ChIP experiments from ESCAPE database.

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).

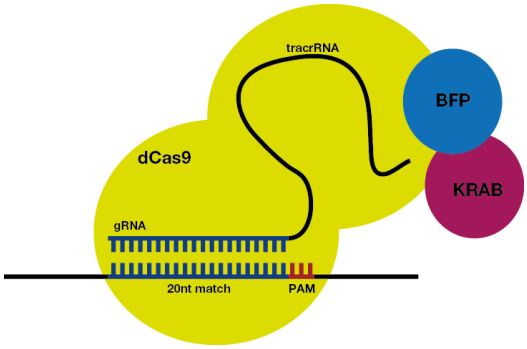


Figure 5.4 Schematic of CRISPRi

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.

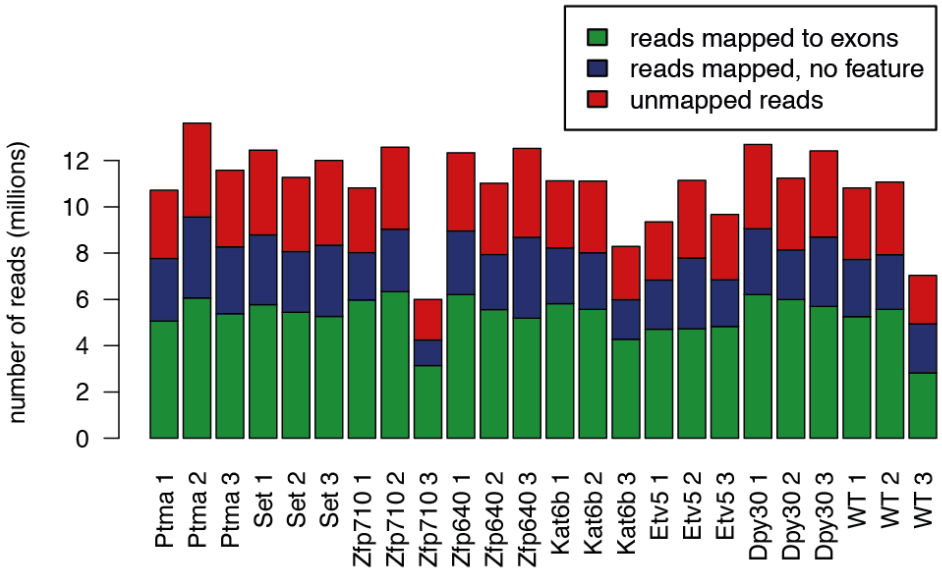


Figure 5.5 Mapping statistics
Barplot shows how many reads map to exons, mouse genome and how many do not map for all samples in three replicates.

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.

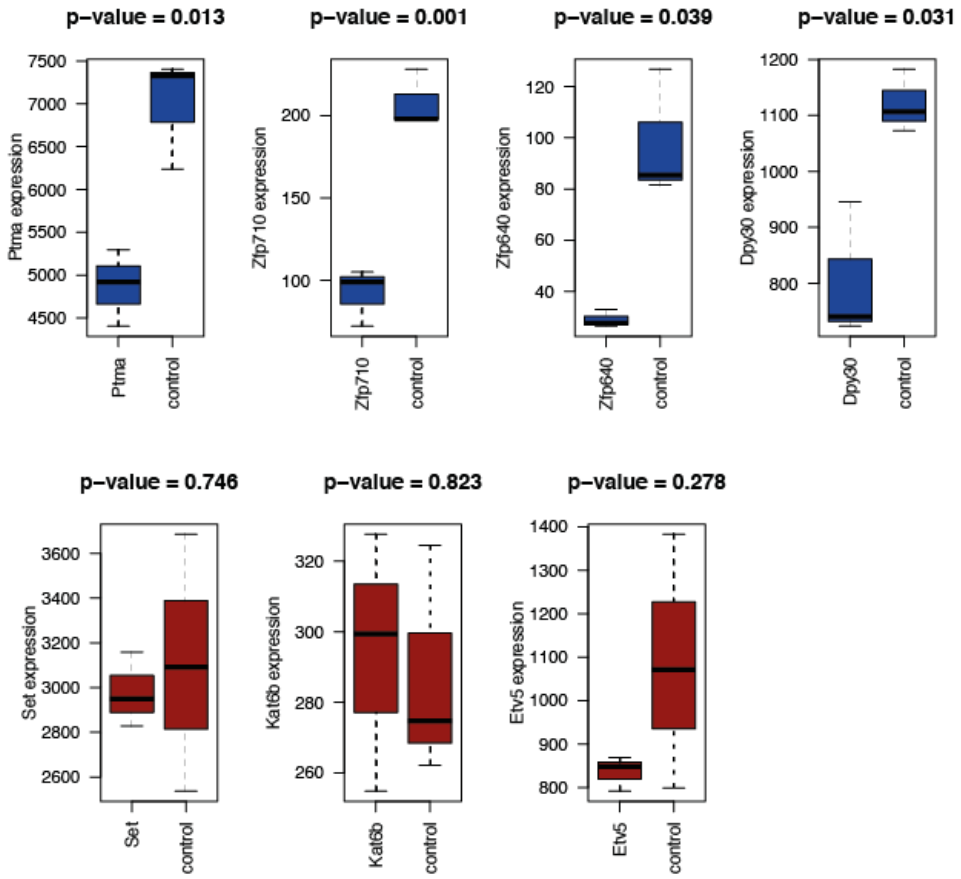


Figure 5.6 CRISPRi results

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*.

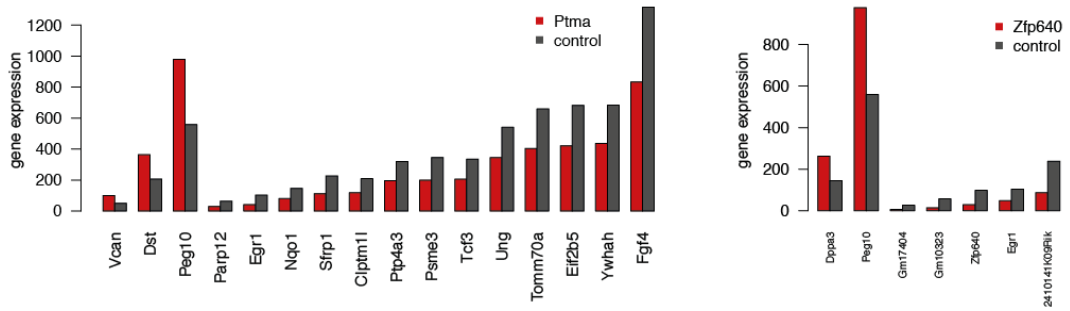
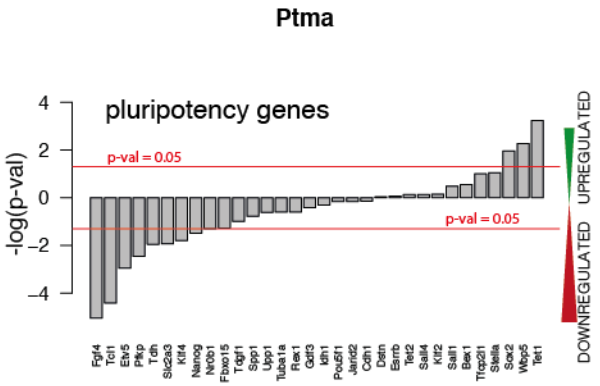


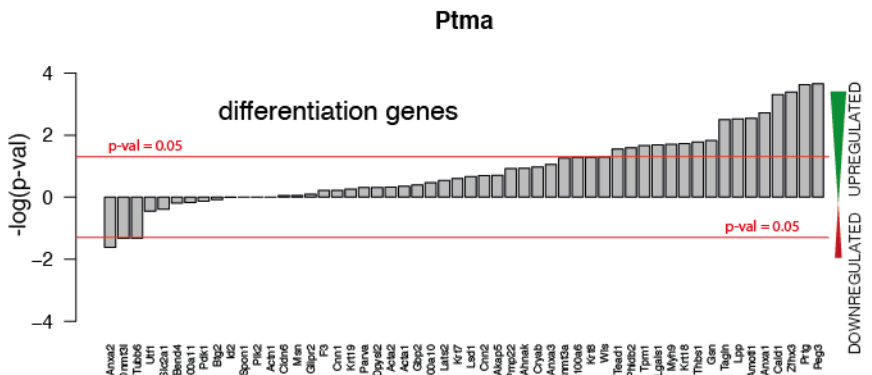
Figure 5.7 Differentially expressed genes in *Ptma* and *Zfp640* downregulated samples

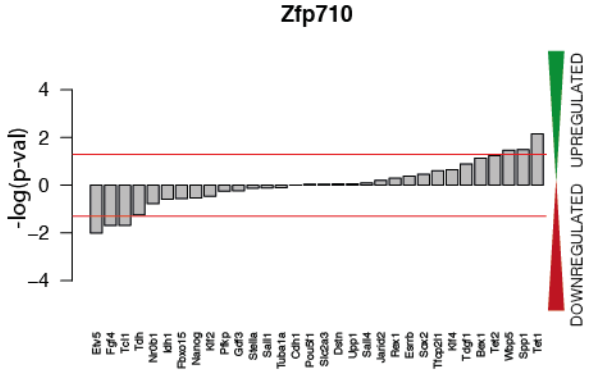
Barplot of gene expression levels of significantly differentially expressed genes in *Ptma* and *Zfp640* repressed samples (DESeq, multiple hypotheses testing adjusted p-value < 0.05).

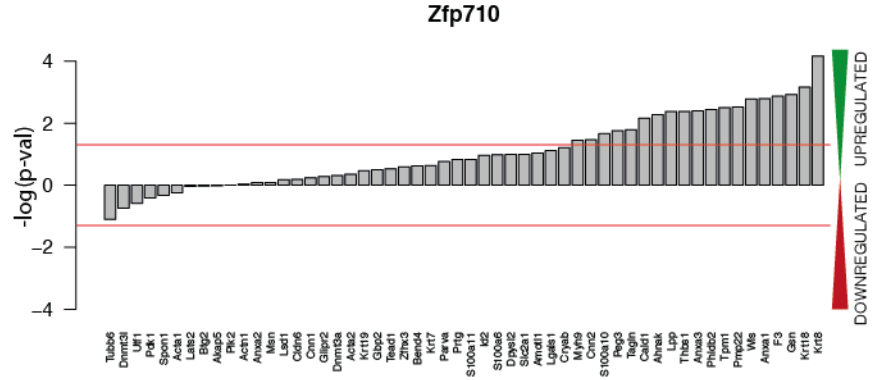
Due to having only three replicates per condition and the relatively low quality of sequencing data I was able to detect only a few significantly differentially expressed genes. To observe if there is a trend for change in expression of major pluripotency and differentiation factors I plotted *p*-values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

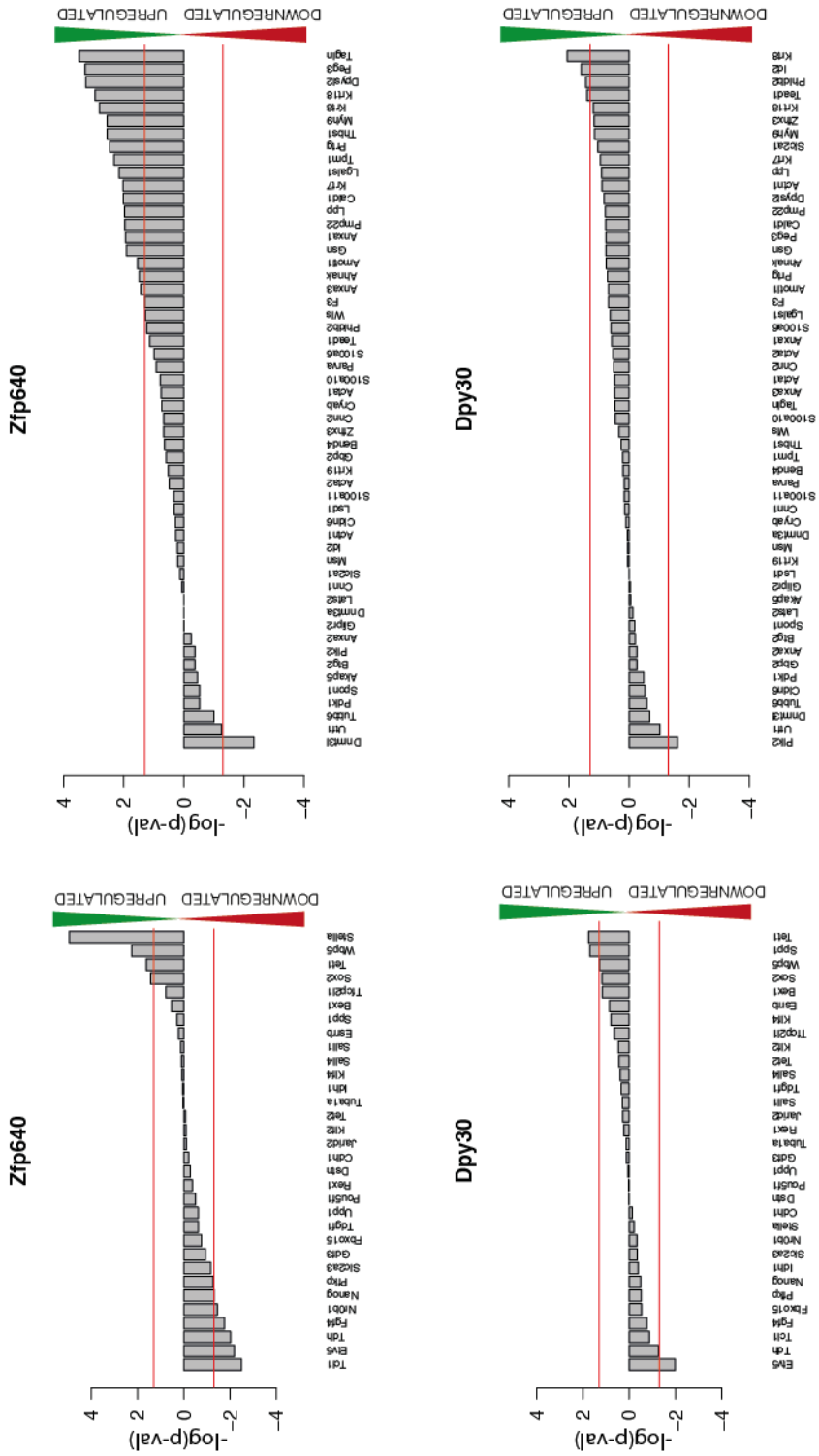


Figure 5.8 Significance of pluripotency and differentiation genes expression changes in knock down samples.

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.