GENE EXPRESSION NETWORK ANALYSIS OF THE ROUTES TO PLURIPOTENCY



Fabio Melo Rosa Amaral BSc (Hons) MSc

School of Biosciences and School of Life Sciences

University of Nottingham

Thesis submitted to the University of Nottingham for the degree of

Doctor of Philosophy

February 2018

I would like to dedicate this thesis to my wife Zaneta and my son Frederico who provided me all the encouragement I needed for achieving this goal and patiently coped with all the hurdles we faced in the last years.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 100,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Fabio Melo Rosa Amaral BSc (Hons) MSc

February 2018

Acknowledgements

Firstly, I would like to thanks Dr Ramiro Alberio and Dr Andrew Johnson for giving me this opportunity to pursue a PhD in their labs and giving me freedom to develop new skills and new ideas. I am very thankful for the Medical Research Council for funding my studentship.

I am also very grateful for Dr. Reinhard Stöger all the useful discussions and support throughout all the stages of my studies.

I big thank you for Professor Tim Somervaille for having confidence on my newly acquired skills and employing me as a bioinformatician in his lab. I would also like to thank Tim for reading my thesis, correcting my grammar and making many useful suggestions.

A very special thanks goes to my wife Zaneta and my son Frederico for their understanding during my absence in a number of occasions and for all the support and encouragement to overcome the difficulties.

Finally I would like to say a big thank you to my parents Marco and Ilka for everything they have taught me in life and for being my role models for correctness and determination to overcome obstacles life throws at us.

Abstract

Great progress has been made towards the understanding of the molecular mechanisms driving factor induced somatic cell reprogramming to pluripotency since the discovery by Takahashi and Yamanaka. However this process remains highly stochastic and inefficient. More study is needed in order to achieve a more deterministic cell fate conversion which could further improve the quality of stem cells generated, essential for prospective therapeutic applications. The work presented here was developed under the premise that natural embryonic development can serve as a guide to achieve more efficient pluripotency induction. It was observed that the histone variant H2A.Z, which has a role in pluripotency in embryonic stem cells, is highly expressed in the oocyte and upon over-expression, together with *Pou5f1*, Sox2, *Klf4* and *Myc*, was able to increase the efficiency of somatic cell reprogramming to induced pluripotency. A gene co-expression network analysis of somatic cells being reprogrammed identified hub genes associated with *H2Af.z* and chromatin remodelling related genes which could be tested for further improving the reprogramming efficiency induced by *H2Af.z* over-expression. Moreover, the study of genetic networks from pre-implantation embryos identified preserved genetic circuits also present during the course of

reprogramming. The most preserved network modules are associated with the final stages of pluripotency induction. However the analysis also identified a genetic network associated with the zygotic genome activation in the totipotent embryo stage which is also found in a sub-population of pluripotent stem cells characterised by the expression of genes from the *Zscan4* family, Tet1, *Etv5* and *Mga* among other genes. This provocative observation led me to hypothesise that during the course of reprogramming the forced activation of hub genes from such network may help improve its efficiency, possibly by recapitulating the natural embryonic processes which induces totipotency prior to pluripotency. The identification of preserved network modules and its hub genes presented in this work may serve a platform for further reprogramming studies in a quest for improved cocktails of reprogramming factors capable of more efficiently generating induced pluripotent stem cells.

Table of contents

List of figures xvii

List of tables xxiii

1 Introduction 1

1.1 The origins of pluripotent stem cells 2

1.1.1 Teratocarcinomas 2

1.1.2 Embryonal carcinoma cells 4

1.1.3 Embryonic stem cells 7

1.1.4 Germ cell derived pluripotent stem cells 11

1.1.5 Human embryonic and epiblast stem cells 12

1.2 Reverting cell fate by nuclear reprogramming 18

1.2.1 Somatic cell reprogramming with defined factors 20

1.2.2 Transcriptional and epigenetic memory 22

1.2.3 Genomic fidelity of reprogrammed cells 24

1.2.4 Factor induced pluripotency vs somatic nuclear transfer . . . 25

1.2.5 Mechanistic studies of iPSC reprogramming 26

1.3 Aims and Objectives of This Study 30

1.3.1 Identification and validation of a new oocyte reprogramming factor 30

1.3.2 Gene network analysis of reprogramming and normal development 31

2 Materials and Methods **33**

2.1 Cell Culture 33

2.1.1 Media and Cell Culture Reagents 33

2.1.2 Cell lines 35

2.1.3 Cell Culture Routine 37

2.1.4 Somatic Cell reprogramming 39

2.2 Molecular Biology 43

2.2.1 RNA Extraction 43

2.2.2 cDNA Synthesis 44

2.2.3 Quantitative PCR 44

2.2.4 Gene Synthesis 45

2.2.5 Gateway cloning 47

2.2.6 Transformations 48

2.2.7 Plasmid Purification 48

2.2.8 Plasmid Sequencing 48

2.2.9 Plasmids 49

2.3 Computational Methods 55

2.3.1 Dataset pre-processing 55

2.3.2 Weighted gene co-expression network analysis (WGCNA) . . 55

3 Mining Oocyte Transcriptome For Reprogramming Molecules 59

3.1 Introduction 59

3.1.1 Maternal reprogramming factors from the oocyte 59

3.1.2 H2A.Z functions 62

3.1.3 H2A.Z is highly conserved in evolution 63

3.1.4 Differences between H2A.Z and H2A 64

3.1.5 Developmental viability depends on H2A.Z 65

3.1.6 H2A.Z roles in mammalian development 66

3.1.7 H2A.Z roles in gene expression regulation 68

3.1.8 H2A.Z and chromatin remodelling complexes 71

3.1.9 Conservation of H2A.Z chromatin remodelling complexes . . 73

3.1.10 H2A.Z roles in stem cell biology 76

3.1.11 H2A.Z and chromatin modifying complexes in pluripotency 78

3.1.12 Objectives 82

3.2 Results and Discussion 83

3.2.1 Determine if H2A.Z is a rate-limiting factor in induced pluripotency 83

3.2.2 Characterization of Nanog null iPS cells reprogrammed with H2A.Z 87

3.2.3 H2A.Z as a rate-limiting factor in reprogramming mouse embryonic fibroblasts to pluripotency 89
3.2.4 Determine the mechanism of deposition of H2A.Z at regulatory regions of key pluripotency genes 93
3.3 Conclusion 97

4 Gene Networks of Somatic Cell Reprogramming towards Pluripotency 99

4.1 Introduction 99
4.1.1 Weighted Gene Co-expression Network Analysis 101
4.2 Results and Discussion 104
4.2.1 Exploratory Data Analysis 104
4.2.2 Weighted Gene Co-Expression Network Analysis 114
4.2.3 Network Modules Based Gene Set Enrichment Analysis . . . 128
4.2.4 Conclusion 132

5 Reproducibility of Reprogramming Gene Networks 143

5.1 Introduction 143
5.2 Results and Discussion 144
5.2.1 Data set manipulation 144
5.2.2 Exploratory Data Analysis 149
5.2.3 Weighted Gene Co-Expression Network Analysis 155
5.2.4 Network Based Gene Set Enrichment Analysis 163
5.2.5 Increased Resolution of Network Based Analysis 174
5.2.6 Network Module Preservation in Reprogramming 176

5.2.7 Conclusion 182

6 Gene Expression Networks in Pre-Implantation Development and their Similarities to Reprogramming Networks 185

6.1 Introduction 186

6.2 Results and Discussion 188

6.2.1 Data Set Manipulation 188

6.2.2 Exploratory Data Analysis 189

6.2.3 Weighted Gene Co-Expression Network Analysis 193

6.2.4 Network Based Gene Set Enrichment Analysis 199

6.2.5 Network Module Preservation in Early Development and Reprogramming 214

6.2.6 Conclusion 224

7 Conclusion 227

References 235

Appendix Supplementary Analysis 269

Appendix Publications 273

List of figures

1.1 Embryonic stem cell differentiation 2

1.2 Pluripotent stem cells can be derived from several sources 4

2.1 Reprogramming Nanog–/– pre-iPS cells 39

2.2 Reprogramming TNG-A pre-iPS cells 40

2.3 *H2afz* gene synthesis 46

2.4 *Hist1h2aa* gene synthesis 46

2.5 *Hist1h2bj* gene synthesis 46

2.6 *Hist1h3c* gene synthesis 47

2.7 *Hist1h4b* gene synthesis 47

2.8 pPB-CAG-DEST-pA-pgk-hph 49

2.9 pCMV-hyPBase 50

2.10 PB-CAG-H2AZ-pA-pgk-hph 50

2.11 pPB-CAG-GTK-H2AZ 51

2.12 PB-CAG-H2A-pA-pgk-hph 51

2.13 PB-CAG-H2B-pA-pgk-hph 52

2.14 PB-CAG-coH3-pA-pgk-hph 52

2.15 PB-CAG-coH4-pA-pgk-hph 53
2.16 PB-CAG-mNanog-pA-pgk-hph 53
2.17 PB-CAG-H2B-Cherry-pA-pgk-hph 54
2.18 pPB-CAG-GTK-H2BCherry 54

3.1 Differences between H2A and H2AZ. 65
3.2 SWR and NuA4 complexes 71
3.3 Nanog requirement is bypassed by the over-expression of H2A.Z/H2B 86
3.4 Morphology comparison of Nanog null cell lines 87
3.5 Expression of pluripotency markers of NS derived iPS cells and its precursors 89
3.6 Reprogramming MEF cells to pluripotency 91
3.7 H2AZ improves MEF reprogramming efficiency 92
3.8 Expression of H2A.Z and chromatin remodelling components in iPS cells and its precursors 95

4.1 Gene network analysis rationale 100
4.2 WGCNA workflow 102
4.3 Gene network analysis of iPS cell generation 104
4.4 Chantzoura data set mapped reads statistics 108
4.5 Chantzoura data set sample clustering 110
4.6 Chantzoura data set singular value decomposition 112
4.7 Chantzoura data set principal component analysis 113
4.8 Selection of power β for the Chantzoura data set 115

4.9 Network module definition 118

4.10 Module merging height 119

4.11 Module eigengenes network for the Chantzoura data set 121

4.12 Network module definition after merges for the Chantzoura data set 123

4.13 Chantzoura data set module eigengene associations with phenotypes126

4.14 Module significance to reprogramming cassette type 127

4.15 Dynamics of Biological Processes in iPS cells reprogramming 131

4.16 Turquoise module characterisation 134

4.17 Green module characterisation 135

4.18 Orange module characterisation 136

4.19 Magenta module characterisation 137

4.20 Purple module characterisation 138

4.21 Red module characterisation 139

4.22 Tan module characterisation 140

4.23 Brown module characterisation 141

4.24 Blue module characterisation 142

5.1 Experimental strategy 144

5.2 FastQC per base quality score distribution for FASTQ file 148

5.3 FastQC per base quality score distribution for BAM file 148

5.4 O'Malley data set mapped reads statistics. 150

5.5 O'Malley data set sample clustering 152

5.6 O'Malley data set singular value decomposition 154

5.7 Principal component analysis of O'Malley data set 154

5.8 Selection of power β for the O'Malley data set 156

5.9 Network module definition for the O'Malley data set 157

5.10 Module merging height 158

5.11 Network module definition after merges 159

5.12 Network module definition after merges 160

5.13 Module eigengene association with cell types and cell markers . . . 162

5.14 Dynamics of Biological Processes in iPS cells reprogramming 166

5.15 Turquoise module characterisation 167

5.16 Green module characterisation 168

5.17 Green module characterisation 169

5.18 Magenta module characterisation 170

5.19 Orange module characterisation 171

5.20 Brown module characterisation 172

5.21 Blue module characterisation 173

5.22 Contingency table DEGs vs Network modules 175

5.23 Chantzoura and O'Malley module genes cross tabulation 177

5.24 Chantzoura and O'Malley data sets module preservation statistics . 179

5.25 Network structures of the pluripotency and intermediate reprogramming chromatin remodelling modules with the most preserved hub genes 181

6.1 Mouse pre-implantation development. 186

6.2 Natural route to pluripotency 188

6.3 Xue data set mapped reads statistics 190

6.4 Xue data set sample clustering 192

6.5 Xue data set singular value decomposition 193

6.6 Xue data set principal component analysis 194

6.7 Selection of power β for the Xue data set 195

6.8 Network module definition 197

6.9 Network module definition after merges 198

6.10 Module eigengene association with cell type 200

6.11 Dynamics of Biological Processes During Early Development 203

6.12 Turquoise module characterisation 205

6.13 Blue module characterisation 206

6.14 Brown module characterisation 207

6.15 Purple module characterisation 208

6.16 Magenta module characterisation 209

6.17 Red module characterisation 210

6.18 Cyan module characterisation 211

6.19 Green module characterisation 212

6.20 Yellow module characterisation 213

6.21 Shared networks between iPS cell generation and early development 214

6.22 Chantzoura and Xue module genes cross tabulation 215

6.23 Module preservation statistics between Xue and Chantzoura data sets 216

6.24 Gene set enrichment analysis of preserved hub genes 218

7.1 Gene Networks in 2C stage and 2C-like cells 231

List of tables

2.1 Applied Biosystems TaqMan RT-qPCR custom assays 45

2.2 Applied Biosystems TaqMan RT-qPCR standard assays 45

4.1 Chantzoura *et al.* RNA-sequencing data set 106

5.1 O'Malley *et al.* RNA-sequencing data set 147

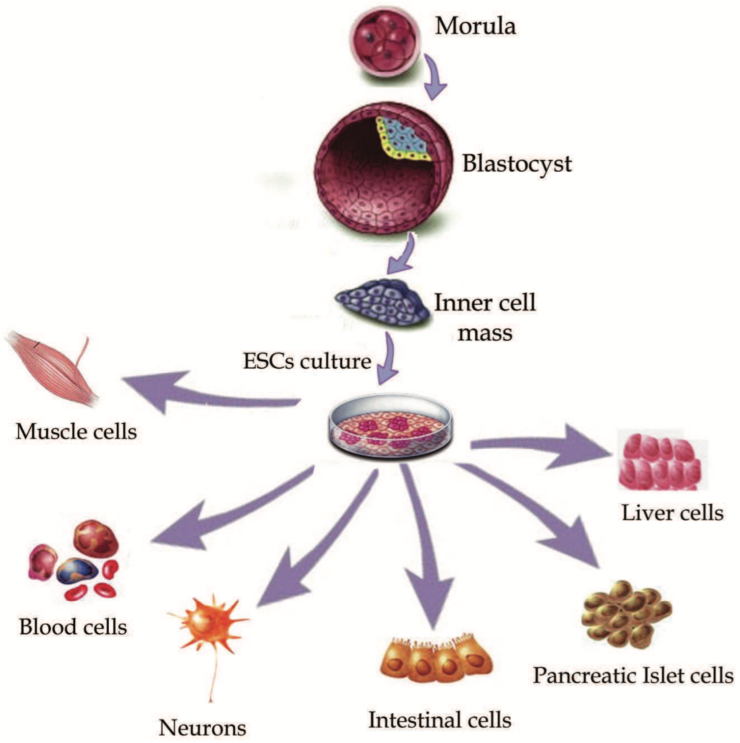
6.1 Preserved hub genes between pre-implantation development and somatic cell reprogramming 219

Chapter 1

Introduction

Pluripotent stem cells are defined by their capacity to self-renew indefinitely while maintaining the ability to differentiate into any of the primary germ layers (ectoderm, mesoderm and endoderm) (Figure 1.1) as well as the germline (De Los Angeles et al., 2015; Smith, 2001). When these stem cells are grafted ectopically (e.g. subcutaneous injection) into a suitable host they produce multi germ layer tumours called teratomas (Evans, 1972; Stevens, 1970) and most importantly they can form a whole new germ-line competent animal when introduced into the pre-implantation embryo stages of the morula or blastocyst (Bradley et al., 1984; Brinster, 1974; Illmensee and Mintz, 1976; Mintz and Illmensee, 1975; Stewart and Mintz, 1981, 1982). These properties have sparked great scientific interest as a means to further explore animal developmental biology, develop cell replacement therapies, disease modelling and drug screening (Karagiannis and Eto, 2016). In order to achieve its full application potential and safety it is imperative to have a

thorough knowledge of how these cells are formed, maintained and differentiated at a systems biology level.



Adapted from Meregalli et al. (2011)

Fig. 1.1 Embryonic stem cell differentiation

1.1 The origins of pluripotent stem cells

1.1.1 Teratocarcinomas

The discovery of pluripotent stem cells originated from the study of rare germ cell tumours known since antiquity and called teratomas (benign) or teratocarcinomas (malignant). These tumours usually appear localised in the gonads, however they can be also occasionally found at ectopic sites (Solter, 2006). Teratoma takes its names from the Greek word *teratos*, which means monster, by virtue of its bizarre

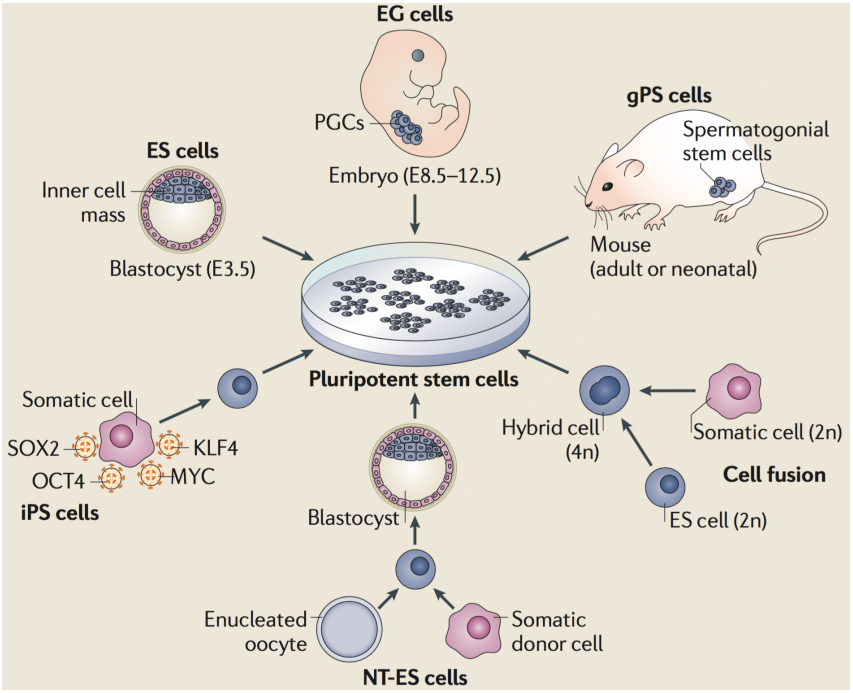
composition of a disorganised mixture of adult tissues such as cartilage, teeth, skin, hair, muscle and neural cells.

The study of this unusual disease started to gain momentum with the observation that the mouse strain 129 sporadically develops testicular teratomas at a frequency of 1% (Stevens and Little, 1954). This mouse strain permitted the development of experimental techniques which led to the observation that new teratomas could be reconstituted by grafting part of these tumours subcutaneously or intraperitoneally, highlighting its malignant properties (Solter, 2006). The high incidence of teratomas has a strong genetic component with a mutation on the gene dead-end homologue 1 (Dnd1) present in the 129/terSv mouse sub-strain being responsible for increasing this tumour incidence to about 30% (Stevens, 1973; Youngren et al., 2005).

In a quest to narrow down the origin of teratoma causing cells it was observed that experimental tumour formation was made even more efficient when fetal genital ridges were used as a grafting donor tissue, allowing the first descriptions of how teratomas are formed (Stevens, 1964). Similarly, re-transplantable tumours could be derived from the engraftment of whole pre-implantation embryos suggesting a close relationship of teratoma and early mouse development (Diwan and Stevens, 1976; Solter et al., 1970; Stevens, 1968). Although teratomas are predominately found in male gonads these tumours may also emerge in females. The mouse strain LT develops ovarian teratomas at a high frequency as a result of spontaneous parthenogenetic development from ovarian oocytes underscoring the connec-

tions between the development of teratomas and early embryos (Stevens and Varnum, 1974).

The pluripotent stem cell component of teratomas became evident with the demonstration that the serial engraftment of a single cell derived from a tumour could generate new teratomas (Kleinsmith and Pierce, 1964). This experiment demonstrated that a cell type contained in teratomas is able to grow indefinitely and also develop into several mature cell types and disorganised adult tissues.



Adapted from Gaspar-Maia et al. (2011)

Fig. 1.2 Pluripotent stem cells can be derived from several sources

1.1.2 Embryonal carcinoma cells

The early efforts to culture explants from both testicular and embryo-derived teratocarcinomas led to the isolation of clonal cell lines which, given their origin,

were denominated embryonal carcinoma (EC) cells. EC cells were characterised by their tendency to form small, tight colonies of cells with a high nuclear cytoplasmic ratio and prominent nucleoli. Remarkably, EC cell lines were able to grow indefinitely while keeping the same capacity to differentiate to diverse adult cell types as the donor tumours both *in vivo*, recapitulating teratoma formation when transplanted to a suitable host, and spontaneously in *in vitro* cultures.

The establishment and maintenance of EC cells were greatly assisted by co-culture with mitotically inactivated fibroblast feeder layers (Kahan and Ephrussi, 1970) and by certain batches of serum supplemented media (Martin, 1980; Martin and Evans, 1975) which soon became the standard culture method for these cells at that time. The importance of feeder cells was evidenced by the observation that, when clonal EC cell lines were removed from feeder supported cultures and transferred to gelatin-coated dishes, some undifferentiated cell foci could still be maintained in regions where EC cell derived fibroblasts had accumulated (Martin and Evans, 1975). Substrate attachment was also seen to be essential for the propagation of undifferentiated EC cells since when these cells are cultured in low adherent substrates, such as plastic, EC cells readily formed aggregates termed embryoid bodies containing a core of stem cells surrounded by a shell of primitive extra-embryonic endoderm cells and would completely differentiate into diverse cell types after prolonged culture (Martin and Evans, 1975).

The identifications of cell markers able to distinguish undifferentiated EC cells was extensively used to assist in the derivation and characterisation of cell lines.

The most widely used procedure for a low resolution identification of pluripotent stem cells to this day is the detection of alkaline phosphatase activity which is highly expressed in murine and human EC cells as well as in the inner cell mass (ICM) of the blastocyst and in the ectoderm and primordial germ cells (PGCs) (Benham et al., 1981; Berstine et al., 1973). The stage-specific embryonic antigen 1 (SSEA1, also known as Fut4) is another widely used surface marker which reliably detects undifferentiated mouse EC cells, ES cells and PGCs. Interestingly this marker is not expressed in undifferentiated human EC cells but only some of their differentiated cell derivatives. Instead, human EC cells are identified by the expression of the surface makers stage-specific embryonic antigen 3 (SSEA3) and 4 (SSEA4) with the former being also a marker of the mouse counterpart (Kannagi et al., 1983; Shevinsky et al., 1982). Two other classic EC cell surface markers are TRA1-60 and TRA1-81 which are also expressed in human ES cells and early human embryos (Andrews and Andrews, 2002; Fenderson et al., 1987; Henderson et al., 2002).

The directed differentiation of EC cells *in vitro* began with the use of chemical compounds such as retinoic acid, dibutyryl cAMP and hexamethylbisacetamid which, depending on their combination, could bias the differentiation towards one or another primitive germ layer derivatives achieving similar results with cells from mouse and human origins (Andrews, 1984; Jakob et al., 1978; Simeone et al., 1990; Strickland and Mahdavi, 1978). The various similarities observed between the properties of EC cells and cells from the early embryo prompted the question whether these tumor derived cells could colonise embryos and contribute to

normal development when injected into the mouse blastocyst cavity. This line of investigation led to a series conflicting results with various degrees of success. The vast majority of studies reported modest animal chimerisms as evidenced by the extent of contribution of pigmented coat in albino hosts (Brinster, 1974). One group reported extensive embryonic contribution including germline chimerism suggesting that the malignant phenotype of tumor derived cells was reversible and that EC cells are equivalent to embryonic cells (Illmensee and Mintz, 1976; Mintz and Illmensee, 1975; Stewart and Mintz, 1981, 1982). However other groups, using different cell lines reported more limited chimerism and also reported frequent development of tumours derived from embryonal carcinoma cells (Papaioannou et al., 1978, 1975; Rossant and McBurney, 1982) underscoring the observation that EC cells are invariably accompanied with various degrees of kariotypic abnormality (Blelloch et al., 2004; Martin, 1980). Despite the mixed outcomes obtained by different labs, mainly due to different levels of chromosomal anomalies of different EC cell lines, the study of these tumor derived stem cells was decisive for the identification of the conditions for isolating pristine embryo derived pluripotent stem cells which were termed Embryonic Stem (ES) cells and are considered the gold standard cell for pluripotency studies to this day.

1.1.3 Embryonic stem cells

The foundational work on teratomas and the empirically defined culture conditions set for EC cells paved the way to the isolation of ES cells. In 1981 two groups

independently reported the derivation of the first ES cell lines from mouse blastocysts cultured on a feeder layer of division-inactivated fibroblasts (Evans and Kaufman, 1981; Martin, 1981). More precisely, the source of ES cells was shown to be the transient inner cell mass (ICM) from the embryonic days 3.5-4.5 (E3.5-E4.5) pre-implantation embryo (Brook and Gardner, 1997; Gardner and Brook, 1997) (Figure 1.2). Interestingly, the cells of the ICM usually have a very transient period of self-renewal *in vivo*. However, this self-renewing period can be further extended during diapause, a condition in which the embryonic development is delayed by arresting at the peri-implantation stage (Smith, 2001). In mice this phenomenon is caused by estrogen deficit due to persistent lactation, ovariectomy, or by action of estrogen antagonists (Nichols and Smith, 2011). In fact the first ES cells were derived from blastocyst in diapause (Evans and Kaufman, 1981) and it has been shown to greatly enhance ES cell generation efficiency (Brook and Gardner, 1997).

These cells proved to be very similar to mouse EC cells in terms of known marker expression, self renewal and differentiation capacities, but most importantly, stable euploid cells could be reproducibly isolated which were significantly more efficient at producing high-grade germline competent chimeras (Bradley et al., 1984; Kuehn et al., 1987; Robertson et al., 1986) and could even form an entire animal when injected into tetraploid blastocyst, which has the development restricted to extra-embryonic tissues (Nagy et al., 1993).These properties sparked great scientific interest as ES cells became an essential tool for manipulating mouse embryos to study mouse genetics, development and physiology (Evans, 2011).

Precise genetic engineering, using mouse ES cells as a vector for custom genetic modifications introduced by homologous recombination, has revolutionised how gene function can be studied (Capecchi, 2008; Smithies, 2008; Thompson et al., 1989). This technology led to a worldwide effort coordinated by the International Knockout Mouse Consortium (IKMC) with the aim of mutating all protein-coding genes in the mouse using a combination of gene trapping and gene targeting (Collins et al., 2007). This initiative was further bolstered by the European Conditional Mouse Mutagenesis Program which has so far generated a resource of over ten thousand mutant ES cell lines and over two thousand transgenic mouse lines all featuring conditional loss of function targeted genes permitting the precise assessment of gene function in a temporal and spatial manner (Skarnes et al., 2011).

The study of pluripotent stem cell properties was greatly enhanced by the extensive refinement of the culture conditions of mouse ES cells with the elimination of the confounding factors coming from serum and feeder cells. The fibroblast feeders were found to be a source of leukemia inhibitory factor (LIF) and the supplementation of this cytokine alone was sufficient to eliminate the need of a co-culture with feeders if ES cells are grown on gelatin coated dishes (Smith et al., 1988; Williams et al., 1988). Furthermore, the observation that ES cells cultured in the presence of LIF without serum results mainly in its differentiation into neural fate led to the discovery that bone morphogenetic proteins (BMPs), which are potent antagonists of neural specification in vertebrate embryos, are sufficient to prevent differentiation of mouse ES cells (Ying et al., 2003; Ying and Smith, 2003).

While LIF and BMP have been demonstrated to be the minimum cytokines/growth factors capable of sustaining pluripotent self-renewal they were later shown to be dispensable altogether for the derivation, propagation and maintenance of pluripotency of mouse ES cells if pro-differentiation signals are inhibited by small molecules (Ying et al., 2008). This regimen involving the compounds PD0325901 and CHIR99021 (2i) is capable of inhibiting of Fibroblast Growth Factor (FGF) / Extracellular Signal-regulated Kinase (ERK) and Glycogen Synthase Kinase 3 (GSK3) respectively in fully chemically defined serum free media (Ying et al., 2008). The culture in 2i/LIF conditions can robustly sustain homogeneously undifferentiated naïve (or ground state) pluripotent mouse ES cells and has made possible the derivation of high grade germline competent ES cells from all murine genetic backgrounds so far attempted (Batlle-Morera et al., 2008; Nichols et al., 2009) as well as from rats which were notorious to be recalcitrant to the conventional methods (Buehr et al., 2008).

With the advent of defined culture systems able to support the maintenance of homogeneously naïve mouse ES cells it was highlighted that this *in vitro* cell state was indeed equivalent to cells from the ICM at E4.5 stage at the transcriptome level (Boroviak et al., 2015). It is now well established that, other than the differentiation properties aforementioned, important defining characteristics of naïve pluripotency are the expression of a specific set of genes; a global state of DNA hypomethylation; two active X-chromosomes in female cell lines; a low level of bivalent histone marks; a capacity to generate energy through both oxidative phosphorylation (associated with a high mitochondrial content) and

glycolysis; the ability to propagate as single cells with high cloning efficiency; a lack of lineage priming or the co-expression of pluripotency-associated genes and lineage-specific markers (Davidson et al., 2015).

1.1.4 Germ cell derived pluripotent stem cells

The realisation that spontaneously occurring teratomas could produce pluripotent EC cells led to the suggestion that these tumors may be derived from primordial germ cells (PGCs) (Stevens, 1967). This hypothesis has been confirmed when embryonic germ (EG) cells with very similar pluripotent characteristics to ES cells were derived by culturing the posterior region of 8.5 dpc mouse embryos on feeder cells in media supplemented with LIF, steel factor and FGF4 (Matsui et al., 1992; Resnick et al., 1992) (Figure 1.2). EG cells can also be derived from cultured explants of 11.5 and 12.5 dpc genital ridges.

Interestingly, EG cells derived at this stage have limited capacity to integrate into chimeras and particularly into the germline (Labosky et al., 1994; Tada et al., 1998). These differences may be due to aberrant imprinting in late derived EG cells. The genes expressed in ES cells exhibit somatic specific imprinting patterns and EG cells derived from migratory PGCs after E9.5 display reduced methylation imprints, probably reflecting the epigenetic reprogramming occurring in the founder PGCs at that stage (Geijsen et al., 2003).

Pluripotent stem cells have also been isolated at lower efficiency from a later germline developmental stage. These germline-derived pluripotent stem (gPS)

cells are derived from spermatogonial stem cells of neonatal and adult testes (Kanatsu-Shinohara et al., 2004; Ko et al., 2009) (Figure 1.2).

EG cells have also been derived from human PGCs when cultured on mouse STO (S, Swiss albino Sandoz Inbred Mouse; T, 6-thioguanine resistant; O, ouabain resistant) fibroblast feeders and media supplemented with LIF, bFGF and forskolin (Shamblott et al., 1998). These human EG cells are remarkably similar to human ES cells however they have not been shown to be able to be continuously passaged for more than 20-25 times (Shamblott et al., 1998).

Despite being considered unipotent (precursors of the gametes), PGCs are thought to be in a dormant pluripotency state evidenced by the observation that, with exception of *Klf4*, virtually all pluripotency genes are up regulated during their specification (Leitch et al., 2013). Clearly PGCs can acquire pluripotency *in vivo* in spontaneous teratomas and *in vitro* by EG cell conversion via a poorly understood epigenetic reprogramming event that may be triggered by FGF (Durcova-Hills et al., 2006). However FGF supplementation seems to be dispensable for EG cells derivation if PGCs are cultured directly in 2i/LIF media conditions (Leitch et al., 2010).

1.1.5 Human embryonic and epiblast stem cells

The derivation of pluripotent stem cells from human pre-implantation embryos only became possible around twenty years after the first reports of the isolation of mouse ES cells (mESCs), which was followed by a significant increase in the

volume of research in the pluripotency field due to its potential for biomedical applications (Thomson, 1998). Interestingly, human and mouse ES cells differ in most, if not all, of the above described characteristics for ground state pluripotent stem cells, including signalling requirements for self-renewal maintenance (Davidson et al., 2015). Similar to human EC cells, the embryo derived counterparts displayed different morphology to mESCs, expressed a distinct set of surface markers, enhanced levels of DNA methylation, X-inactivation, was unable to grow at clonal density and was dependent upon FGF and activin to self-renew instead of LIF.

The many differences between mouse and human embryo-derived pluripotent stem cells were initially suggested to be due to intrinsic species differences. However this notion was put to question with the isolation of pluripotent stem cells from mouse post-implantation epiblast, which were termed epiblast stem cells (EpiSCs) (Brons et al., 2007; Tesar et al., 2007). It has been shown that mouse EpiSCs (mEpiSCs) share strikingly similar properties with the standard human ES cells (hESCs) indicating that the latter may actually correspond to a more advanced stage of development than the blastocyst, termed primed pluripotency (Brons et al., 2007; Nichols and Smith, 2009, 2011; Rossant, 2008; Tesar et al., 2007). In agreement with this notion mEpiSCs shows increased lineage priming and little or no ability to participate in chimera formation (Hackett and Surani, 2014).

Careful comparison between gene expression in mEpiSCs derived at different stages of post-implantation development and the mouse epiblast from E5.5 to

E8.5 shows that mEpiSCs are most similar to the ectoderm of the anterior primitive streak, a relatively late stage of development (Kojima et al., 2014). By modifying the culture conditions for mEpiSCs, Wu et al. were able to derive mEpiSCs that were more similar to late primitive streak-stage epiblasts (Wu et al., 2015). Both of these studies indicate that mEpiSCs are poised at a fairly late stage of epiblast development.

The primed pluripotent state of EpiSCs lacks the expression of *Rex1* and *Klf4*, which are diagnostic markers of naïve ES cells, and are characterised by the expression of Fgf5 and *brachyury* (T) and one inactive X chromosome in female cells (Brons et al., 2007; Tesar et al., 2007). It has been demonstrated that primed state EpiSCs can be converted back to naïve state by ectopic expression of pluripotency factors such as *Klf4*, restoring the most stringent properties of naïve state pluripotency (Guo et al., 2009; Hall et al., 2009). This conversion can also be induced *in vitro* in the presence of feeders and LIF, at a very low efficiency however, indicating a level of plasticity.

The notion that primed pluripotent stem cells can be reset to a naïve state similar to that of mESCs and the cellular and molecular similarities between hESCs and mEpiSCs are central to the hypothesis that it must be possible to derive human naïve pluripotent stem cells. Following this line of thought several recent studies have been reporting the acquisition of human or non-human primate naïve pluripotent stem cells using reprogramming approaches similar to what has been described for mouse cells (Buecker et al., 2010; Chan et al., 2013; Fang et al., 2014;

Gafni et al., 2013; Hanna et al., 2010; Takashima et al., 2014; Theunissen et al., 2014a; Valamehr et al., 2014; Wang et al., 2014, 2011; Ware et al., 2014). The initial studies relied on the continuous expression of extra reprogramming factors such as Klf4 to maintain naïve pluripotency after reprogramming but most recent reports have focused on conditions independent of exogenous transgenes.

Whereas these published protocols vary significantly, most culture conditions are based on the use of small molecule inhibitors of MEK/ERK and GSK3B (2i) along with LIF and some extra compounds. Many use FGF2 and some form of transforming growth factor beta (TGFβ) agonist, and most employ mouse embryonic fibroblast feeder cell support and complex protein additives such as knockout serum replacement (KSR).

The resultant cells also show significant differences which became evident with a systems biology analysis approach (Weighted Gene Co-Expression Network Analysis, WGCNA) comparing the gene co-expression patterns in mouse and human naïve and primed cells, and pre-implantation embryos of both species (Huang et al., 2014a). This study has reached some important conclusions reflecting our lag in knowledge about early embryo development and pluripotency characteristics in the human in comparison to the most well established studies in mouse. As expected naïve mESCs and primed mEpiSCs correlate well to the late mouse blastocyst and post-implantation epiblast respectively. Critically this analysis has shown that mESCs grown in serum/LIF have no significant correlation with any

stage of mouse embryonic development indicating an *in vitro* culture adaptation to serum (Huang et al., 2014a).

Surprisingly, the gene co-expression network profiles between human standard ES cells, which were used to compare with the naïve cell, were proven to be significantly heterogeneous and resembles blastocyst outgrowths, suggestive of adaptation to culture. This heterogeneity is surprising given these are well established cell lines maintained with standard culture conditions by several laboratories around the world. This observation put into question the current standards for defining the baseline of human primed pluripotency.

Moreover, the various human naïve cells generated with different protocols were shown to be quite different from each other reflecting the different signalling requirements for each system. The most conserved gene co-expression modules among several studies analysed are associated with RNA processing and ribosomal biogenesis, as well as mitochondrial genes, suggesting that a common metabolic state might be a crucial component for human naïve pluripotency.

Importantly, even considering all the variation between studies, the recently established human naïve pluripotent stem cells showed a higher correlation with the human late blastocyst than standard human ES cells, with the cells produced by the protocols of Theunissen et al. (2014b) and Takashima et al. (2014) scoring the highest similarities with the human blastocyst. Furthermore, the human naïve cells from these two studies were the only ones showing a significant correlation to mouse naïve ESCs and mouse late blastocyst which has been widely used as

the gold standard for naïve pluripotency. Following from the initial success at resetting primed pluripotency to naïve from pre-established human standard ESCs, the culture conditions used by Takashima and colleagues was used to derive naïve ESCs directly from the pre-implantation human embryo (Guo et al., 2016).

It is likely that the culture conditions for maintaining stable human naïve pluripotent stem cells will continue to be refined and become more standard and independently replicated. However a number of these culture conditions recently published were reported as not being reproducible (Theunissen et al., 2014a) indicating that the criteria for assessing the human naïve pluripotency is still far from a consensus.

We have learned so far that the naïve pluripotency network which has been established in human cells is similar but not identical to mouse naïve ESCs. The human naïve pluripotency network is much less stable than that which is seen in the mouse and this difference is likely due to species specific developmental differences. One key difference is the absence of diapause in the human, which in the mouse is correlated with more efficient ESC derivation (Pera, 2014). This may explain why the activation of a regulatory network that can keep cells in a pluripotent state against their natural tendency to differentiate is more difficult in human cells (Rossant, 2015). Further systematic comparative studies of primate pre-implantation and post-implantation and more extensive data on *de novo* derivation of naïve human PSCs are needed to more definitively address the signalling requirement for human naïve pluripotency maintenance.

1.2 Reverting cell fate by nuclear reprogramming

The discovery that a differentiated somatic cell can be reverted to a ground state embryonic stage, able to develop into a whole new living being, is arguably one of the greatest scientific breakthroughs of all time. Our ability to artificially revert cell fate has caused a shift in a long standing paradigm dating back to the 19th century that development can only occur in one direction (Weismann et al., 1894). The classic view of normal embryonic development was metaphorically described by Waddington's epigenetic landscape which compared the unidirectionality of development as balls rolling down a hill with branching paths, finding their terminal fate at the bottom of the slope (Waddington, 1957).

The irreversibility of development was challenged by Briggs and King who first performed cloning experiments by nuclear transfer from early tadpole embryonic cells into enucleated frog's eggs with a goal of studying nuclear differentiation (Briggs and King, 1952). They observed that the nuclei of embryonic cells at the earliest stages are better suited for cloning than from cells of later stages, which they believed would gradually lose the capacity to revert their developmental state.

This view began to change when John Gurdon was able to clone a frog by transplanting the nucleus of a tadpole intestinal cell into an enucleated frog egg (Gurdon, 1962; Gurdon et al., 1958). The resetting of a specialised cell to full totipotency was a major breakthrough ahead of its time which showed that the complete set of genetic information necessary for development is retained even after cell specialisation.

However, the final proof of genomic equivalence was provided nearly four decades later by Keith Campbell and Ian Wilmut with the cloning of a sheep. In this landmark experiment they created the first ever mammal clone from an adult somatic cell nuclear donor (Wilmut et al., 1997). This feat was accomplished by culturing cells from the mammary gland of an adult pregnant ewe in conditions that kept the nuclei of these cells at the resting cell cycle stage (G_0), and after a few passages they electrofused these cells with enucleated oocytes from a different sheep strain. From the 277 oocytes initially used, one sheep developed to term, which was named Dolly after the American country singer Dolly Parton.

Cloning by somatic cell nuclear transfer (SCNT) was soon reproduced with several other animals such as cows (Kato et al., 1998), mice (Wakayama et al., 1998) and more recently even with human, for the isolation of SCNT derived hESCs overcoming decades of technical challenges (Tachibana et al., 2013). The use of mature mouse lymphocytes as nuclei donors provided an undisputed evidence of the reprogramming of terminally differentiated cells (Hochedlinger and Jaenisch, 2002). The genetic rearrangements that occur during lymphocyte maturation presented a means of categorically proving that donated nuclei came from a fully differentiated adult cell and not from some persisting stem cell which could have contaminated the samples.

Apart from the academic interest in understanding this fundamental biological phenomenon, SCNT mediated reprogramming has opened a whole new range of potential applications for this technology. Opportunities to *rescue* extinct or en-

dangered species and to use cloned human embryos as a source of isogenic hESCs for biomedical applications (therapeutical cloning) have both created great expectations as well as heated ethical debates. One of the most controversial issues shared both by therapeutical cloning and derivation of human ES is the destruction of embryos and debates about this question may never reach a consensus due to cultural and religious diversity.

1.2.1 Somatic cell reprogramming with defined factors

While somatic cell reprogramming studies via SCNT technology progressed at a slow pace due to its technical difficulties, the discovery of induced pluripotency with defined factors took the scientific community by storm. Inspired by the seminal discovery that a single transcription factor (MyoD) was able to reprogram fibroblasts into muscle cells (Davis et al., 1987) Takahashi and Yamanaka set out to assess the reprogramming potential of genes important for the maintenance of pluripotency in mouse embryonic stem cells. Starting with 24 candidate genes they were able to reduce the transcription factor cocktail to four factors *Oct4*, *Sox2*, Klf4, and *c-Myc* (hereafter referred to as OSKM). The resulting cells termed induced pluripotent stem cells (iPSCs) exhibited the morphology and growth properties of ESCs, expressed ESC marker genes and produced teratomas containing a variety of tissues from all three germ layers after subcutaneus transplantations in mice (Takahashi and Yamanaka, 2006a). Strikingly the same reprogramming factors were able to reprogram human somatic cells to pluripotency (Takahashi

et al., 2007). This astonishing discovery has revolutionised the field of regenerative medicine and sparked a great scientific interest on the paradigm shift questions about cellular identity it created. However before iPSC technology could fulfil its much anticipated role in autologous cell replacement in regenerative medicine as well as a tool for drug discovery screenings and developmental biology studies several technical hurdles had to be overcome. Genetic mutations, chromosomal and epigenetic aberrations often detected in some iPSC clones have raised concerns about their tumorigenicity potential which can jeopardise their clinical application (Yamanaka, 2012).

The first generation of iPSCs was created with retroviral transduction as a form of transgene delivery and were found to be only partially reprogrammed as they often were unable to reactivate the endogenous expression of key pluripotency factors and shut down the expression of the somatic genes as well as the introduced transgenes. These inconsistencies could be circumvented by the selection of reprogrammed clones which reactivated Nanog expression and remarkably these cells displayed a similar gene expression and global DNA methylation patterns to ESCs and were able to produce germ-line competent adult mice chimeras (Okita et al., 2007). Alarmingly, approximately 20% of the offspring developed tumours attributable to reactivation of the *c-Myc* transgene (Okita et al., 2007).

The following years after the iPSC technology inception several refinements were published addressing alternative transgenic factor delivery methods such as excisable polycistronic lenti-viruses and piggyBac transposons as well as non-

integrating approaches such as adeno and Sendai viruses, transfection of episomal DNA, plasmid DNA, mRNA and protein transduction (Bian et al., 2009; Carey et al., 2009; Fusaki et al., 2009; Kim et al., 2009; Okita et al., 2008; Sommer et al., 2009; Stadtfeld et al., 2008; Warren et al., 2010; Woltjen et al., 2009; Yu et al., 2009; Yusa et al., 2009; Zhou et al., 2009). Despite the low efficiency of these methods it became possible to produce iPSC clones completely free from exogeneous material. Importantly the choice of reprogramming method has decisive impact on the observed transcriptional and epigenetic patterns of iPSC clones.

1.2.2 Transcriptional and epigenetic memory

Initial comparisons between human iPSCs (hiPSCs) and hESCs have highlighted that hiPSCs often retained a transcriptional and epigenetic memory of the donor cells indicating incomplete reprogramming but also displayed aberrant epigenetic patterns at some hiPSCs loci which were considered to be reprogramming errors (Chin et al., 2009; Doi et al., 2009; Ghosh et al., 2010). Up to 80% of the differences in the gene expression profile between hESCs (H9 cell line) and hiPSCs are considered as a result of residual reprogramming factor expression from integrative vectors and this variation is consistently diminished after transgene excision (Soldner et al., 2009) alleviating initial concerns that hiPSCs and hESCs were two different pluripotent cellular subtypes (Chin et al., 2009). Unexpectedly it has been observed that the use of Sendai virus causes a permanent change in

the hiPSC transcriptional pattern even after the loss of the vector despite being a non-integrative virus with transient transgene expression (Choi et al., 2015).

Another factor often overlooked in many reprogramming protocols is the stoichiometry of the factors involved in the reprogramming process which has been shown to affect the final epigenetic state of iPSCs (Carey et al., 2011). High expression of *Oct4* and *Klf4* combined with lower expression of c-Myc and *Sox2* is able to produce better quality iPS cells with higher capacity to generate "all-iPSC mice" by tetraploid (4n) complementation (Carey et al., 2011). Nevertheless, it has been shown that somatic cell memory tends to be erased at the transcriptional and epigenetic levels in prolonged cultures of hiPSCs as reprogramming appears to continue gradually (Krijger et al., 2016; Polo et al., 2010). Moreover, it has been demonstrated that the genetic background is responsible for most of the gene expression pattern differences previously observed between hESCs and hiPSCs by systematically comparing isogenic cell lines (Choi et al., 2015). The genetic background was also shown to be responsible for most of the observed transcriptional heterogeneity in hiPSCs often attributed to residual somatic cell memory (Rouhani et al., 2014). It was demonstrated that hiPSC lines derived from different tissues of the same person are more similar than hiPSC lines generated from the same type of cell but from different donors (Rouhani et al., 2014). The differences in genetic background also have an impact on the differentiation potential of the resultant hiPSC line regardless of the type of somatic cell used. (Kyttälä et al., 2016). Therefore current evidence suggests hiPSCs and hESCs are not significantly

different from each other when taking into account optimum reprogramming protocols.

1.2.3 Genomic fidelity of reprogrammed cells

The genomic stability of reprogrammed cells is also an important concern, especially in the context of clinical applications. Given that the generation of iPSCs requires a considerably long time in non-physiological culture conditions the probability of acquiring potentially tumorigenic genomic mutations has raised apprehension about the safety of these cells for medical application. However, this concern is not exclusive to iPSCs as ESCs also have a tendency of picking up genomic abnormalities over time in culture. In fact a cytogenetic study of a large number of hiPSC and hESC lines have shown that the frequency of abnormal karyotypes was similar in both cell types and the types of chromosomal aberration indicates a process of culture adaptation (Taapken et al., 2011). Recurrent partial gain of chromosome 12 which contains the pluripotency factor NANOG is a feature observed in both hiPSCs and hESCs. On the other hand trisomy of chromosome 8 and 17 tends to occur more in hiPSCs and hESCs, respectively, and these differences were suggested to be a consequence of different culture conditions (Pyle et al., 2006). A study using karyotyping single nucleotide polymorphism (SNP) arrays suggested a higher rate of abnormalities in hiPSCs than in hESCs (Hussein et al., 2011) however the frequency of *de novo* copy number variation appearing in hiPSCs using next-generation sequencing based methods

was very low or absent and suggested that the reprogramming process of iPSCs does not promote mutagenesis is the source of new genomic alterations is most likely inherited from the donor cells or acquired during prolonged culture (Abyzov et al., 2012; Cheng et al., 2012; Young et al., 2012).

1.2.4 Factor induced pluripotency vs somatic nuclear transfer

The efficiency of iPSC reprogramming technology has seen a considerable increase since its discovery however it still remains a poorly understood phenomenon. Several studies have compared reprogramming efficiency with other methods such as that of SCNT and a combined knowledge from different methods might help us better understand the reprogramming process and further improve its efficiency. The quality of reprogrammed cells by iPSC and SCNT technologies has been compared by several groups and the levels of imprinted methylation defects and frequencies of *de novo* point mutations between hiPSCs and SCNT ESCs were found to be very similar (Johannesson et al., 2014; Ma et al., 2014), suggesting that the aberrations observed are unrelated to the reprogramming event and most likely a result of extended cell culture (Cheng et al., 2012).

A global DNA methylation analysis revealed a similar percentage of differentially methylated regions (DMRs) in SCNT ESCs and iPSCs in comparison with ESCs, with some DMRs shared between the two reprogrammed cells indicating the existence of some regions that are resistant to reprogramming regardless of the method used. However the percentages of DMRs associated with donor cell

epigenetic memory may be similar to the total number of DMRs in SCNT ESCs is 5-fold less than in iPSCs indicating that SCNT may be a superior method with regards to erasure of epigenetic memory (Ma et al., 2014). These results go against the conclusions from two other studies that suggest that the major determinant of the differences between iPSCs and SCNT ESCs are their genetic background (Johannesson et al., 2014; Kyttälä et al., 2016). Further studies are needed to resolve the issue of whether SCNT are superior reprogramming technique.

The effect of epigenetic memory on developmental potential has been investigated by serial cloning using both SCNT and iPSC technologies. While mice could be generated for six generations via serial rounds of iPSC reprogramming and tetraploid complementation (Gao et al., 2015), mice could be cloned for 25 generations via SCNT (Wakayama et al., 2013). These studies suggest that we have still a lot to learn about the reprogramming properties from the oocyte. A combination of newly identified oocyte factors may help boost iPSC generation efficiency which often is the first choice in reprogramming approach due to its lower technical and infrastructural requirements without the ethical dilemma of embryo manipulations.

1.2.5 Mechanistic studies of iPSC reprogramming

Despite extensive work around the world on iPSC reprogramming, mechanistic studies have been hampered by the low efficiency and asynchronous nature of the current protocols. Many models have been proposed to explain how the

majority of cells are able to initiate reprogramming but only few cells overcome the barriers required to achieve full pluripotency (Yamanaka, 2009). The first model suggested a stochastic process in which all starting cells have the capacity to generate pluripotent progeny with different time latencies which are amenable to acceleration (Hanna et al., 2009). This model was later refined separating the initial stochastic stage of reprogramming from a hierarchical phase which are more predictable (Buganim et al., 2012).

A number of reprogramming roadblocks have been described as their elimination or attenuation increase the efficiency and kinetics of cell fate conversion. Remarkably the MBD3/NuRD repressor complex has been described as the main epigenetic barrier and the depletion of MBD3 protein permits a synchronised and fully deterministic process of reprogramming in which virtually all induced cells from mice and human origin are reprogrammed within 7 days (Rais et al., 2013). These results have been the source of much debate as other groups were unable to reproduce the methodology with such efficiency (Lujan et al., 2015). In direct contradiction the MBD3/NuRD repressor complex has also been suggested to facilitate somatic cell reprogramming (dos Santos et al., 2014). Nevertheless the supplementation of the reprogramming media with ascorbic acid and GSKβ inhibitor was reported to facilitate more synchronous and rapid iPSC formation from several mouse cell types with a remarkable possibility of generating chimera-competent iPSCs after as little as 48 h of OKSM expression (Bar-Nur et al., 2014). Similarly a study using a combinatorial modulation of signaling pathways approach with small molecules revealed that the inhibition of TGF-β together with

activation of Wnt signaling in the presence of ascorbic acid greatly facilitated reprogramming of mouse embryonic fibroblasts to pluripotency at a rate over 80%. Moreover mouse hepatic and blood progenitors required only TGF-β inhibition or canonical Wnt activation, respectively, achieving reprogramming efficiencies of nearly 100% (Vidal et al., 2014).

The combination of cell surface markers and reporter genes has greatly facilitated the identification and isolation of reprogramming intermediate populations which permitted the delineation of epigenetic and transcriptomic events on the route to pluripotency (O'Malley et al., 2013a; Polo et al., 2012). Surprisingly the process of induction of the pluripotency network is not constrained to activation of only pluripotency genes at the levels observed in ESCs. In some cases pluripotency genes must first be down-regulated before being up-regulated later on (Polo et al., 2012; Tiemann et al., 2014). RNA-sequencing analysis of these populations demonstrates two waves of pluripotency gene up-regulation (O'Malley et al., 2013a). However, to further complicate the understanding of the sequence of events occurring during reprogramming, gene expression signatures totally unrelated to the identity of both the starting and ending cell types can be observed at intermediate stages of reprogramming. Transient expression of several epidermis specific markers have been reported before the mesenchymal to epithelial transition (MET) at an early phase of reprogramming of mouse embryonic fibroblasts demonstrating that reprogramming is not simply the reversal of the normal developmental processes (O'Malley et al., 2013a). Conversely several groups have described the transient activation of developmental regulators at intermediate

stages of reprogramming suggesting that the route to pluripotency follows normal development stages but in reverse order (Takahashi et al., 2014), although this observation is not replicated in more efficient reprogramming systems (Bar-Nur et al., 2014; O'Malley et al., 2013a; Polo et al., 2012) suggesting that the activation of of these developmental regulators might be noise inherent in stochastic action of reprogramming factors.

These and many other inconsistencies are most likely due to differences in reprogramming methods and or starting cell types. Critically many genes and/or cellular functions considered to be absolutely essential for the reprogramming process are being now reported to be dispensable and system specific (Chantzoura et al., 2015). For example, the *Nanog* gene was widely considered the gateway to pluripotency and without its expression the acquisition of pluripotency was deemed impossible (Silva et al., 2009). Recent studies have now shown that Nanog null mouse embryonic fibroblasts (MEFs) can give rise to iPSCs in the presence of Vitamin C (VitC) (Schwarz et al., 2014) or with 100-fold less efficiency compared to wild-type MEFs in the absence of VitC (Carter et al., 2014). Nanog reprogramming dependency could also be bypassed by the combined enhanced activation of STAT3 and *Klf4* expression (Stuart et al., 2014).

1.3 Aims and Objectives of This Study

Understanding the molecular mechanisms underlying somatic cell reprogramming to pluripotency is key for the predictable and efficient generation of high-grade pluripotent stem cells and may be useful for therapeutic applications. Despite the significant progress since the seminal discovery by Takahashi and Yamanaka the current methods for factor induced pluripotency are still inefficient and non-deterministic. The natural route to pluripotency in early embryos must occur within a short time frame for development to progress successfully and somatic cells can follow this path as demonstrated by somatic cell nuclear transfer experiments. Therefore reprogramming lessons can be learned from this natural process.

The objective of this study was to identify new factors capable of improving the efficiency and kinetics of reprogramming somatic cells to pluripotency by harnessing the information from the commonalities between gene expression networks from early embryo development and somatic cells being reprogrammed to pluripotency. This work was divided in to two main parts:

1.3.1 Identification and validation of a new oocyte reprogramming factor

The first phase of this study focused on the analysis of an amphibian oocyte transcriptome from which a maternal reprogramming candidate factor was identified.

The selected gene was experimentally tested and validated to be able to improve the speed and reprogramming efficiency of different mouse somatic cell types when over-expressed together with conventional Yamanaka factors *Pou5f1*, *Sox2*, *Klf4* and *Myc*.

1.3.2 Gene network analysis of reprogramming and normal development

The second phase of this study focused on the gene co-expression network analysis (Zhang and Horvath, 2005) of publicly available RNA-Seq data sets of different induced pluripotency reprogramming studies as well as for the initial stages of mouse development up to morula stage. This analysis was able to reveal conserved network modules which were highly reproducible between reprogramming data sets and some of these modules were also present during normal development. The hub genes of the preserved network modules were identified as a mixture of some well known pluripotency factors along with a number of genes with no reported links to pluripotency which could be used as new reprogramming factor candidates.

Chapter 2

Materials and Methods

2.1 Cell Culture

2.1.1 Media and Cell Culture Reagents

ES Cell Standard Medium - Serum/LIF

1 x Glasgow Minimum Essential Medium (GMEM) (Sigma cat. no. G5154)
2 mM L-glutamine (Invitrogen cat. no. 25030-024)
1 mM sodium pyruvate (Invitrogen cat. no. 11360-039)
1 x non-essential amino acids (Invitrogen cat. no. 11140-035)
10% (v/v) fetal bovine serum (Invitrogen cat. no. 10270-106)
0.1 mM β-mercaptoethanol
LIF ,1000U/mL (ESGRO Mouse LIF Millipore, ESG1106)

2i/LIF Medium

1:1 Neurobasal (Gibco 21103-049) :DMEM:F-12 (Gibco, 21331-020)
Non-essential amino acids (100x), 1:100 (Gibco, 11140-035)
N2 Supplement (100x), 1:200 (Gibco, 17502-048)
B27 Supplement (50x), 1:100 (Gibco, 17504-044)
β-mercaptoethanol, 1:1000 of stock solution (Sigma, 60-24-2)
Penicillin-Streptomycin, 1:1000 (Gibco, 15070-063)
PD0325901, 1 μM (Signalling Technologies, University of Dundee)
CHIR99021, 3 μM (Signalling Technologies, University of Dundee)
LIF ,1000U/ml (ESGRO Mouse LIF Millipore, ESG1106)

Freezing medium

Added DMSO (dimethyl sulphoxide, tissue-culture grade, Sigma cat. no. D-2650) to ES cell standard medium to a final concentration of 10% (v/v). Filter sterilised. Made fresh before use.

PBS

Dulbecco's phosphate-buffered saline (PBS) without calcium and magnesium (Invitrogen cat. no. 14190-094).

β-mercaptoethanol stock solution

Added 70 μl of β-mercaptoethanol (Sigma cat. no. M-7522) to 20 ml of distilled, deionised water (Invitrogen cat. no. 15230-071). Filter sterilised and stored at 4°C for up to 2 weeks.

Trypsin solution

Added 100 mg of EDTA tetrasodium salt (Sigma cat. no. E-6511) to 500 ml of PBS. Filter sterilised and added 10 ml (for 1 x) or 20 ml (for 2 x) of 2.5% trypsin solution (Invitrogen cat. no. 15090-046) and 5 ml of chicken serum (Invitrogen cat. no. 16110-082). Stored in 20 ml aliquots at -20°C. (Note: 2.5% trypsin solution should be aliquoted and stored at -20°C to avoid multiple freeze-thawing cycles.)

0.1% gelatin

Add 25 ml of 2% bovine gelatin solution (Sigma cat. no. G1393) to 500 ml of PBS. Store at 4°C.

Antibiotics

Puromycin (Sigma, P9620): 1.0 μg/ml

Hygromycin B (Gibco, 10687-010): 400 μg/ml

Geneticin - G418 sulphate (Thermo Fisher, 10131035): 200 μg/ml

2.1.2 Cell lines

E14Tg2a

Mouse embryonic stem cell line derived from 129/Ola mice (ATCC CRL-1821). This cell was used as a reference pluripotency gene expression.

BT12

E14Tg2a-derived ES cells carrying a targeted deletion of Nanog (Chambers et al., 2007a). Express hygromycin resistance under the control of the endogenous Nanog-promoter. Undifferentiated ES cells can be selected using hygromycin. These cells were used as a reference for pluripotency gene expression in a *Nanog* –/– background.

Nanog –/– pre-iPS cells

Nanog –/– pre-iPS cells are Nanog–/– neural stem (NS) cells which were transduced with pMXs-based retroviral reprogramming factors (Takahashi and Yamanaka, 2006b) and have not completed reprogramming to pluripotency which can be rescued by the forced expression of a *Nanog* transgene (Silva et al., 2008). These cells constitutively express green fluorescent protein (GFP). Cells were a gift from Jose Silva, Cambridge University.

TNG-A pre-iPS cells

TNG-A pre-iPS cells are mouse embryonic fibroblasts from a genetically modified mouse which were transduced with pMXs-based retroviral reprogramming factors (Takahashi and Yamanaka, 2006b) and have not completed reprogramming to pluripotency. These cells have a GFP-IRES-Pac reporter/selection cassette knocked in one of the Nanog alleles (Chambers et al., 2007b) which can be used for monitoring and selecting reprogrammed cells in naïve pluripotency. These

cells complete reprogramming to pluripotency at a very low rate and are used for assessing transgenes for their capacity of improving reprogramming efficiency. Cells were a gift from Jose Silva, Cambridge University.

2.1.3 Cell Culture Routine

All cell lines were maintained in 37°C, humidified incubators maintained at 7% CO2.

Passage and expansion of mammalian cell cultures

Mammalian cell lines were routinely cultured on gelatin-coated tissue culture T25 flasks or 6 well plates and were passaged when they reached about 80-90% confluency. Cells growing in culture flasks/dishes were washed twice with 1X PBS and incubated in 0.05% Trypsin (Sigma) for 2-3 minutes at 37 °C until cells were uniformly dispersed into small clumps.. Cells were dissociated to single cells by pipetting the solution up and down a couple of times. The cell solution was transferred into a 20 ml centrifuge tube containing 5-10 trypsin volumes of media to inactivate the trypsin and centrifuged at 300 x g for 3 minutes. Following centrifugation, the supernatant was aspirated, pelleted cells were resuspended in fresh culture media and counted using a haemocytometer. Then, cells were transferred into a new cell freshly gelatinized culture T25 flask, splitting the passaged flask in a 1:3 ratio.

Cryopreservation of cell lines

Following Trypsin incubation and counting, cells were centrifuged at 300 x g for 10 minutes. Then, pelleted cells were quickly resuspended in freshly prepared pre chilled freezing medium. Cells were resuspended at concentration of 0.5-3.0 x 10^6 and 1 ml aliquots of cells transferred into cryovials (Nunc, 377224) then placed inside a passive freezer (Nalgene Mr. Frosty, product no. C1562) filled with isopropyl alcohol and incubated at –80 °C overnight before being transferred into liquid nitrogen for long-term storage.

Thawing cell lines

A single dish of a 6-well plate was coated per sample with 0.1% gelatin and aspirated off immediately before use. Frozen cell vials were thawed in a 37°C water bath and then transferred to 10 ml of pre-warmed ES cell medium and mixed well by swirling. The cells were pelleted by spinning for 3 minutes at 1200 rpm in a bench-top centrifuge.The medium was aspirated off and the cells gently resuspended in 5ml of pre-warmed medium and transferred to the 6-well plate and grown at 37°C in a humidified 7% CO2 incubator. The medium was changed the following day to remove dead cells and residual DMSO.

2.1.4 Somatic Cell reprogramming

Reprogramming *Nanog* –/– pre-iPS cells

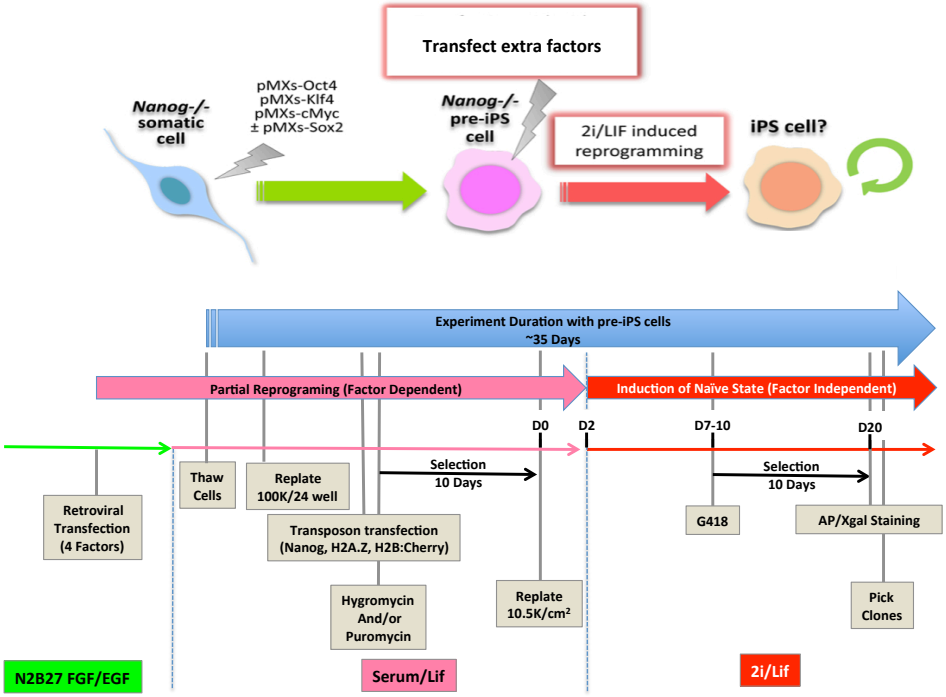


Fig. 2.1 Reprogramming Nanog–/– pre-iPS cells

A graphical representation of the reprogramming assay of *Nanog* null cells is shown in Figure 2.1. Frozen pre-iPS cells were thawed in ES cell medium (serum/LIF) and two days later $1.5x10^5$ cells were replated per well of a 24 well plate. Transfection with transposon vectors was performed on the following day and a 10 day drug selection initiated 48 hours later with the appropriate antibiotics. The stably transfected cells were replated at $1.5x10^4$ cells per centimetre square density to perform the reprogramming assay. Two days later the media was replaced with serum free N2B27 2i/LIF. G418 selection [200 μg/ml] was started after 7 to 10

days of replating. G418 selects for complete reprogrammed cells via the selection cassette inserted into one of the *Nanog* locus of the *Nanog* –/– cells and is used for eliminating background pre-iPS cells. No iPS cells emerged in empty vector transfectants whether or not geneticin was applied. After 10 days of selection alkaline phosphatase and X-gal staining were performed at on technical replicate wells and colonies were picked for further analysis.

Reprogramming TNG-A pre-iPS cells

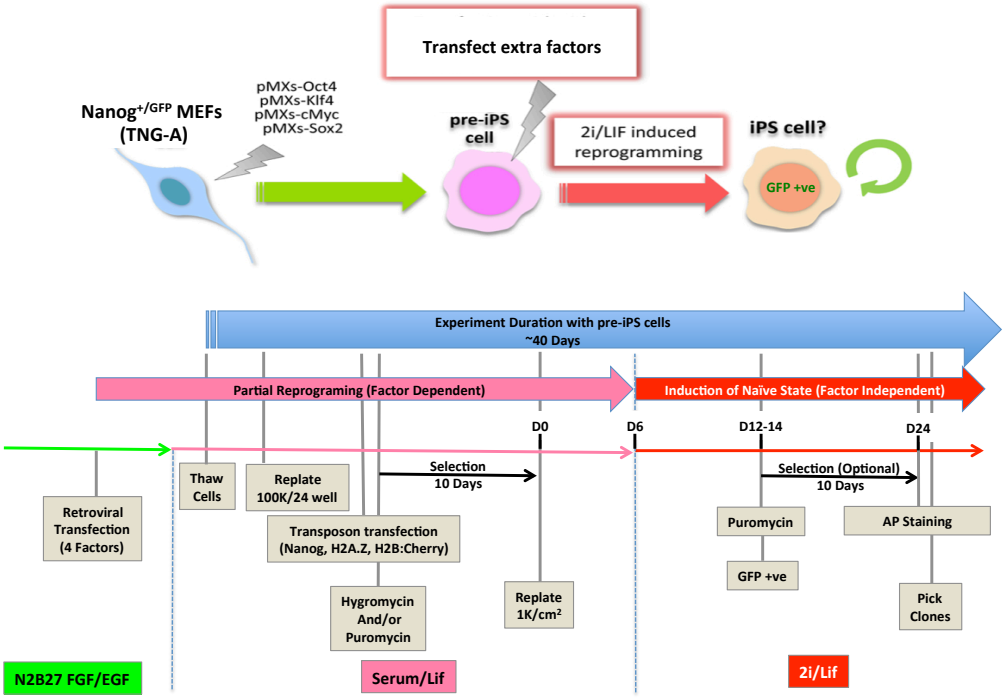


Fig. 2.2 Reprogramming TNG-A pre-iPS cells

A graphical representation of the reprogramming assay of MEF derived pre-iPS cells is illustrated in Figure 2.2. Frozen pre-iPS cells were thawed in ES cell medium (serum/LIF) and two days later $1x10^5$ cells were replated per well of a 24 well plate.

Transfection with transposon vectors was performed on the following day and a 10 day drug selection initiated 48 hours later with the appropriate antibiotics. The stably transfected cells were replated at $1.5x10^3$ cells per centimetre square density to perform the reprogramming assay. Six days later the media was replaced with serum free N2B27 2i/LIF. GFP positive cells began appearing after 6 to 8 days after media switch and at this time point puromycin selection [1μ/ml] was applied to select for expression of the GFP-IRES-Pac reporter transgene from *Nanog* locus. Alkaline phosphatase staining was performed at day 24 on technical replicate wells and colonies were picked for further analysis.

Transfection – Lipofection

Pre-iPS cells were stably transfected with 1μg of PB-CAG-loxP-Transgene-loxP-PGK-Hygromycin plus 2μg PBase expression vector, pCMV-hyPBase (Yusa et al., 2011) using Lipofectamine LTX Reagent with PLUS Reagent (Invitrogen, 15338030) following manufacturer's protocol.

Alkaline Phosphatase Staining

Cells were stained for alkaline phosphatase using the Alkaline Phosphatase (AP) kit (Sigma 86R-1KT) following standard manufacturer's protocol. Stained colonies were scored by eye using a dissecting microscope.

X-gal Staining

0.1M phosphate buffer (pH 7.3) for X-gal staining:

3.74 g monobasic sodium phosphate, (NaH2PO4-H2O; FW = 137.99)

10.35 g dibasic sodium phosphate (Na2HPO4; FW = 141.96)

1L H_2O

Fix buffer for X-gal staining:

0.1M phosphate buffer (pH 7.3)

5 mM EGTA (Sigma cat. E4378), pH 7.3

2 mM MgCl2

0.2% glutareldahyde (Sigma act. num. G-7776)

Wash buffer for X-gal staining:

0.1M phosphate buffer (pH 7.3)

2 mM MgCl2

X-gal staining buffer:

0.1M phosphate buffer (pH 7.3)

2 mM MgCl2

5mM potassium ferrocyanide (K4Fe(CN)6-3H20 Sigma cat. num. P-9287)

5mM potassium ferricyanide (K3Fe(CN)6 Sigma cat. num. P-8131)

Before use, X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) was added to a

final concentration of and filter sterilised to minimise the formation of crystals in the staining reaction.

X-gal stock solution (50mg/ml):

Added 10 ml dimethylformamide (BDH cat. No.103224J) to a 500 mg bottle of X-gal.

Cell culture medium was aspirated off and replaced with X-gal Fix buffer and kept for 15 minutes. The cells were washed twice (5 minutes each) with X-gal Wash buffer. X-gal staining buffer containing 1 mg/ml X-gal was added just enough to the cover cells and the covered plate was then placedinside a humidified chamber and incubated overnight at 37°C. The next day I aspirated off the staining solution and added Wash buffer.

2.2 Molecular Biology

2.2.1 RNA Extraction

0.5-2 million cells were harvested by trypsinisation, washed in PBS and lysed in buffer RLT (Qiagen RNeasy kit). Lysates were snap-frozen in Ethanol/dry ice and stored at -80C or used directly for RNA extraction. Lysates were homogenised by spinning through a Qiashredder (Qiagen, 79654) and subjected to RNA extraction using the RNeasy Kit (Qiagen, 74104) according to the manufacturer's protocol and

incorporating an on-column DNase digest (Qiagen, 79254) to remove genomic DNA. Eluted RNA concentration was determined by spectrophotometry.

2.2.2 cDNA Synthesis

cDNA was synthesized using the Invitrogen Superscript III First-Strand Synthesis Kit (Invitrogen, 11752050) according to the manufacturer's protocol using oligo-dT primers. 0.5-2g RNA was used as template. cDNA was diluted 1:10 in sterile water and used for PCR and qRT-PCR.

2.2.3 Quantitative PCR

Gene expression levels were determined by quantitative PCR (qPCR) carried out with Applied Biosystems 7500 Fast Real-Time PCR System using the TaqMan Fast Universal PCR Master Mix (Applied Biosystems) and FAM-labelled TaqMan gene expression assays following manufacturer's protocol. Average threshold cycles were determined from triplicate reactions using the Applied Biosystems software and the relative levels of gene expression were normalized to GAPDH (VIC-labeled endogenous control assay) and arbitrarily assigning a sample as the reference to which all other samples were compared. Errors were calculated from technical replicates.

Table 2.1 Applied Biosystems TaqMan RT-qPCR custom assays

Gene	**Forward primer** (5'–3')	**Reverse primer** (5'–3')	**Probe** (FAM 5'–3' MGB)
Endogenous Oct4	TTCCACCAGGCCCCC	GGTGAGAAGGCGAAGTCTGAAG	CCCACCTTCCCCATGGCT
Retroviral Oct4	TGGTACGGGAAATCACAAGTTTGTA	GGTGAGAAGGCGAAGTCTGAAG	CACCTTCCCCATGGCTG

Table 2.2 Applied Biosystems TaqMan RT-qPCR standard assays

Gene	**Probe ID**
Gapdh	4352339E
Esrrb	Mm00442411_m1
Nanog	Mm02384862_g1
Nr0b1	Mm00431729_m1
Rex1	Mm03053975_g1
Tfcp2l1	Mm00470119_m1
H2afz	Mm02018760_g1
Vps72	Mm00656969_g1
P400	Mm00551485_m1
Srcap	Mm00613213_m1

2.2.4 Gene Synthesis

The coding sequences for H2A.Z, H2A, H2B, H3 and H4 with flanking Gateway sites were synthesised by the Invitrogen GeneArt Gene Synthesis service and their sequences are illustrated below.

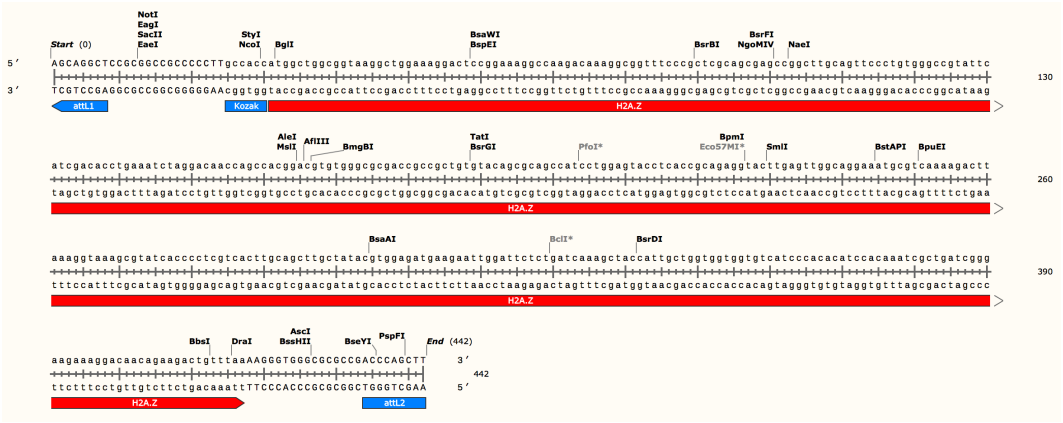


Fig. 2.3 *H2afz* gene synthesis.



Fig. 2.4 *Hist1h2aa* gene synthesis.



Fig. 2.5 *Hist1h2bj* gene synthesis.

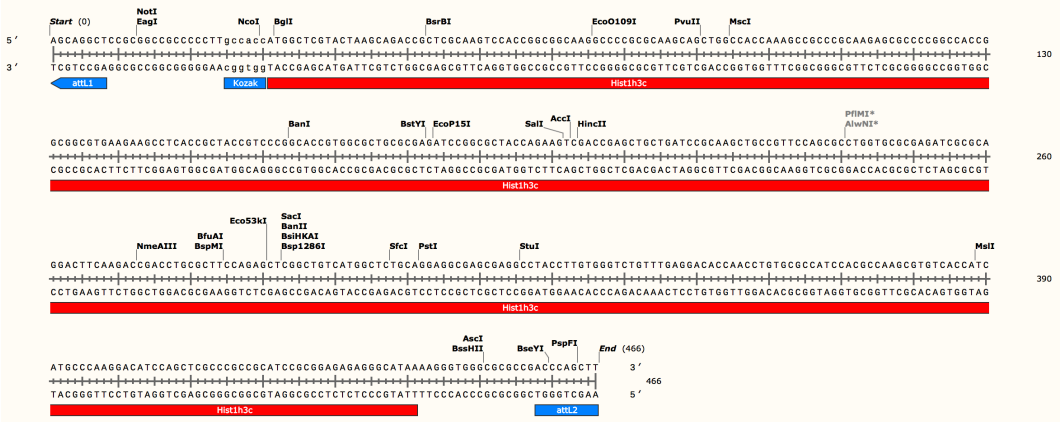


Fig. 2.6 *Hist1h3c* gene synthesis.



Fig. 2.7 *Hist1h4b* gene synthesis.

2.2.5 Gateway cloning

The Gateway adapted synthetic gene fragments were cloned into the piggyBac expression vector pPB-CAG-DEST-pA-pgk-hph (Gift from Jose Silva, Cambridge University), which contains compatible Gateway sites. Gateway cloning was performed using the Gateway LR Clonase II Enzyme Mix following manufacturer's protocol (Invitogen, Catalog No. 11791-020).

2.2.6 Transformations

Plasmid DNA was transformed into One Shot TOP10 Chemically Competent *E. coli* (Invitogen, Catalog No. C404010) 50μL cells were thawed on ice, incubated on ice with 1μL plasmid DNA for 20 minutes, heat shocked for 30 seconds at 37°C, placed on ice for 2 minutes, 1ml Luria Broth (LB) media added and tubes shaken at 37°C for 1 hour before plating 10 or 100μL on the appropriate selective plates. Plates were incubated overnight at 37°C.

2.2.7 Plasmid Purification

Single bacterial colonies were picked from fresh plates into the 5ml of LB containing the appropriate selection (100μg/ml ampicillin) and shaken overnight at 37C. Plasmid DNA was isolated by Miniprep (Qiagen, 27106) or Maxiprep (Qiagen, 12162) according to the manufacturer's protocol.

2.2.8 Plasmid Sequencing

The sequencing primer BGH_rev_primer (5'-TAGAAGGCACAGTCGAGG-3'), which anneals at the Bovine Growth Hormone poly-adenylation sites of all piggyBac transposon vector constructs, was used for sequencing the cloned transgenes.

2.2.9 Plasmids

The Gateway adapted piggyBac destination vector pPB-CAG-DEST-pA-pgk-hph and the *(*Nanog) piggyBac expression vector were a gift from Jose Silva, Cambridge (Theunissen et al., 2011). The pCMV-hyPBase vector for expression of the hyperactive piggyBac transposase was a gift from Kosuke Yusa, Cambridge (Yusa et al., 2011). The maps of the plasmids used are illustrated in figures below.

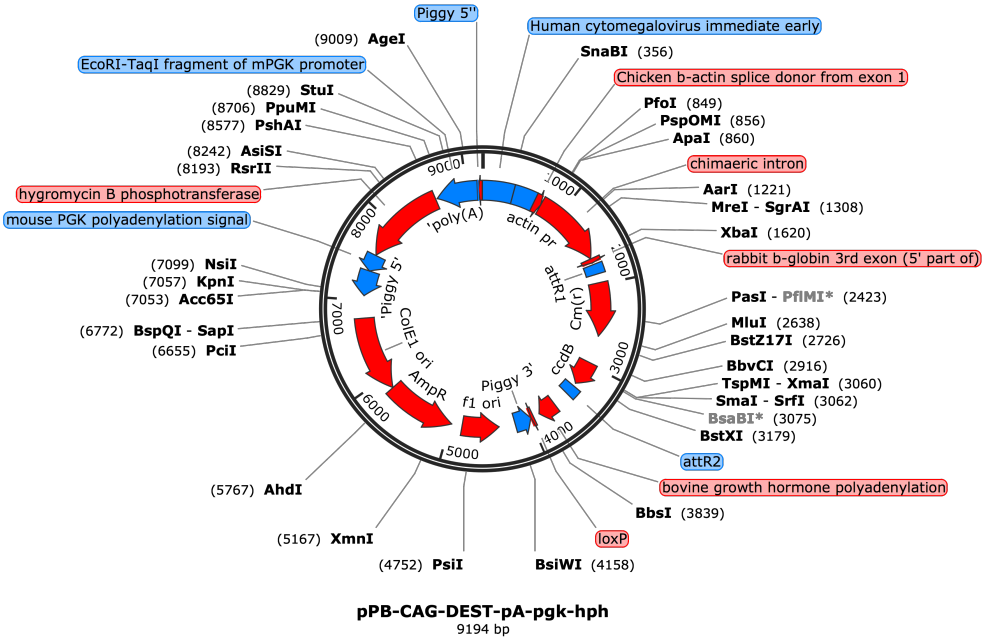


Fig. 2.8 pPB-CAG-DEST-pA-pgk-hph

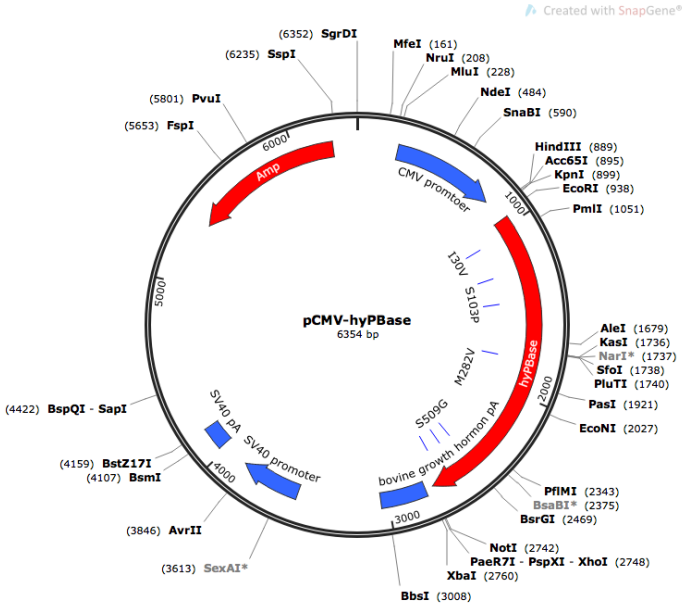
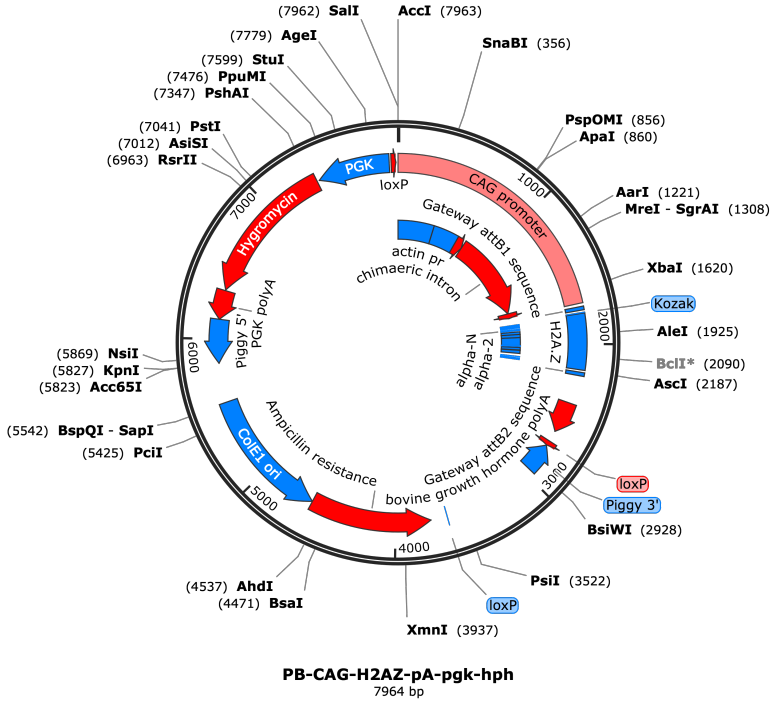


Fig. 2.9 pCMV-hyPBase



PB-CAG-H2AZ-pA-pgk-hph
7964 bp

Fig. 2.10 PB-CAG-H2AZ-pA-pgk-hph

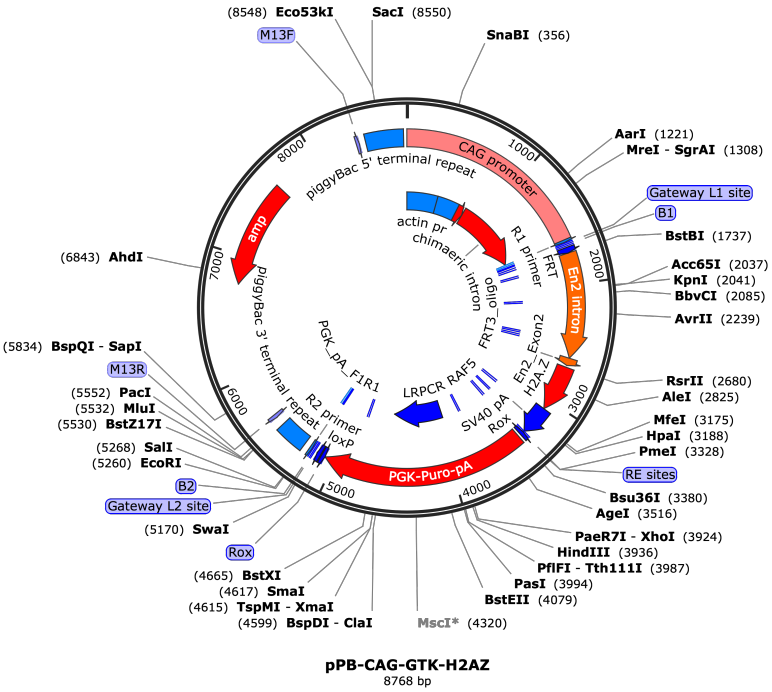


Fig. 2.11 pPB-CAG-GTK-H2AZ

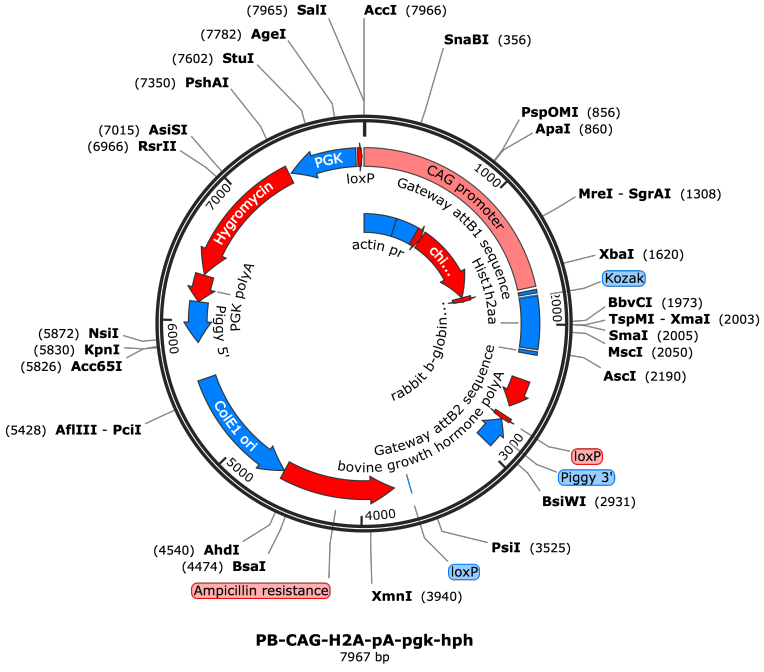


Fig. 2.12 PB-CAG-H2A-pA-pgk-hph

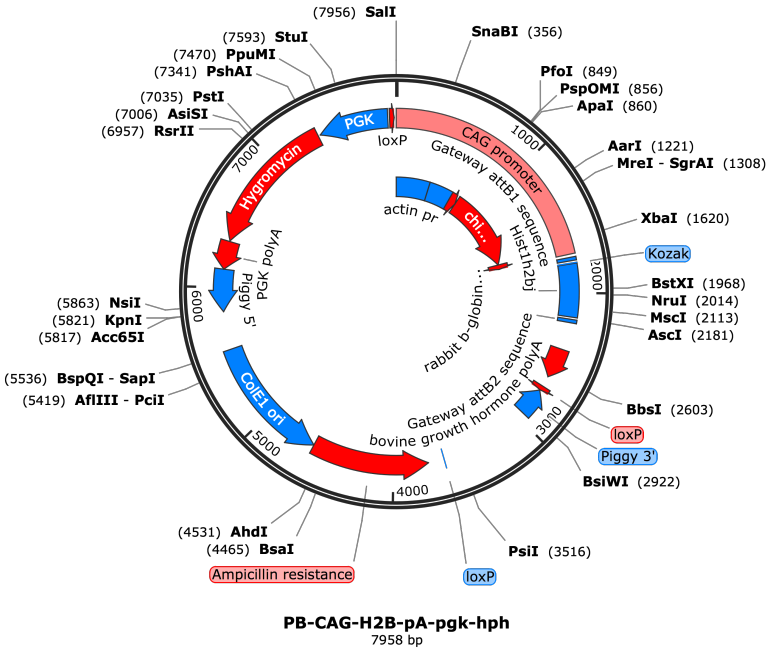


Fig. 2.13 PB-CAG-H2B-pA-pgk-hph

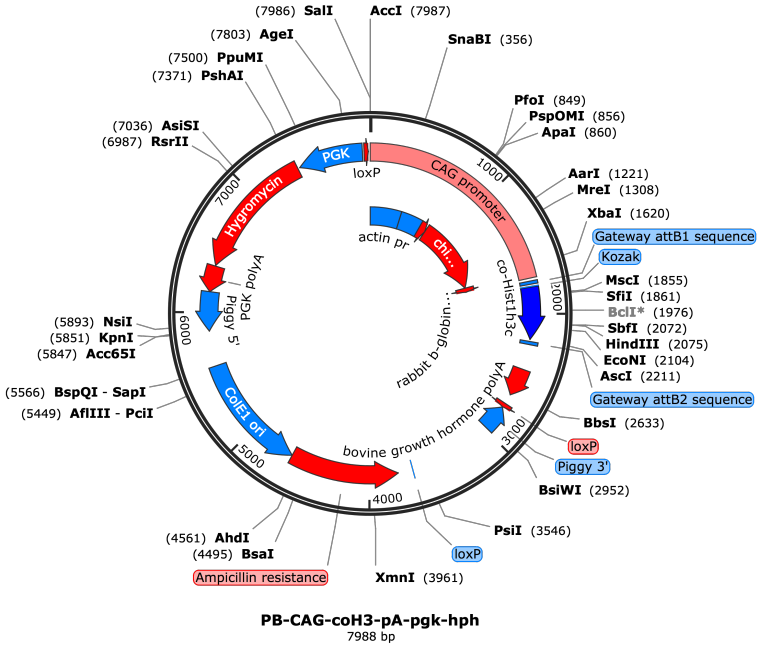


Fig. 2.14 PB-CAG-coH3-pA-pgk-hph

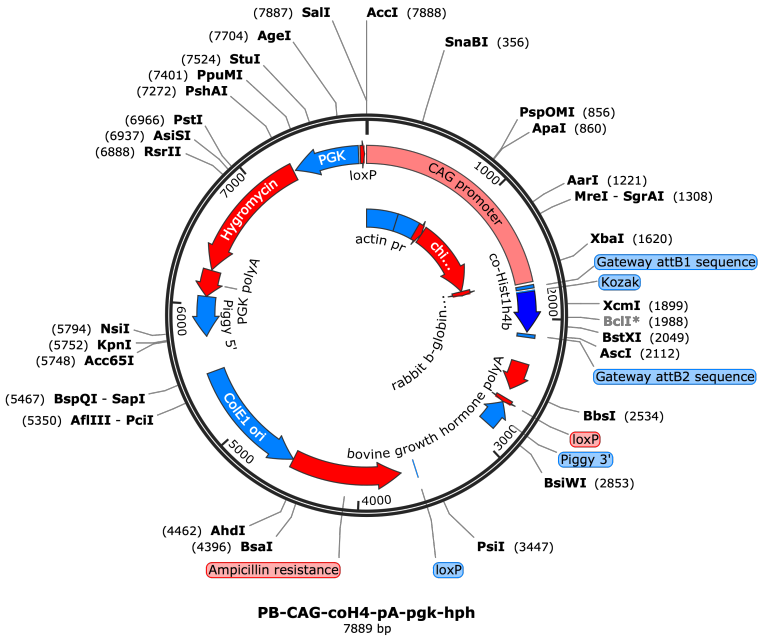


Fig. 2.15 PB-CAG-coH4-pA-pgk-hph

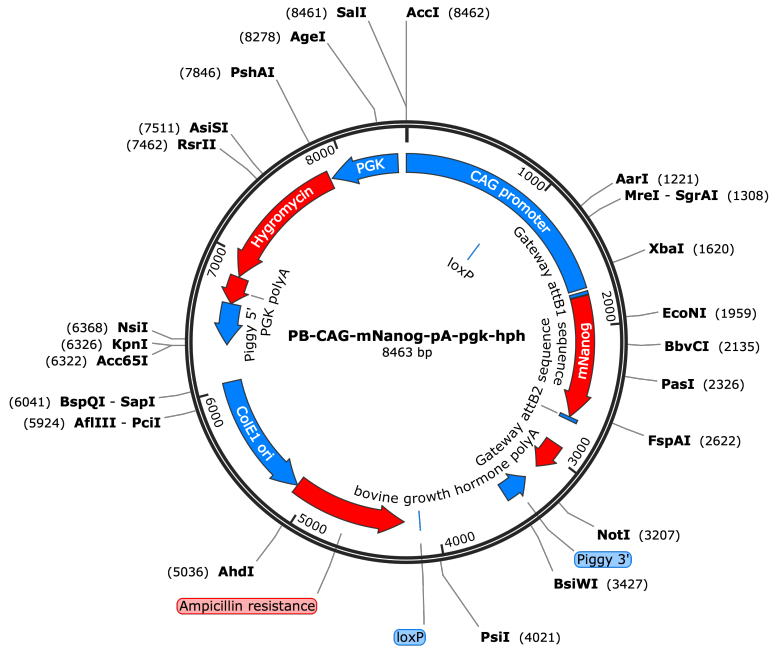
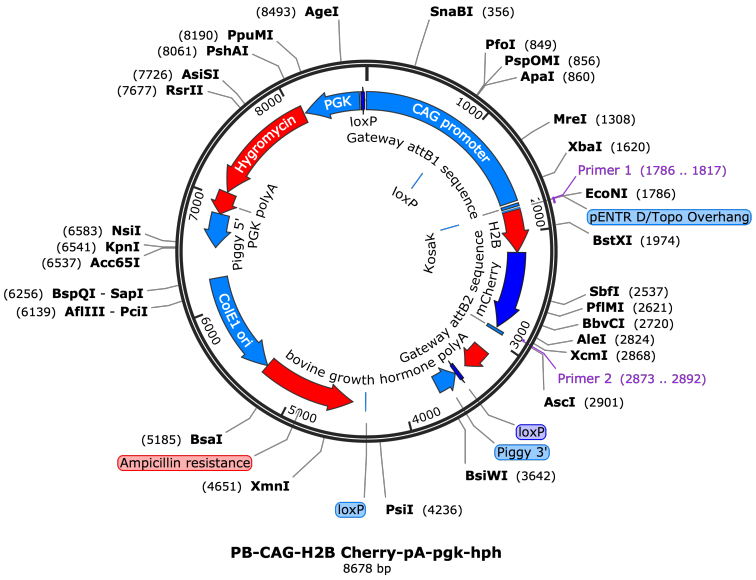
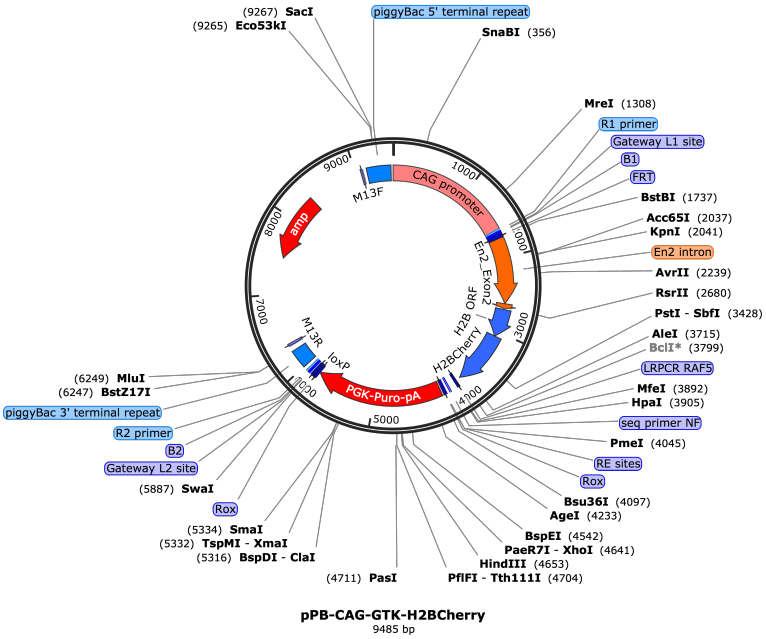


Fig. 2.16 PB-CAG-mNanog-pA-pgk-hph



PB-CAG-H2B Cherry-pA-pgk-hph
8678 bp

Fig. 2.17 PB-CAG-H2B-Cherry-pA-pgk-hph



pPB-CAG-GTK-H2BCherry
9485 bp

Fig. 2.18 pPB-CAG-GTK-H2BCherry

2.3 Computational Methods

2.3.1 Dataset pre-processing

All expression data sets were downloaded from Gene Expression Omnibus (GEO) or ArrayExpress repositories. The data were processed using Bioconductor packages in R. Different data sets were processed slightly differently. Data from (Chantzoura et al., 2015) was downloaded as gene raw counts data (Supplementary data from GSE70816) and normalized by regularised log transformation using the DESeq2 R package (Love et al., 2014). The genes with counts of less than 10 in 20% of the samples were removed. Data from O'Malley et al. (2013b) was downloaded as raw FASTQ files and processed as described in the data set manipulation section of Chapter 4. Data from Xue et al. (2013) was downloaded as a pre-processed RPKM expression data (Assession GSE44183) and genes with low expression in all samples (average RPKM < 0.1) were filtered out, followed by quantile normalisation. Specific details are given in the respective chapters.

2.3.2 Weighted gene co-expression network analysis (WGCNA)

All three downloaded data sets were analysed with weighted gene co-expression network analysis through the R package WGCNA (Langfelder and Horvath, 2008). Analyses was carried out following the methodologies described in by Horvath (2011). A signed weighted gene co-expression networks were made for each data set separately. The nodes of such a network correspond to gene expression profiles,

and edges between genes are determined by the pairwise correlations between gene expressions. To construct the co-expression networks, I firstly produced adjacency matrices A based on Pearson correlations between all gene pairs. Their component A_{ij} was measured by the function, $A_{ij} = (0.5 + 0.5 \times \mathrm{cor}(x_i, x_j)/2)^{\beta}$, where $\mathrm{cor}(x_i, x_j)$ denotes Pearson correlation for expression profiles of gene i and j and β is a user defined value selected in order to emphasise strong correlations and suppress noisy small correlations. The power β of 12, which is the default value, is interpreted as a soft-threshold of the correlation matrix. The adjacency matrices are then transformed into topological overlap matrices (TOM), which is a robust and biologically meaningful measure of network interconnectedness, as it reflects not only direct adjacency in gene pairs but also their interconnections. The degree of overlaps was used to build hierarchical trees with average linkage method. The tree branches were then classified into gene modules using the Dynamic Tree Cut algorithm (Langfelder et al., 2007). Then I merged modules when their eigengenes (the first principal component of module expression pattern) show strong correlation. Based on the properties of different data sets, slightly different parameters were used which is described in more detail in their respective chapters.

Module preservation analysis

To test whether gene modules identified in a data set were conserved in other data sets, I assessed their preservation degrees with the following two methods. I firstly examined overlaps between modules derived from different data sets

and estimated their statistical significance from the hypergeometric distribution (Fisher's Exact Test). I then utilised a composite preservation statistics proposed by WGCNA so as to calculate $Z_{summary}$ scores. Module preservation analysis uses a permutation test to define a test statistic that summarises the evidence that the network connectivity (intramodular hub gene status) and density (how tight interconnections among genes in a module are) of the module is preserved in an independent data sets (Langfelder et al., 2011) The $Z_{summary}$ scores correspond to degrees of how much their composition and structure are preserved in the network between test and reference data sets. In general this module preservation statistics can be interpreted as $Z_{summary}$ scores > 10 indicate high preservation, $Z_{summary}$ scores < 2 generally no evidence of preservation. $Z_{summary}$ scores which are above 2 and below 10 indicate the modules are weakly to moderately preserved.

Hub gene analysis

Eigengenes summarise the expression profiles for modules, and show strong correlation with a number of genes in their modules. Therefore in order to choose intramodular hub genes, I calculated the module eigengene based connectivity (κME) also known as module membership (MM) values. κME = cor(E_q,x(i)), where E_q denotes eigengene of module q and x(i) denotes expression pattern of gene i. Because I used signed co-expression networks I expect that module genes have significant positive MM values. The advantage of using a correlation to quantify module membership is that this measure is naturally scaled to lie in the interval [-1, 1] and a corresponding statistical significance measure (P value) can be easily

computed. Genes were considered module hubs when their MM (κME) values were greater than 0.80.

Gene Set Enrichment Analysis

I performed gene set enrichment analysis using the selected top hub genes (κME > 0.8) of each module and their corresponding κME scores using Enrichr (Chen et al., 2013a), a web-based gene list enrichment analysis tool.

Chapter 3

Mining Oocyte Transcriptome For Reprogramming Molecules

3.1 Introduction

3.1.1 Maternal reprogramming factors from the oocyte

The powerful reprogramming capacity of the oocyte is widely recognised. Immediately after fecundation the sperm chromatin is decondensed, the sperm protamines are replaced with oocyte histones and the paternal DNA gets demethylated and gene expression networks get activated. All those events must occur in a fast and orchestrated way for the embryo development to progress appropriately. In nuclear transfer experiments the donor nucleus goes through very similar epigenetic reprogramming events. In order to be successful therefore the study

of maternal factors and its mechanisms of chromatin reprogramming is highly relevant for the understanding and improvement of somatic cell reprogramming methods.

A number of oocyte factors responsible for these early reprogramming events have been identified. The maternal chaperone nucleoplasmin has been shown to promote the replacement of paternal protamines for maternal histones H2A and H2B before DNA replication (Gillespie, 2000). Moreover the histone H3.3 was shown to be deposited in the paternal chromatin by the maternal chaperone Hira and this event was required for DNA replication and ribosomal transcription in the zygote (Lin et al., 2014). These active chromatin remodelling events have also been observed in early SCNT experiments (Prather et al., 1990). The replacement of somatic histones for oocyte specific variants such as murine H1foo and *Xenopus laevis* B4 was shown to precede the transcriptional activation of core pluripotency factors such as Oct4 and Sox2 (Gao et al., 2004; Jullien et al., 2010; Teranishi et al., 2004).

The study of mammalian oocytes in reprogramming has been hampered by their restricted availability and alternative sources such as amphibian oocytes have provided valuable insights. Much work has been done with *Xenopus laevis* eggs (Metaphase II) and oocytes (Prophase I) in the context of somatic nuclear reprogramming and it has been shown that both have the capacity for transcriptional reactivation of pluripotency genes

While *Xenopus* oocytes have been successfully used in reprogramming experiments, evidence from our lab points to the use of the Salamander Axolotl (*Ambystoma mexicanum*) as a preferred alternative. The reactivation of Nanog expression during reprogramming of somatic cells is considered one of the most stringent landmarks of successful reprogramming to pluripotency (Silva et al., 2009). Importantly, Nanog expression fails to be reactivated when mammalian fibroblasts are incubated in *Xenopus* oocyte extracts (Bian et al., 2009; Halley-Stott et al., 2010) but it is readily up-regulated even before Oct-4 reactivation when incubated in axolotl oocyte extracts (Alberio et al. personal communication). Indeed, Axolotl oocytes have been more efficient than frog oocytes at promoting the removal of repressive chromatin marks H3K9me3, H3K27me3 and the chromatin-associated protein HP1 as well as the the demethylation of promoters from pluripotency associated genes (Bian et al., 2009). Remarkably the Axolotl oocyte already expresses Oct4 (Bachvarova et al., 2004), Sox2 and Nanog (Dixon et al., 2010), and these factors are stored as proteins in the oocyte (Alberio et al. personal communication), giving this reprogramming model a distinct advantage of combining the chromatin remodelling activities inherent in oocytes and a unique pluripotency gene network resembling that of an embryonic stem cell. The transcriptome of Axolotl oocyte has been recently sequenced and the histone variant *H2a.fz* stood out as one of the most expressed genes (Unpublished data, Loose M., Johnson A.). The unusual properties of this histone variant and its association with a pluripotency regulatory network prompted us to study its possible role in somatic cell reprogramming.

3.1.2 H2A.Z, one histone variant with many important functions

The histone H2A.Z is a variant of the canonical histone H2A (West and Bonner, 1980) and its implication in a large variety of biological functions has attracted massive interest from the scientific community in recent years (Mehta et al., 2010; Svotelis et al., 2009; Zlatanova and Thakar, 2008).

This non-allelic variant has been associated with functions at times contradictory, such as transcription activation (Abbott et al., 2001), repression (Guillemette et al., 2005), transcription elongation (Krogan et al., 2003), insulation of euchromatin from spread of heterochromatin (Meneghini et al., 2003), silencing of X chromosome (Greaves et al., 2006) and centromere (Greaves et al., 2007; Rangasamy, 2003), higher order chromatin organization (Fan et al., 2002, 2004) and mediation of RNA interference (Zofall et al., 2009).

H2A.Z has also been shown to be important for the maintenance of genome stability (Papamichos-Chronakis et al., 2011) with roles in DNA double strand break (DSB) repair (Kalocsay et al., 2009; Madigan et al., 2002; Xu et al., 2012), homologous recombination (Fritsch et al., 2004; Wu et al., 2007), nucleotide excision repair (NER) (Yu et al., 2013) and proper chromosome segregation (Krogan et al., 2003; Morillo-Huesca et al., 2010; Rangasamy et al., 2004; Sharma et al., 2013).

Cell proliferation control is another key function of H2A.Z (Svotelis et al., 2009). This histone has been found to have a role in cell cycle regulation by directly controlling the gene expression of cyclins (Dhillon et al., 2006). Importantly, it is also involved in the suppression of senescence by repressing the p53/p21

pathway (Bellucci et al., 2013; Gallant-Behm et al., 2012; Gevry et al., 2007; Lee et al., 2012).

Due to its important roles in gene expression regulation, genome stability and cell cycle regulation, its not surprising that H2A.Z has increasingly been implicated in the development of oncogenesis. The over-expression of this histone variant and/or its aberrant genomic incorporation has been found in several types of cancer, especially at the metastatic stage (Chevillard-Briet et al., 2014; Conerly et al., 2010; Dunican et al., 2002; Hua et al., 2008; Rangasamy, 2010; Rhodes et al., 2004; Zucchi et al., 2004). Indeed, high levels of H2A.Z facilitate activation of estrogen-responsive genes in metastasis-associated events (Dalvai et al., 2012; Hua et al., 2008; Marques et al., 2013; Rangasamy, 2010; Svotelis et al., 2010).

3.1.3 H2A.Z is highly conserved in evolution

The histone variant H2A.Z was first identified in mammalian cells in 1980 (West and Bonner, 1980). The protein structure of this non-allelic histone is highly conserved in evolution, which suggests an important role in chromatin structure and function (Thatcher and Gorovsky, 1994). Orthologs of H2A.Z have been identified in a variety of organisms as diverse as the budding yeast *Saccharomyces cerevisiea* (Htz1) (Illingworth et al., 2012), the fission yeast Schizosaccharomyces pombe (Pht1) (Carr et al., 1994), protozoans *Tetrahymena termophila* (hv1) (Allis et al., 1986) and the malaria parasite *Plasmodium falciparum* (H2A.Z) (Miao et al., 2006; Trelle et al., 2009), the fruit fly *Drosophila melanogaster* (H2AvD) (van Daal

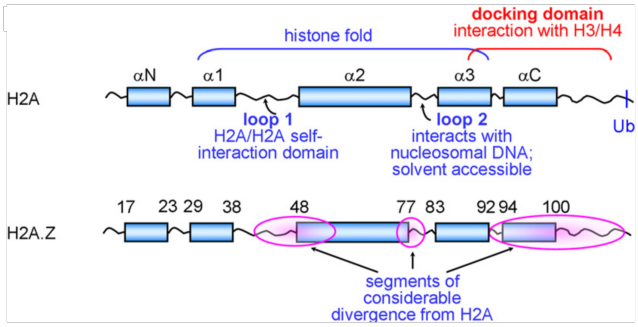
and Elgin, 1992), *Caenorhabditis elegans* (Htz-1) (Whittle et al., 2008), *Xenopus leavis* (H2A.ZI) (Iouzalen et al., 1996), Axolotl (Personal communication, Andrew Johnson), chicken (H2af/z) (Harvey et al., 1983), pig (H2AFZ) (Zhang et al., 2009) and even the flowering plant *Arabidopsis thaliana* (HTA8, 9, 11) (March-Díaz et al., 2007; Yi et al., 2006) to name a few common experimental models.

3.1.4 Differences in structure and expression regulation between H2A.Z and H2A

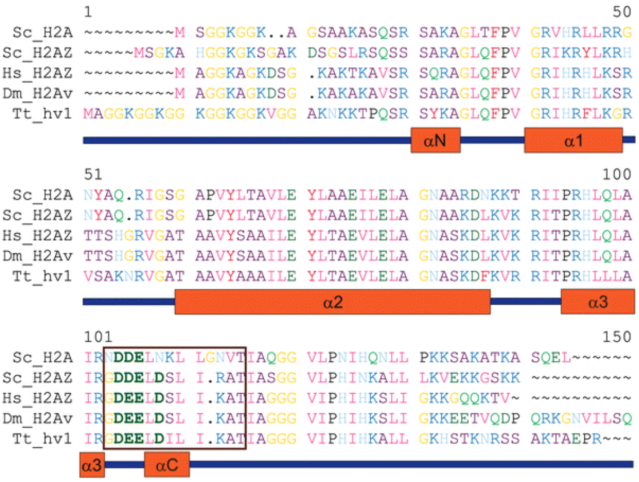
Phylogenetic studies indicate that the histone variant H2A.Z appeared once very early in evolution and is more closely related to other H2A.Z orthologs than to the canonical histone H2A from the same species (Thatcher and Gorovsky, 1994). Between species the H2A.Z protein sequence identity can be up to around 90% but only about 60% in relation to H2A in the same organism (Iouzalen et al., 1996) (Figure 3.1). H2A and H2A.Z diverge most significantly in two key structural regions: a) the boundaries of the Loop1/-helix 2 and -helix 2/Loop2 within the canonical histone fold which are involved in H2A/H2A self-interaction and nucleosome/DNA interaction; b) The C-terminal tail and the acidic patch in the C-terminal -helix which serve as a docking domain for the interaction with H3/H4 (Zlatanova and Thakar, 2008) (Figure 3.1).

Throughout the cell cycle, the gene expression regulation of the histone variant H2A.Z also contrasts markedly with the regulation of its canonical counterpart.

While histone H2A is only synthesized in the G0 and S-phase, the non-allelic H2A.Z histone is also expressed in G1 and G2 mammalian cells (Wu et al., 1982).



(a) H2A and H2AZ structural differences. Adapted from (Zlatanova and Thakar, 2008)



(b) H2A and H2AZ alignment. Adapted from (Jensen et al., 2011)

Fig. 3.1 Differences between H2A and H2AZ.

3.1.5 H2A.Z divergent structural domains are essential for developmental viability

Despite not being essential for yeast viability (Carr et al., 1994; Santisteban et al., 2000), an indispensable role for H2AZ has been demonstrated for the early development of many organisms. The protozoan Tetrahymena cannot survive after the

deletion of the histone hv1 (Liu et al., 1996) and this critical role has been shown to be dependent on acetylatable lysine residues at the N-terminal histone tail (Ren and Gorovsky, 2001).

Furthermore H2AZI *Xenopus* morphants were shown to produce defective gastrulation, leading to malformed short trunk embryos whereas the over expression slows down development after gastrulation (Ridgway et al., 2004). This effect could be reversed by altering the amino acid residues of the C-terminal extended acidic patch of H2AZI (Ridgway et al., 2004).

Likewise in *Drosophila* the homozygous deletion of H2AvD is lethal at larval stage and amino acid swaps for H2A equivalent residues at the acidic patch were sufficient to induce such a phenotype (Clarkson et al., 1999) .

H2A.Z null mice embryos die at the time of implantation presumably as a consequence of defect in proliferation and differentiation of the pluripotent inner cell mass (ICM) (Faast et al., 2001).

3.1.6 H2A.Z roles in mammalian development

In the pre-implantation mouse embryo, H2A.Z was initially thought to begin its expression at the blastocyst stage and be restricted to the trophoblast (Rangasamy, 2003). The ICM presumably lacked H2A.Z and it would only be produced at the onset of differentiation when this histone variant exerts a role in the establishment of heterochromatin (Rangasamy, 2003).

The lack of H2A.Z in the initial phases of mammalian development seems surprising, especially in the pluripotent ICM, given that this histone variant appears to be overall the most abundant histone in R1 embryonic stem (ES) cells (Kafer et al., 2010). Furthermore H2A.Z containing nucleosomes are enriched at the transcription start site (TSS) with direct correlation with transcriptional activity in ES cells (Creyghton et al., 2008). This is in agreement with the initial proposed role of this histone variant. H2A.Z was first linked to transcriptional activation in early Tetrahymena studies due to evidence of its localization at the transcriptionally active macronucleus (Allis et al., 1986; Wenkert and Allis, 1984; Whittle et al., 2008). Similarly studies with chicken erythrocytes suggests a role of H2A.Z acetylation in transcriptional activation (Bruce et al., 2005).

These discrepant observations are possibly due to differences in experimental procedures employed. Indeed a recent study has shown that, apart from the 2 cell stage when zygotic genome activation (ZGA) takes place, acetylated H2A.Z is readily detected in mice embryos throughout the pre-implantation stages and is significantly detected in the ICM euchromatin (Boskovic et al., 2012). Interestingly the same study showed that bovine acetylated H2A.Z embryonic expression pattern also includes the 8 cell stage when ZGA is triggered (Boskovic et al., 2012). This demonstrates that the H2A.Z dynamics may not be conserved in mammalian development (Boskovic et al., 2012).

However, the H2A.Z expression pattern during germ cell development is similar between mice and humans. During mouse germ line development H2A.Z is

detected at high levels in migratory primordial germ cells (PGC) and is rapidly down-regulated after colonization of the genital ridges (Hajkova et al., 2008) and then reactivated during the meiotic transition of spermatogenesis when it can be detected at some regions of facultative heterochromatin (Banaszynski et al., 2010; Greaves et al., 2006). Similarly H2A.Z is highly expressed in human PGCs with highest expression at around 7-9 weeks in the testis and 7.5 weeks in the ovary. Moreover, H2A.Z expression also gets depleted when PGCs reach the gonads, only reappearing at around 17 weeks in some VASA +ve cells in both genders (Gkountela et al., 2013).

3.1.7 H2A.Z as a regulator of gene expression

Genome-wide studies have shown that H2A.Z is generally associated with nucleosomes located in and around promoter and enhancer regions with a role in transcriptional activation of many genes (Hu et al., 2013; Marques et al., 2010). However it appears that H2A.Z may perform this function via distinct mechanisms varying from gene to gene (Sevilla and Binda, 2014). Upon transcription induction the levels of promoter bound H2AZ at RAR- and ER-regulated genes are increased (Dalvai et al., 2013; Hu et al., 2013) whereas it decreases at the regulatory regions of p21 and CCND1 in mammalian cells (Bellucci et al., 2013; Dalvai et al., 2013).

H2A.Z containing nucleosomes are enriched at the flanks of a nucleosome-free region (NFR) at TSSs (Zlatanova and Thakar, 2008) and has been shown to be

able to promote the recruitment of the RNA polymerase II (RNAPII) in yeast and mammals (Adam et al., 2001; Hardy et al., 2009). Moreover, a recent study has shown that H2A.Z facilitates the access of activation or repression complexes to chromatin and is required for both self-renewal and proper differentiation of ES cells (Hu et al., 2013).

Conversely H2A.Z has been previously proposed to be dispensable for ES cell self-renewal and only required for lineage commitment by interfering with the regulation of Polycomb complex target genes (Creyghton et al., 2008; Illingworth et al., 2012). In fact the global H2A.Z landscape around all TSSs in ES cells and multipotent neural stem cells reveled the enrichment at the promoters of active and poised genes and absence from stably silent ones (Ku et al., 2012).

The seemingly contradictory associations of H2A.Z with active and facultative silenced chromatin and its requirement for ES cell self-renewal might be partially explained by post-translational modifications such as acetylation, mono-ubiquitylation and methylation which provides a further layer of gene expression regulation (Binda et al., 2013; Draker and Cheung, 2009; Hu et al., 2013; Sarcinella et al., 2007; Sevilla and Binda, 2014; Talbert and Henikoff, 2010).

The acetylation of the N-terminal of H2A.Z is a hallmark of active genes (Adam et al., 2001; Allis et al., 1986; Nakatani et al., 2009; Santisteban et al., 2000; Shahbazian and Grunstein, 2007) and directly correlates with the level of gene expression in ES cells (Hu et al., 2013). H2A.Z can have its N-terminal tail acetylated at five lysine residues (K4, K7, K11, K13 and K15) (Bonenfant et al., 2007; Boyne et al.,

2006; Bruce et al., 2005; Dryhurst et al., 2009; Ishibashi et al., 2009; Valdes-Mora et al., 2012). However residues 4 and 7 can instead be mono-methylated by the lysine methyltransferase SETD6 and this modification has been proposed to be a silencing mark required for E14 mouse ES cell self-renewal (Binda et al., 2013).

The Polycomb repressive complex 1 (PRC1) subunit RING1B E3 ubiquitin ligase is responsible for ubiquitinating both H2A.Z and its canonical counterpart H2A (Cao et al., 2005; De Napoles et al., 2004; Sarcinella et al., 2007). While H2A can be ubiquitinated at lysine 119, H2A.Z can be mono-ubiquitinated at any one of the three possible C-terminal lysine residues (K120, K121 and K125) (Ku et al., 2012; Sarcinella et al., 2007). This modification was suggested to function as a repressive mark since it is found at transcriptionally silent facultative heterochromatin of the inactive X chromosome (Sarcinella et al., 2007). Furthermore dually modified H2A.Z by both acetylation and ubiquitylation has been found enriched at bivalent chromatin providing further evidence for their proposed regulatory roles (Ku et al., 2012).

SUMOylation is another post-translation modification that can be found in H2A.Z with no known function in gene expression control. This mark does not seem to be conserved in evolution and has only been identified in budding yeast. In *Saccharomyces cerevisiea* the SUMOylation of Htz1 has a role in DSB repair by repositioning the damaged chromatin to the nuclear periphery (Kalocsay et al., 2009).

3.1.8 H2A.Z and chromatin remodelling complexes

Although H2A.Z was first identified in humans most of the functional studies have been performed in *Saccharomyces*, which provides a convenient platform for biochemical studies of H2A.Z function since it is not essential for yeast viability.

Unlike H2A the histone variant H2A.Z is incorporated in the chromatin at non-random, specific positions in a replication independent manner. The molecular mechanism responsible for the site specific recruitment of the H2A.Z histone variant began to unravel around 10 years ago when a novel multi-protein complex was discovered in a genetic screen for chromatin modifying proteins and transcription elongation in yeast (Krogan et al., 2003). Orthologs of most of its components have now been identified in several species including humans (Figure 3.2) (Bao and Shen, 2007; Eirín-López, 2007; Guillemette and Gaudreau, 2006; Korber and Hörz, 2004; Lu et al., 2009; Morrison and Shen, 2009; Svotelis et al., 2009).

	S. cerevisiae	*S. pombe*	*C. elegans*	*A. thaliana*	*D. melanogaster*	Human TIP60	Human SRCAP
SWR1-C	Swr1	swr1	ssl-1	PIE1	Domino	p400	SRCAP
	Bdf1	bdf1	F13C5.2*		dBrd8	Brd8	
	Swc2	swc2	C17E4.6	AtSWC2	dYL-1	YL1	YL1
	Swc3						
	Swc5	swc5					
	Swc6	SPBC29A3.05	CD4.7*	AtSWC6			ZnF-HIT1
	Swc7						
	Arp6	arp6	C08B11.6	AtARP6			Arp6
	Rvb1	rvb1*	T22D1.1	Ruv*	dPontin	TIP49a	TIP49a
	Rvb2	rvb2*	C27H6.2	RuvB*	dReptin	TIP49b	TIP49b
	Htz1	pht1	H2A.Z	HTA8, 9, 11	H2Av	H2A.Z	H2A.Z
		msc1					
	Yaf9	SPAC17G8.07	gfl-1*	AtGAS41	dGas41	GAS41	GAS41
	Swc4	swc4*	ekl-4*		dMAP1	DMAP1	DMAP1
	Arp4	Alp5	ZK616.4*	AtARP4	BAP55	BAF53a	BAF53a
	Act1	act1	ACTin	Actin-1/3	Act87E	Actin	Actin
NuA4	Esa1	mst1	mys-1	HAM2*	dTip60	Tip60	
	Yng2	png1*	ing-3*		dIng3	ING3	
	Epl1	epl1*	epc-1		E(Pc)	EPC1	
	Tra1	tra1*	trr-1	AtPI3,4K*	dTra1	TRRAP	
	Eaf1	vid21*	ssl-1*				
	Eaf3	alp13*	mrg-1*	AtMRG*	dMrg15	MRG15	
	Eaf5						
	Eaf6	SPAC6F6.09*		At4G14385.1p*		hEaf6	
	Eaf7		ZK1127.3		dMrgBP	MRGBP	

Fig. 3.2 SWR and NuA4 complexes have shared evolutionarily conserved components with overlapping functions. Adapted from (Lu et al., 2009)

This complex belongs to the INO80 subfamily of chromatin remodelling complexes, formed by the Swi2/Snf2-related complex (SWR-C) and INO80 complex (INO80-C) which are characterised by a split ATPase domain in the core subunit Swr1 and Ino80, required for H2A.Z deposition and eviction respectively (Bao and Shen, 2007; Kobor et al., 2004; Krogan et al., 2004; Mizuguchi et al., 2004; Papamichos-Chronakis et al., 2011).

SWR-C is composed of 15 subunits including Swr1, the main element (Lu et al., 2009). Swr1 protein is the largest component of the complex with two important regions, the ATPase domain and HSA (helicase–SANT-associated) which are required for the interaction with many of the SWR-C subunits (Wu et al., 2005, 2009). The interaction with H2A.Z is mediated by the subunit Swc2, but may also occur via a N-terminal domain of Swr1 (Wu et al., 2005, 2009).

Several lines of evidence indicate that the site-specific recruitment of SWR-C can be mediated by transcription factors and some particular post-translational modification of histones in the target nucleosomes. The SWR-C share four subunits with NuA4, another chromatin modifying complex responsible for the acetylation of the histones H2A, H2A.Z and H4 indicating related functionality (Allard, 1999; Babiarz, 2006; Choi et al., 2009; Galarneau et al., 2000; Keogh et al., 2006; Kobor et al., 2004; Krogan et al., 2003; Millar et al., 2006; Mizuguchi et al., 2004; Zhang et al., 2004).

It has been shown that the acetylation of H4 is one key epigenetic mark prior to H2A.Z deposition (Altaf et al., 2010). It has an important role in the recruitment

of bromodomain containing SWR-C which catalyzes the ATP dependent histone dimer exchange H2A/H2B for H2AZ/H2B (Kobor et al., 2004; Krogan et al., 2004; Mizuguchi et al., 2004). Consistent with this, lower H2A.Z levels in promoter regions and reduced global association of SWR-C is observed when NuA4 catalytic subunit Eaf1 is mutated (Babiarz, 2006; Durant and Pugh, 2007). A similar effect is observed with H4 mutants with N-terminal residues that prevent acetylation. Moreover, mutations in the primary NuA4-dependent acetylation sites of H4 (H4 K5, 12R) cause a similar loss of H2A.Z from chromatin (Raisner et al., 2005).

The acetylation of H3 histone N-terminal tail has also been linked to having a role in SWR-C recruitment and H2A.Z deposition (Raisner et al., 2005). This modification is catalyzed by Tra1, a member of the Spt-Ada-Gcn5-acetyltransferase (SAGA) histone acetyltransferase (HAT) complex (Auger et al., 2008; Grant et al., 1998). In fact Tra1 interacts with the NuA4 HAT complex via the SANT domain of Eaf1 subunit (Auger et al., 2008; Grant et al., 1998) and can also contribute to H2A.Z acetylation on lysine 14 (Babiarz, 2006; Keogh et al., 2006; Millar et al., 2006).

3.1.9 H2A.Z related chromatin remodelling complexes are widely conserved in metazoans but may have two distinct orthologs of the yeast SWR-C

Virtually all of the components of SWR-C, NuA4 and Tra1 SAGA complex are conserved from yeast to humans (Figure 3.2) (Lu et al., 2009). However it has been suggested that Swr1 and Eaf1, subunits of SWR-C and NuA4 respectively, may have

been fused into one single protein in the course of higher eukaryote evolution to form the core of the TIP60 (Tat-interactive protein, 60 kDa) complex (Lu et al., 2009) which was first identified as an interacting cellular component with the HIV transactivator Tat protein (Kamine et al., 1996).

This fusion protein has been identified as P400 which contains the ATPase and HSA domains from Swr1 and SANT domain from Eaf1 (Lu et al., 2009). The TIP60 complex was initially identified as a NuA4 related HAT in higher eukaryotes (Doyon et al., 2004; Ikura et al., 2000; Kimura and Horikoshi, 1998; Kusch et al., 2004) but later also shown to be capable of catalyzing the exchange of H2A for H2A.Z (Cai et al., 2003; Doyon et al., 2004; Gevry et al., 2007; Martinato et al., 2008; Svotelis et al., 2009).

Similar to the yeast SWR-C, the TIP60 complex can be recruited to chromatin by transcription factors such as E2F1, MYC, p53 and KAI1 with a role in gene expression and cell cycle regulation of human cell lines (Chan et al., 2005; Frank et al., 2003, 2001; Kim et al., 2005; Nagl et al., 2007; Taubert et al., 2004). MYC was shown to be able to recruit TIP60 and SAGA complexes by direct interaction with p400 and TRRAP (ortholog of yeast Tra1) respectively and these complexes have a positive effect on MYC dependent histone acetylation and affinity for target chromatin (Frank et al., 2003, 2001; Martinato et al., 2008).

The SNF-2-related CREB-binding protein activator protein (SRCAP) complex is the second SWR-C ortholog capable of H2A.Z deposition in chromatin in mammals (Cai, 2005; Eissenberg et al., 2005; Gevry et al., 2007; Kobor et al., 2004; Marti-

nato et al., 2008; Ruhl et al., 2006; Wong et al., 2007). While the TIP60 and SRCAP complexes seem to have the same general role in chromatin deposition of H2A.Z they may control different non-redundant functions by different mechanisms (Lu et al., 2009).

These two complexes share many components such as VPS72/YL1, which is the ortholog of yeast swc2 responsible for recruitment by direct association of free H2A.Z/H2B dimers (Cai, 2005; Wu et al., 2005). However they also have some distinguishing unique subunits (Figure 3.2). On one hand, the bromodomain containing subunit Brd8 as well as the HAT NuA4 and SAGA related subunits Tip60/KAT5, ING3, EPC1, TRRAP, MRG15, hEaf6 and MRGBP are only found associated with the TIP60 complex suggesting that SRCAP does not rely on histone acetylation (Auger et al., 2008). On the other hand ZnF-HIT1 and the actin-related protein Arp6 are unique to the SRCAP complex (Lu et al., 2009).

An emerging idea is that while TIP60 targets chromatin for a more localized H2A.Z deposition, SRCAP would have a more global activity (Lu et al., 2009). SRCAP/DOM has been found to have a role in Notch signalling in organisms as diverse as *Drosophila* and humans (Eissenberg et al., 2005). Furthermore, TIP60 (TRRAP/P400) and SRCAP complexes have been proposed to associate with inactive and active promoter in human cell lines respectively, in agreement with the proposed lack of functional redundancy (Gevry et al., 2007; Wong et al., 2007).

3.1.10 H2A.Z roles in stem cell biology

Chromatin states are widely recognised as being central to cell identity by having a role in establishing gene expression patterns important for cell specific functions. Histone variants have been linked to crucial roles in stem cell biology and their study is gaining momentum in the recent few years. ES cells are known to be transcriptionally hyperactive and have a hyper-dynamic chromatin (Hager et al., 2009; Meshorer et al., 2006).

An important player in these biological signatures is the histone variant H2A.Z which has also been shown to be important for self-renewal and differentiation of ES cells (Creyghton et al., 2008; Hu et al., 2013). Of note H2A.Z displays the highest expression levels over all the canonical histones and its variants in ES cells (Kafer et al., 2010). However it remains largely unknown how this histone variant performs its functions in stem cells and contradictory claims have been frequent.

One of the first ChIP-seq genome-wide analyses suggested an exclusive localisation at Polycomb complex target genes adding a new layer of regulation of developmental genes in ES cells (Creyghton et al., 2008). Loss of function assays indicated that H2A.Z would be responsible for mediating targeting of Polycomb repressor complex 2 (PRC2) 2 and repression of early developmental genes, essential for ES cell differentiation (Creyghton et al., 2008). It was proposed that H2A.Z activity in lineage commitment lays in the divergent amino acids residues Gly92, Asp97, and Ser98 from the c-terminal acidic patch (Subramanian et al., 2013a).

In these studies H2A.Z knockdown had no effect on Oct4 expression levels and did not seem to affect the self-renewal capacity of ESCs even after prolonged culture (Creyghton et al., 2008; Subramanian et al., 2013a). Further more it has been shown recently that neither PRC1 nor PRC2 are required for the targeting of H2A.Z to developmental genes in ES cells (Illingworth et al., 2012).

However the genomic landscape and functionality of H2A.Z seems to be more complex than first anticipated. Hu and colleagues demonstrated that H2A.Z is not only enriched at the promoters and enhancers of developmental bivalent genes but also at active genes which globally comprise over two thirds of all genes (Hu et al., 2013). Similar H2A.Z distribution was previously reported in murine ES and neural stem cells and in human ES cells (Ku et al., 2012) and human multipotent hematopoietic stem cells (Cui et al., 2009).

Interestingly H2A.Z was also shown to be required for both efficient self-renewal and differentiation of ES cells and may function as a general mediator of activation and repression complexes (Hu et al., 2013). Similar to previous reports, knockdown of H2A.Z had little effect on Oct4 expression but compromised significantly OCT4 binding to its target genes (Hu et al., 2013). Moreover after 2 days of H2A.Z depletion there was a significant reduction in the expression of core pluripotency factors such as Sox2 , Esrrb , Tbx3 , and Klf4 and an increase in the expression of differentiation genes such as Brachyury , Hand1 , Nkx2-5 , Bmp4 , Runx1, Wnt7b, Sox4 and Nestin. The inefficient self-renewal was further demonstrated by change in ES cell morphology and reduction of SSEA1 expression (Hu et al., 2013).

Retinoic acid induced differentiation was also compromised by reduced RAR binding and activation of differentiation markers, and by reduced repression of pluripotency genes after H2A.Z depletion (Hu et al., 2013). Similarly RAR has been shown to be essential for both targeting and depletion of Suz12 and H2A.Z in the chromatin of ES cells during differentiation (Amat and Gudas, 2011).

Furthermore, the role of H2A.Z in differentiation and lineage commitment has been demonstrated to work together with the transcription factor Foxa2 by mediating nucleosome depletion at the regulatory sequences of genes being activated during ES cell differentiation (Li et al., 2012). It was shown that Foxa2 and H2A.Z recruit nucleosome disassembly complexes (Nap1l1/SWI/SNF/INO80) which were shown to be required for efficient differentiation (Li et al., 2012).

3.1.11 Contrasting roles of H2A.Z may be determined by chromatin modifying complexes

Although there is no clear consensus regarding the roles of H2A.Z in stem cell biology the consistent requirement of the chromatin modifying complexes involved in H2A.Z deposition, removal and post-translational modifications for the maintenance of pluripotency indicate their essential role in determining the context dependent function of this histone variant.

In a comprehensive loss of function screen for chromatin regulators with a role in mouse ES cell self-renewal identified numerous subunits of the Tip60-p400 complex. The knockdown of these components resulted in reduced proliferation,

alteration in cell-cell adhesion, loss of pluripotency markers and hindered teratoma formation demonstrating their requirement for efficient self-renewal in ES cells (Fazzio et al., 2008). Surprisingly the loss of H2A.Z seemed not to affect the maintenance ES cell pluripotency.

This complex was suggested to act mainly as a repressor of developmental regulator genes based on loss of function results but localizes to over half of promoters in ES cells in ChIP-chip, with stronger correlation with highly expressed genes (Fazzio et al., 2008).

Fazzio and colleagues also observed that the knockdown of Nanog hampered p400 binding to a significant overlapping set of target genes. The targeting of p400 was suggested to be dependent of the histone mark H3K4me3 and Nanog would promote p400 binding indirectly since no interaction of p400 and Nanog proteins could be confirmed. However a recent interactome study reported the physical interaction of Nanog with a number of TIP60-TRRAP-P400 components (Gagliardi et al., 2013).

Interestingly TIP60-TRRAP-P400 components have also been reported to interact with several other important pluripotency factors such as Oct4 and Tcfcp2l1 in ES cells and Myc, Klf4 and Stat3 in cancer cell lines (Ai et al., 2007; Frank et al., 2003, 2001; Martinato et al., 2008; van den Berg et al., 2010; Xiao et al., 2003). Therefore it is tempting to speculate that H2A.Z and its associated chromatin remodeling complexes could have a central role in pluripotency as a mediator of the transcriptional regulatory network.

Consistently TRRAP has been linked to have role in the maintenance of pluripotency, to restrict differentiation in ES cells (Sawan et al., 2013) and is essential for mouse early development (Herceg et al., 2001). Similarly Tip60 was shown to be essential at late blastocyst stage as the loss of its function leads to pre-implantation lethality (Gorrini et al., 2007; Hu et al., 2009). Also recently the SET domain-containing protein 6 (SETD6), which monomethylates H2A.Z lysine 7, was shown to be required for the maintenance of ES cell self-renewal (Binda et al., 2013). Furthermore it has been suggested that Tip60-p400 may be modulated by Hdac6 which does not appear to function as a histone deacetylase in mouse ES cells (Chen et al., 2013b).

These chromatin remodelers are also important for the maintenance of adult stem cells. Hematopoietic stem/progenitor cells (HSC) maintenance is dependent on the histone acetyltransferase cofactor Trrap working via p53 and Myc factors (Loizou et al., 2009). The H2A.Z recruiting factor Vps72 of the TIP60 complex can boost HSC self-renewal when over-expressed (Deneault et al., 2009; Umemoto et al., 2012). Furthermore p400 plays a key role in HSC proliferation by controlling the expression of cell-cycle-regulatory genes (Fujii et al., 2010).

Interestingly some histone variants and chromatin remodelers have also been shown to be important in cell fate restriction upon differentiation, and represent an impediment for reprogramming towards pluripotency. For instance several recent reports have shown that the histone variant macroH2A helps maintain

differentiated cell identity and represents a barrier for the acquisition of pluripotency (Barrero et al., 2013; Gaspar-Maia et al., 2013; Pasque et al., 2011).

Likewise, according to a recent claim the depletion of Mbd3, a core member of the Mbd3/NuRD (nucleosome remodelling and deacetylation) repressor complex, which has been previously shown to prevent ES cell differentiation (Kaji et al., 2006), was able to improve both the kinetics and efficiency of the conventional induced pluripotency reprogramming from some rare events over several weeks (Takahashi and Yamanaka, 2006a) to virtually 100% conversion with 7 days (Rais et al., 2013). However this claim has fulled fierce debate about the validity of its observations (Bertone et al., 2015) and contradictory results (dos Santos et al., 2014) suggest that this effect may at least be context dependent.

Ultimately it would be of great interest to dissect the role of the histone variant H2A.Z in the acquisition and maintenance of pluripotency. Its role in transcriptional activation and cross-talk with key pluripotency factors make H2A.Z an attractive reprogramming factor candidate and may provide new insights about embryonic development including the ontogeny of the germ line.

3.1.12 Objectives

The role of H2A.Z in the maintenance of pluripotency has been a controversial theme and its involvement in the establishment of this stem cell state has not been assessed so far. Significant evidence from early blastocyst, germline development and embryonic stem cell studies suggests an important role in pluripotency. Therefore the specific aims of the work presented in this chapter was to evaluate if H2A.Z is a rate-limiting factor in induced pluripotency by assessing somatic cell reprogramming efficiency and kinetics by using this histone variant as an extra reprogramming transgene factor to the already established Yamanaka factors.

3.2 Results and Discussion

3.2.1 H2A.Z as a rate-limiting factor in reprogramming neural stem cells to pluripotency

To start assessing if the levels of expression of *H2A.Z* are a limiting factor in the establishment of pluripotency I made use of a reprogramming assay established by Theunissen et al. (2011). A graphical representation of the reprogramming assay is illustrated in the Methods chapter Figure: 2.1. This system is based on the premise that a somatic cell cannot be reprogrammed to naïve pluripotency without the expression of the transcription factor *Nanog* (Silva et al., 2009). Therefore a mutant *Nanog* null somatic cell can be reprogrammed with the conventional Yamanka factors only if *Nanog* (Theunissen et al., 2011) or downstream factors such as *Esrrb* are genetically reconstituted (Festuccia et al., 2012).

A constitutive PiggyBac vector (Guo et al., 2009; Kaji et al., 2009; Woltjen et al., 2009) for *H2A.Z* was stably co-transfected with *Nanog* into *Nanog* null mouse neural stem cells which had been partially reprogrammed through infection with retroviruses expressing *Oct4, Klf4* and *Myc* (Cells and vectors were kindly given by Jose Silva, Cambridge University). As experimental controls I transfected vectors for *Nanog* only, *H2A.Z* only and the same for a histone reporter fusion *H2B:Cherry*, which was intended to be a negative control for *H2A.Z* and also for transfection/expression efficiency controls. To be able to select both factors in co-transfected cells I used different PiggyBac expression vectors with a distinct

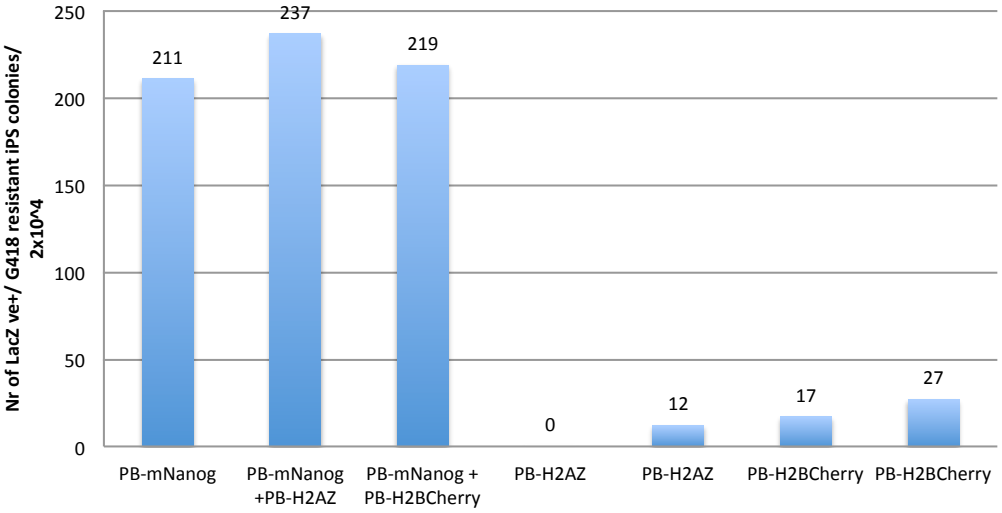
selection cassette for the second factor. Vector maps can be found in the Methods chapter. Of note this vector contains a modified CAGS promoter, which contains a fragment of second intron and second exon of *En2* gene inserted into the end of the of the chimeric intron of the standard CAGS promoter. This modification can drive the transgene expression to a much higher level than the standard CAGS promoter (Dr Barry Rosen, personal communication). The experimental plan is illustrated in the Methods chapter Figure 2.1.

In the first experiment *H2A.Z* over-expression seemed to improve slightly the rescue of *Nanog* -/- towards pluripotency reprogramming (Figure 3.3a). Surprisingly however, both *H2A.Z* and *H2B:Cherry* were able to produce iPS colonies without *Nanog* co-expression. To my knowledge this was unprecedented and of significant importance. While the over-expression of *H2B:Cherry* with both CAGS and modified CAGS could produce some iPS colonies, *H2A.Z* over-expression could only produce iPS colonies with the conventional CAGS expression vectors suggesting that this activity is dependent on expression levels.

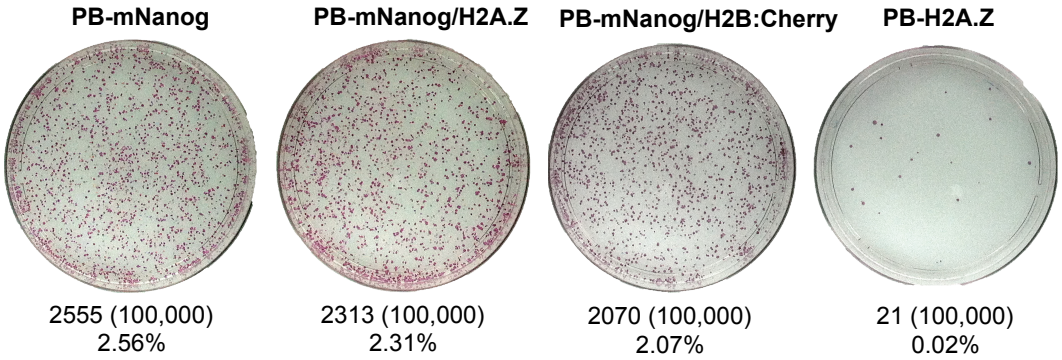
The over-expression of *Nanog* in the pre-iPS cells promoted extensive cell proliferation throughout the reprogramming process and efficiently produced numerous colonies, which may even be underestimated since many colonies were fused to each other. In order to try to get more significant numbers these stably transfected pre-iPS cells were plated at a lower density in larger tissue culture dishes (Figure 3.3c). This time the number of colonies was proportionally lower for the co-delivered *Nanog*/Histone in relation to *Nanog* alone. *H2A.Z* only colony numbers

were also proportionally lower than that of the first experiment, which may also indicate that the cell seeding density might also interfere with the reprogramming efficiency in this conditions. Nevertheless the forced expression of *H2A.Z* together with *Oct4*, *Klf4* and *c-Myc* was able to reprogram *Nanog* null somatic cells.

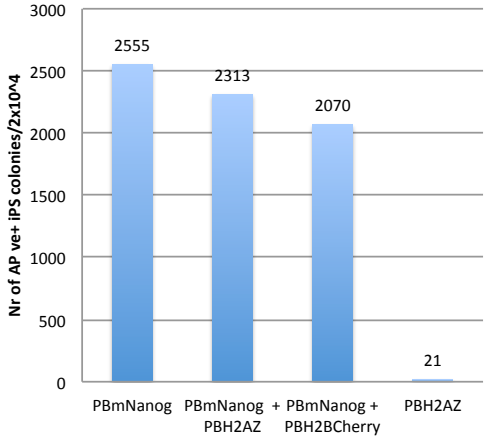
The fact that both *H2A.Z* and *H2B:Cherry* forced expression was able to produce reprogrammed colonies from *Nanog* null somatic cells prompted me to hypothesise that these histones may compete with other endogenous histones with repressive epigenetic marks and result in a reduction in the threshold of endogenous pluripotency gene expression reactivation. If this hypothesis would be proven true, the over-expression of other canonical histones could produce similar effects. I then repeated this experiment with the following set of histones: H2A.Z, H2A, H2B, H2B:Cherry, H3 and H4. The background controls, empty vector and fluorescent protein Cherry, were also included. However only *H2A.Z*, *H2B:Cherry* and *H2B* were able to produce any iPS colonies in a *Nanog* null genetic background confirming the first experiments. Moreover testing for *H2A.Z* as a rate limiting factor was inconclusive in the context of *Nanog* transgene co-delivery given that *Nanog* over expression alone is already a very efficient reprogramming condition and may mask any added effect produced by *H2A.Z*. Nevertheless these experiments revealed the unexpected reprogramming of somatic cells deficient in *Nanog* expression widely regarded as absolutely essential for the acquisition of pluripotency (Silva et al., 2009).



(a) LacZ positive colonies



(b) Alkaline positive colonies



(c) Summary of alkaline positive colonies

Fig. 3.3 **Nanog requirement is bypassed by the over-expression of H2A.Z/H2B:** Nanog null pre-iPS cells used in the reprogramming assay contain a β-geo cassette knocked in into one of *Nanog* loci which permits both the selection with G418 and the staining with Lac-Z when the loci becomes active. (a) Quantification of the total number of LacZ-positive/G418 resistant iPS cell colonies ; (b) Alkaline phosphatase staining of emerging Nanog–/– iPS cell colonies; (c) Quantification of the total number of AP-positive iPS cell colonies.

3.2.2 Characterization of Nanog null iPS cells reprogrammed with H2A.Z

To further characterise these reprogrammed cells, single colonies were picked, expanded and evaluated. As expected, the *Nanog* over-expressing iPS cells appeared to proliferate faster than the *Nanog* null iPS cells reprogrammed with histones only. Interestingly the growth properties of these *Nanog* null reprogrammed cells resemble *Nanog* null ES cells, that is, slower growth and higher propensity to differentiate even in 2i/LIF conditions when no G418 selection is used (Figure 3.4). G418 selects for cells that have reactivated the *Nanog* locus, which selection cassete was knocked in to delete *Nanog* coding sequence (Chambers et al., 2007b). Therefore this selection also prevents differentiation of these mutant cells.

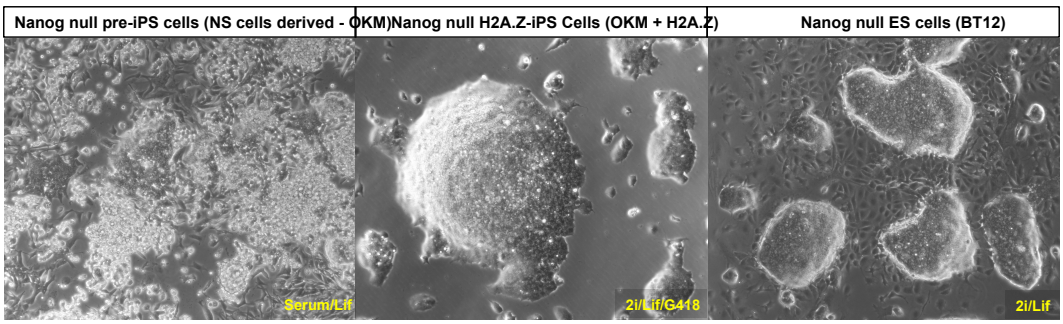


Fig. 3.4 **Morphology comparison of *Nanog* null cell lines:** (**Left**) *Nanog* null pre-iPS cells derived form neural stem (NS) cells which were transduced with retroviruses expressing *Oct4*, *Klf4* and *Myc*. Their morphology is very heterogeneous and does not get reprogrammed without *Nanog* expression. (**Centre**) *Nanog* null iPS cells reprogrammed with *Oct4*, *Klf4* and *Myc* plus H2A.Z. Produces bulky colonies and tend to spontaneously differentiate. G418 selects undifferentiated colonies such as in the figure. (**Right**) *Nanog* null ES cells (BT12) have the same genetic background as the *Nanog* null pre-iPS cells used in the reprogramming assay. Picture illustrates the spontaneous differentiation of these cells when G418 is not applied. Pictures were taken under the same magnification

Clones of reprogrammed iPS cells were also used for assessing their transcriptional expression profiles by real time quantitative polimerase chain reaction (qRT-PCR) for endogenous naïve pluripotency factors and silencing of retroviral reprogramming factors (Figure 3.5). The iPS cells obtained by rescue with *Nanog* ectopic expression displayed equivalent expression levels of the naïve pluripotency markers *Nanog*, *Nr01b* and *Esrrb* in relation to wild type ES cells and *Rex1* expression level was reduced by around two thirds. These levels were more or less similar between the iPS clones reprogrammed with *Nanog*/*H2A.Z* and *Nanog*/*H2B*. The iPS clones reprogrammed with either histone alone had all assessed endogenous pluripotency genes reactivated, however at a lower level in comparison to wild type ES cells. Importantly the expression levels of these genes were remarkably similar between an available Nanog null ES cell line (BT12) and the histone reprogrammed iPS cells. Furthermore all iPS clones had undetectable levels of the transgenic retroviral *Oct4* expression demonstrating efficient silencing of reprogramming factor even in the absence of *Nanog*. The reprogramming of *Nanog* deficient pre-iPS cells by over expression of *H2A.Z* was unprecedented and a complete surprise.

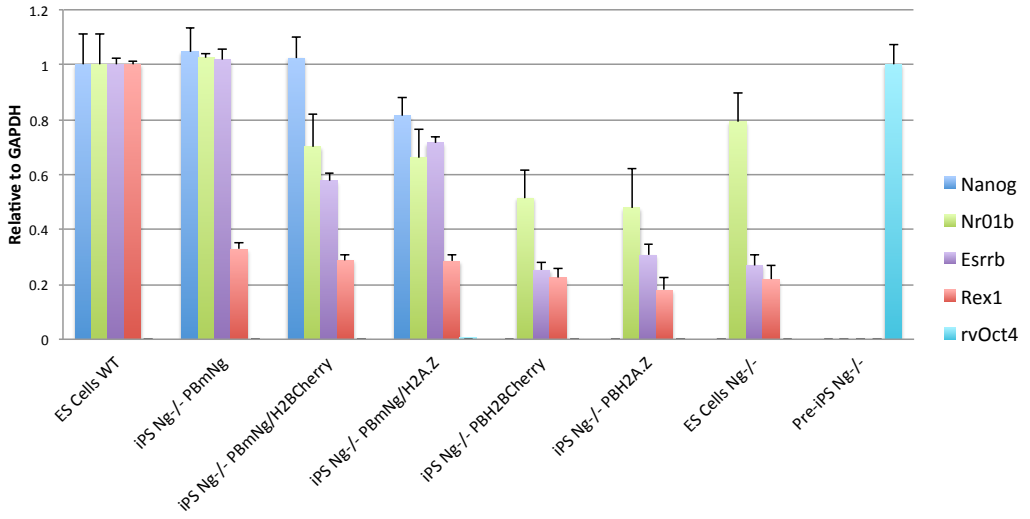


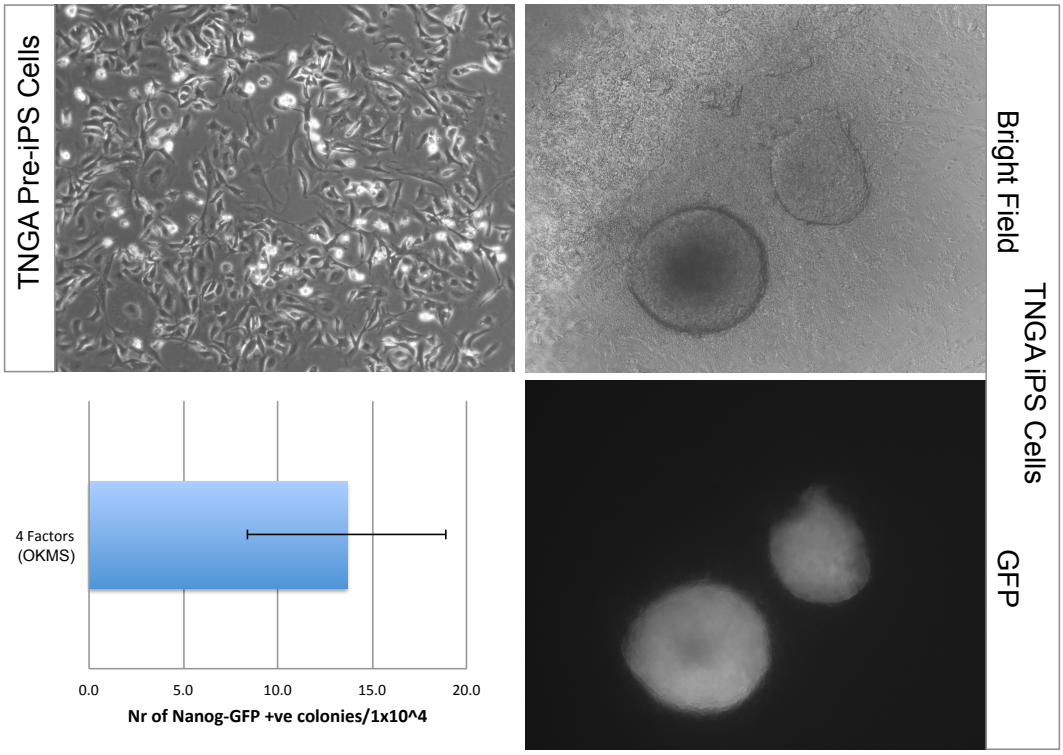
Fig. 3.5 **Expression of pluripotency markers of NS derived iPS cells and its precursors:** qRT-PCR analysis for pluripotency genes and retroviral transgene (rv*Oct4*) expression in *Nanog*–/– pre-iPS cells, *Nanog*–/– ES cells, *Nanog* –/– iPS cells reprogrammed with different combinations with *Nanog*, *H2A.Z* and *H2B:Cherry*. Wild-type ES cells are used as a reference for the expression levels of pluripotency genes and pre-iPS cells are used as a reference for retroviral transgene expression of rv*Oct4*. Data are relative levels to *Gapdh*. Error bars indicate standard errors from three technical replicates

3.2.3 H2A.Z as a rate-limiting factor in reprogramming mouse embryonic fibroblasts to pluripotency

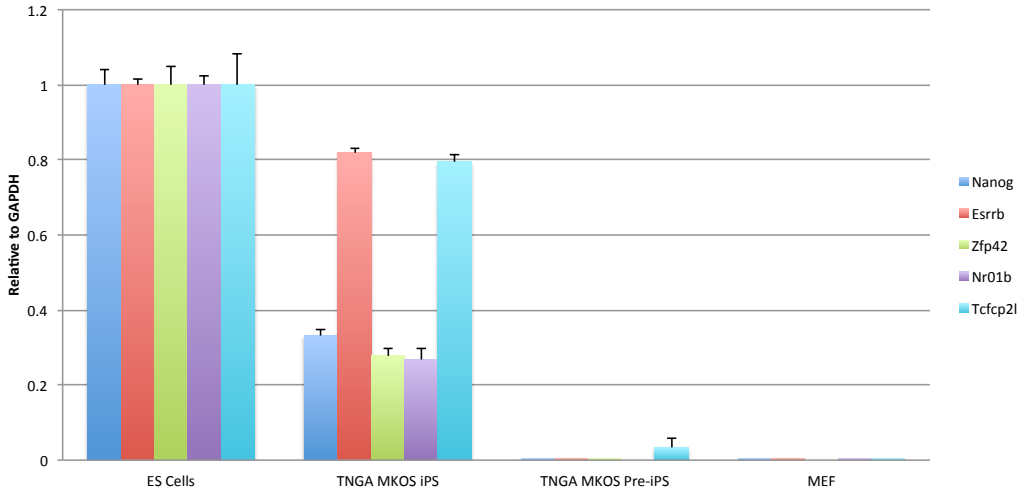
To further test the capacity of enhancing the efficiency of standard reprogramming by over-expression of histones I repeated the previous experiments with mouse embryonic fibroblasts from a genetically modified mouse. A graphical representation of the experimental scheme is illustrated in the Methods chapter Figure: 2.2. These cells have a GFP-IRES-Pac (*Pac* gene confers resistance to puromycin) reporter/selection cassette knocked in one of the *Nanog* alleles (Chambers et al., 2007b) which provide a convenient means of monitoring and selecting repro-

grammed cells in naïve pluripotency (Figure 3.6a) (TNG-A Pre-iPS cells was kindly provided by Jose Silva, Cambridge University). These cells were transduced with the four Yamanaka factors and reprogrammed at a very low efficiency into naïve pluripotency when switched from serum/LIF growth conditions to 2i/LIF serum free media. This conversion took between three to four weeks and the analysed clones displayed reactivation of endogenous *Nanog*, *Esrrb*, *Zfp42*, *Nr01b* and *Tcfcp2* (Figure 3.6b). These factors were only detectable in fully reprogrammed cells with exception of *Tcfcp2*, which was found at very low level in partially reprogrammed cells.

TNG-A MEF derived pre-iPS cells were then used to assess the effect of over-expressing the various histones during iPS cell reprogramming. Consistent with the previous experiments *H2A.Z* over-expressing cells produced significantly higher colony numbers in comparison to all conditions tested (Figure 3.7). They began expressing GFP around 4 days after the positive control *Nanog* forced expressing cells, at which point the selection with puromycin was started. After 10 days of selection colonies were observed only in *H2A.Z* cells and to a much-reduced extent also with *H2A* (Figure 3.7). Consistent with *H2A.Z* reported roles in cell cycle regulation and proliferation, the *H2A.Z* treated cells appeared to proliferate extensively during the period of reprogramming and this feature was not seen with any treatment with other histones (Figure 3.7). Also unique to *H2A.Z* over-expressing cells was the formation of multiple colony foci even before the appearance of *Nanog:GFP* expression, indicating an earlier change in cell adhesion properties.



(a)



(b)

Fig. 3.6 **Reprogramming MEF cells to pluripotency:** TNGA pre-iPS cells have one of *Nanog* alleles knocked in with a GFP expressing cassette which can be used to monitor reprogramming completion. The second *Nanog* allele is intact. (a) Quantification of Nanog-GFP positive colonies. Bar represent the average number of colonies from three experiments and error bar the standard error. (b) Expression of pluripotency markers of MEF derived iPS cells and its precursors. qRT-PCR analysis for pluripotency genes expression in pre-iPS cells and iPS cells reprogrammed with conventional Yamanaka 4 factors. Wild-type ES cells are used as a reference for the expression levels of pluripotency genes. Data are relative levels to *Gapdh*. Error bars indicate standard errors from three technical replicates

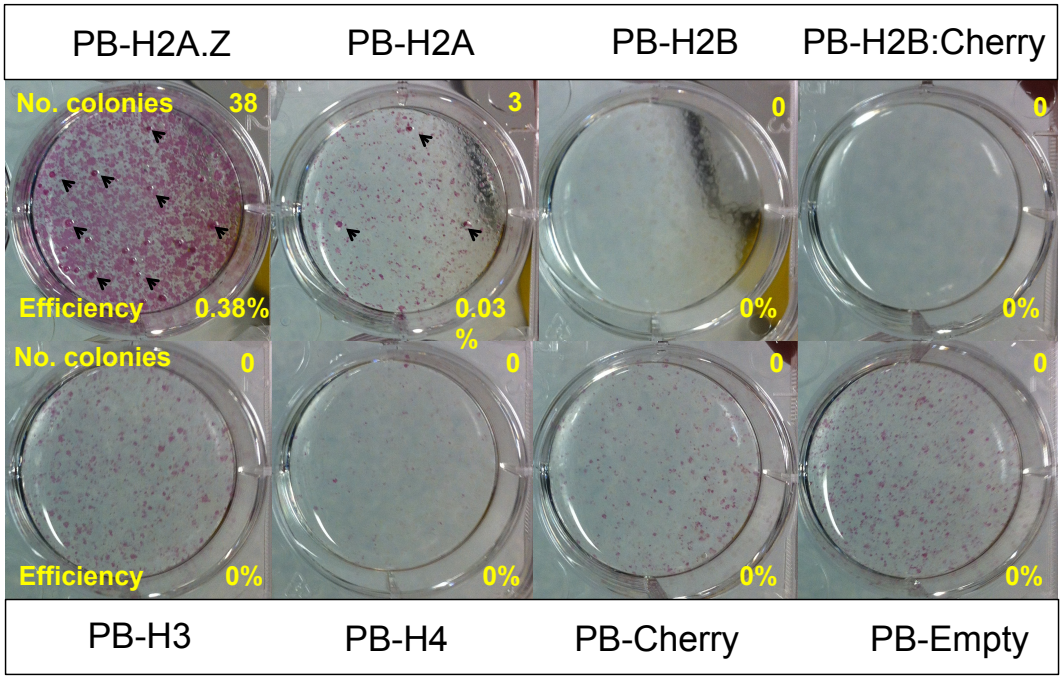


Fig. 3.7 **H2AZ improve reprogramming efficiency:** Alkaline phosphatase (AP) staining of iPS cell colonies generated by over-expressing different histones. Empty piggyBac (PB) vector and PB-Cherry are reprogramming base level controls. Staining was performed on the first day GFP positive colonies were observed in any condition. Arrow heads indicate such GFP positive colonies which differ from AP background cells also by their bulky colony morphology.

The fact that the canonical histones H2A and H2B also have some positive effect in reprogramming was not expected, however it highlights the importance of nucleosomal topology in transcriptional activation. The chromatin remodelling complexes involved in deposition and removal of the histone variant H2A.Z are believed to do so by the exchange of the dimers H2A/H2B for H2A.Z/H2B or vice versa. It is tempting to speculate that either H2A or H2B can become a limiting factor at some point of the cell cycle given that their expression is tightly associated with the G0/S phase (Wu et al., 1982) as opposed to the constitutive nature of the non-canonical histone variant H2A.Z.

Thus, all the results so far, with two different somatic cell types, indicate a role of H2A.Z in the establishment of pluripotency and elucidating its mechanisms would be a significant contribution to the study of pluripotency and may provide important insights about early embryo development including the germline.

3.2.4 Possible mechanisms of deposition of H2A.Z at regulatory regions of key pluripotency genes

In mammals H2A.Z is thought to be deposited at its targets by two possible related chromatin remodelling complexes, Tip60/Trrap/P400 and SRCAP (Gevry et al., 2007; Ruhl et al., 2006). In one hand the complex Tip60/Trrap/P400 is believed to be recruited to chromatin targets via the interaction with transcription factors and the recognition of some histone marks: (Hu et al., 2013), H3K18Ac (Gervais and Gaudreau, 2009) and H4K16ac (Kimura et al., 2002; Suka et al., 2002). On the other hand SRCAP complex seem to deposit H2A.Z more globally and independent of histone acetylation (Lu et al., 2009).

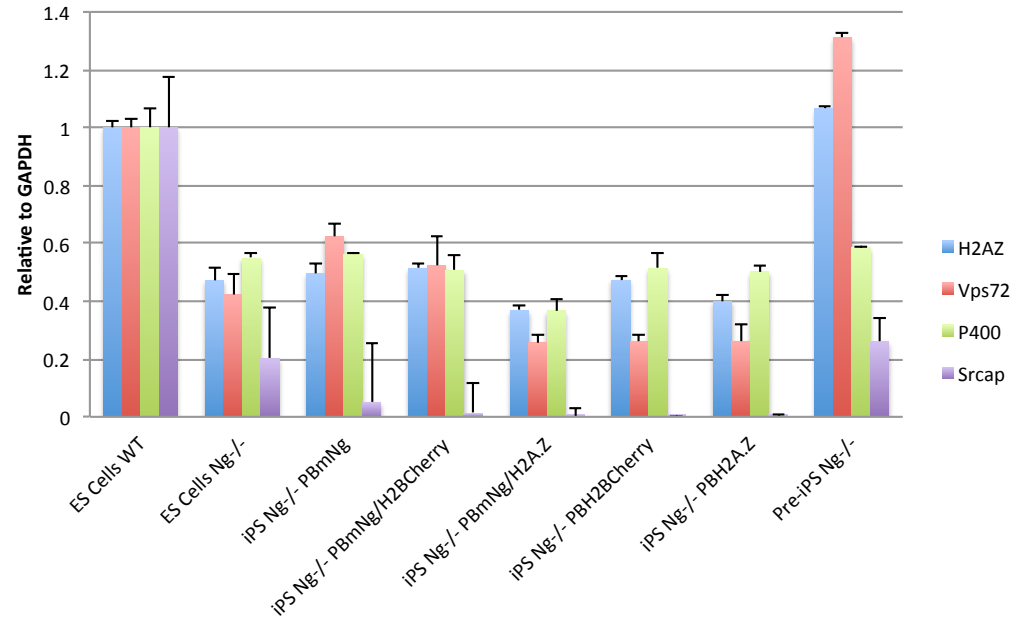
Both core ATPase sub-units of these chromatin remodelling complexes, *P400* and *Srcap*, are expressed in ES, NS and MEF cells (Figure 3.8). However there was a significant difference between the expression of these ATPase sub-units in NS derived iPS cells. *Srcap* was barely detectable in Nanog null iPS cells generated with 4 factors plus a histone *H2A.Z* or *H2B*. Interestingly the expression profile of *H2A.Z*, *Vps72*, *P400* and *Srcap* was very similar between a *Nanog* null ES cells and these *Nanog* null iPS cells. Furthermore *Srcap* levels of expression were also

very low in all NS derived iPS cells, even when Nanog was over-expressed. This data may suggest that the main chromatin remodelling complex involved in the reprogramming of NS cells is P400 and not SRCAP. Indeed P400 and SRCAP may not have redundant roles in pluripotency. The loss of function of P400 in ES cells leads to loss of self-renewal and is not compensated by SRCAP expression (Fazzio et al., 2008). It would be interesting to assess the capacity of reprogramming NS cells to pluripotency after SRCAP depletion.

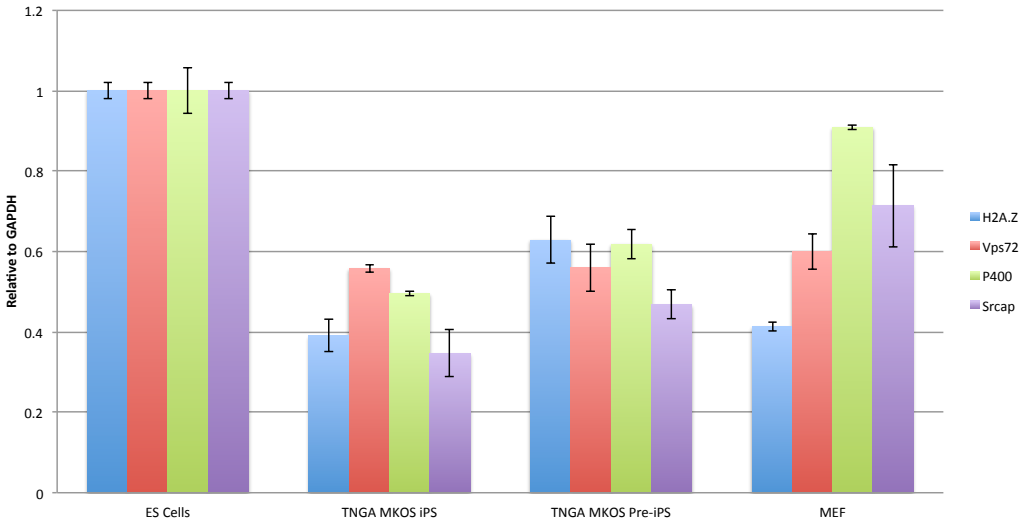
The reprogramming of MEF cells with only 4 factors resulted in less variable expression of *H2A.Z*, *Vps72*, *P400* and *Srcap* when iPS, pre-iPS and MEFS were compared. Importantly this experiment was performed without any histone over-expression. It will be interesting to assess the expression profile in the same stages of MEF reprogramming with 4 factors and *H2A.Z* over-expression. It might be possible that at high levels of *H2A.Z* the expression of *Srcap* would be down regulated.

One of the proposed models for the mechanism of transcription activation mediated by H2A.Z is via the acetylation of lysine residues at the C-terminal tail of this histone variant (Draker and Cheung, 2009) and its hyper acetylation directly correlate with the levels transcription in ES cells (Hu et al., 2013). Therefore the efficiency of this post-translational modification may be a rate-limiting event in the establishment of stable transcriptional activity of the pluripotency circuitry.

To test this hypothesis it would be possible to design two mutant H2A.Z in which the codons for the acetylatable lysine residues (K4, K7, K11, K13 and K15) (Bo-



(a)



(b)

Fig. 3.8 **Expression of H2A.Z and chromatin remodelling components in iPS cells and its precursors:** qRT-PCR analysis for *H2A.Z* and chromatin remodelling components *Vps72*, *P400* and *Srcap* in (a) *Nanog* null NS derived iPS cells reprogrammed with different combinations with *Nanog*, *H2A.Z* and *H2B:Cherry* and (b) MEF derived iPS cells reprogrammed with standard Yamanaka 4 factors. Wild-type ES cells are used as a reference for the expression levels. Data are relative levels to *Gapdh*. Error bars indicate standard errors from three technical replicates

nenfant et al., 2007; Boyne et al., 2006; Bruce et al., 2005; Dryhurst et al., 2004; Ishibashi et al., 2009; Valdes-Mora et al., 2012) would be replaced to code for glutamine mimicking acetylated residues. Conversely, to test the requirement of H2A.Z acetylation in the reactivation of the pluripotency network the codons for these acetylatable lysine residues would also be replaced for a codon of arginine, which cannot be acetylated. These mutant variants can be co-expressed along with an inducible RNA interference vector to knockdown the endogenous H2A.Z during the process of somatic cell reprogramming.

Other potential approaches could be by the over-expression of the *Kat5* sub-unit, known to be the HAT catalytic subunit of the TIP60/P400 (Cai, 2005; Doyon et al., 2004; Squatrito et al., 2006). Furthermore the recently reported dramatic improvement in the efficiency and kinetics of somatic cell reprogramming to pluripotency by depletion of MBD3 sub-unit of NURD complex histone deacetylase (Rais et al., 2013) may be instrumental in dissecting the role of acH2A.Z in reprogramming. For this purpose one could use an inducible vector to co-express *H2A.Z* (and designed mutants) together with inducible RNA interference for *Mbd3*.

3.3 Conclusion

The results presented in chapter gives evidence for a role of the histone variant H2A.Z in the increase of reprogramming efficiency of two types of somatic cells to pluripotency. Unexpectedly, the over-expression of *H2A.Z* was able to reprogram mutant somatic cells lacking *Nanog* expression. Since *Nanog* has always been deemed an essential gateway to pluripotency (Silva et al., 2009) these findings were very interesting and with high impact potential for the field of pluripotency and reprogramming.

During the progress of this project two back to back publications were reported presenting alternative ways of bypassing *Nanog* requirement during reprogramming. The first study demonstrated that *Nanog* is dispensable for the generation of induced pluripotent stem cells in optimised culture conditions in media containing high levels of ascorbic acid and the resultant *Nanog* null iPS cells are transcriptionally highly similar to wild-type iPSCs and support the generation of teratomas and chimeric mice (Schwarz et al., 2014). The second report identified that a cross-talk between Nanog and Klf4 via Stat3 signalling activation and that an enhanced signalling via the Klf4 axis was sufficient to bypass Nanog requirement in the acquisition of full pluripotency (Stuart et al., 2014). It would be interesting to investigate the connections between Nanog bypass mechanisms and where H2A.Z lies in the order of events.

At a similar time the above mentioned studies were published another relevant study reported that the over-expression of *TH2A* and *TH2B*, two histone vari-

ants highly expressed in oocytes, could increase the efficiency of somatic cell reprogramming to pluripotency when co-expressed with the four conventional Yamanaka factors, *Oct4, Sox2, Klf4* and *Myc* (Shinagawa et al., 2014). Interestingly the gene for the histone H2A used in the current project *Hist1h2aa* which its other name is *TH2A* and it resulted in a very slight increase in iPSC efficiency in my experiments. This demonstrated the involvement of maternal histone variants in the reprogramming of the somatic nucleus. Finally another report demonstrated that the reprogramming of somatic cells to induced pluripotency is dependent on *H2A.Z* and its over-expression can increase reprogramming efficiency (Wang et al., 2015). This study further proposed that the pluripotency activity of Nanog requires biochemical stabilisation of the histone variant H2A.Z.

Despite promising initial results the number of new reports demonstrating similar finding to what I have been working on and the limited resources available, the viability of my project became compromised. Therefore it was decided to invest my efforts in alternative ways of studying the reprogramming of somatic cells to pluripotency and these approaches will be described in the next chapters.

Chapter 4

Gene Networks of Somatic Cell Reprogramming towards Pluripotency

4.1 Introduction

During normal development the progression towards pluripotency presumably must happen in a deterministic fashion for the success of pre-implantation development. On the other hand, somatic cell reprogramming towards pluripotency is much less efficient, slow and goes through a series of stochastic events which constitute a bottle neck for the success of complete reprogramming.

However, improvement in somatic cell reprogramming should be possible because a somatic nucleus transferred into an oocyte can be fully reprogrammed,

re-engaging in deterministic normal embryonic development. Therefore the identification of genetic networks overlapping natural development and artificial reprogramming processes could potentially lead to more efficient recipes for somatic cell reprogramming towards pluripotency. This rational assumption forms the basis of the investigation described in this and the next chapters. (Figure 4.1)

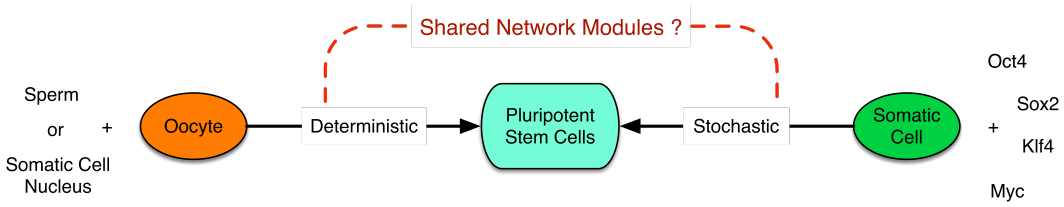


Fig. 4.1 Gene network analysis rational.

The aim of this study was to investigate the gene expression networks involved in somatic cell reprogramming and in normal embryo development, trying to identify common features and possible bottlenecks for reprogramming leading to a selection of new gene candidates which could facilitate induced pluripotency. The wealth of publicly available high quality transcriptome data from controlled time course somatic cell reprogramming studies and normal pre-implantation embryo development provide us with an unprecedented opportunity to perform a systematic transcription network analysis which could not only identify limiting reprogramming factors but also help us better understand mammalian development.

4.1.1 Weighted Gene Co-expression Network Analysis

The strategy chosen for the dissection of the genetic networks operating during somatic cell reprogramming and mammalian pre-implantation development is based on weighted gene co-expression network analysis (WGCNA) (Zhang and Horvath, 2005).

A key feature of such analysis is that it focuses on network modules (clusters) of genes instead of individual genes as is traditionaly the case. It provides a statistical advantage by alleviating the multiple testing problem. Figure 4.2 illustrates a general overview of the WGCNA pipeline used in this study as described below:

1. **Construction of a network:** The pairwise gene co-expression correlation coefficients are transformed into adjacency measures by raising their values to a selected power value (soft thresholding approach) designed to preserve the continuity of the data and to satisfy a scale free topology. The adjacency values are then used for the calculation of a topological overlap matrix (TOM) which represent a weighted gene co-expression network.

2. **Identification of Modules:** The network modules are defined by using an average linkage hierarchical clustering coupled with the topological dissimilarity measures derived from topological overlap matrix. An important approach used in WGCNA is the use of a hybrid dendrogram cutting algorithm combined with partitioning around medoids (dynamic tree cutting). This is a robust method that retains the advantages of both methods. The

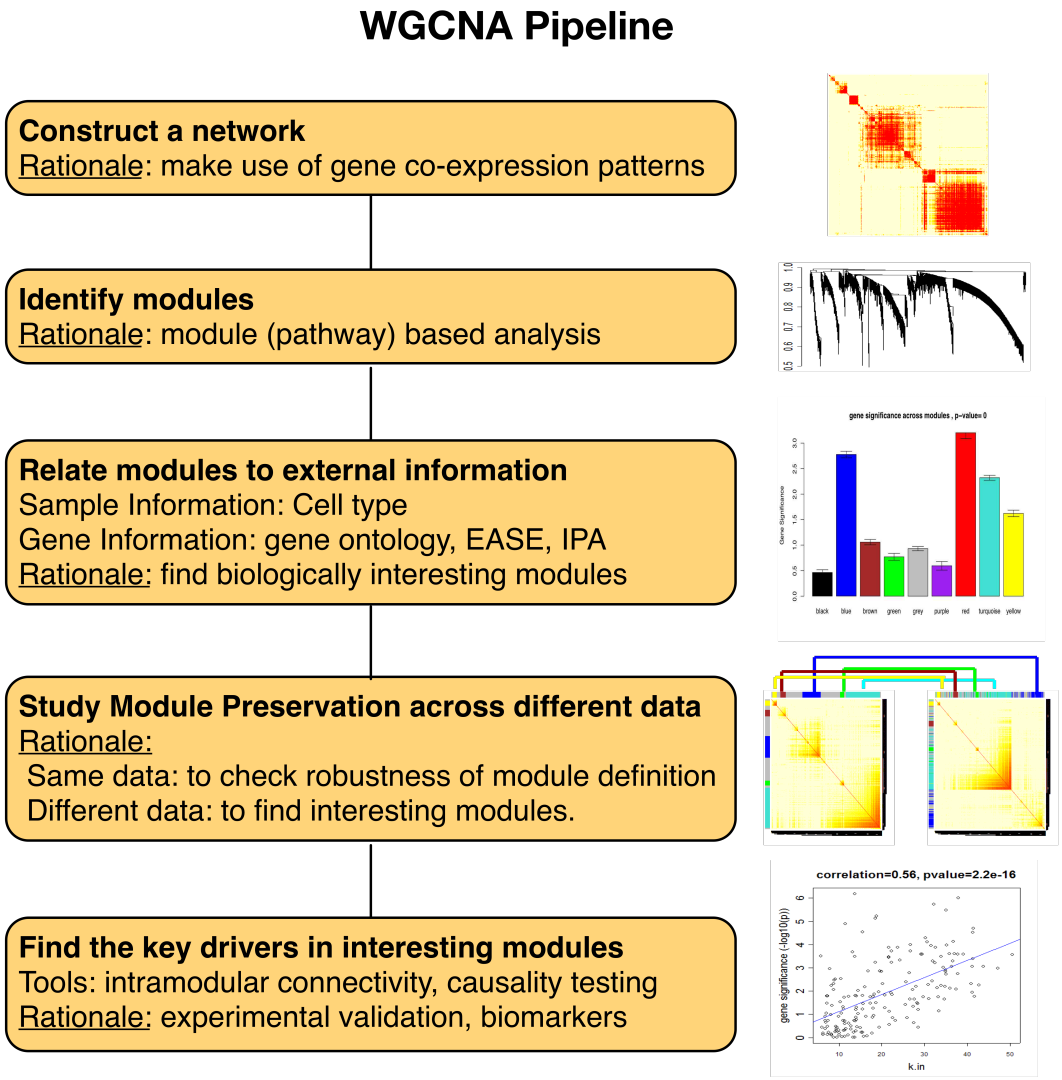


Fig. 4.2 WGCNA workflow. Adatpted from a WGCNA lecture by Steve Horvath at Network Analysis Workshop 2013 (https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/WORKSHOP/2013/

identified modules are then decomposed into eigengenes which are dimension reduced representations of the modules.

3. **Relate Modules to External Information:** The module eigengenes from cell reprogramming allows us to measure gene module memberships and then measure their correlation with the different samples types (different cell types) allowing us to efficiently identify key network modules differing between samples. These modules are then annotated based on databases for gene ontology and pathway analysis. Moreover the correlation with a known biological trait gives rise to a gene significance measure which is a good parameter for the prioritisation of candidate genes.

4. **Module Preservation Network Analysis:** The identified network modules are then used for module preservation analysis to test the robustness of defined modules with independent data sets. This permits us to identify network features that are conserved between different systems as well as the features that are unique to each data set.

5. **Find Key Drivers in Interesting Modules:** The genes with the highest module membership and highest gene significance for relevant cellular traits for each module are then used further for the identification of key driver genes. With these measures we identify the intra-modular hub genes which have a high probability of being the most important genes of a given module network.

4.2 Results and Discussion

For the first part of the analysis I collected published time series gene expression data sets from iPS cell generation and used the algorithms from the WGCNA package to construct weighted gene co-expression networks present in the reversal of somatic fate towards pluripotency (Figure 4.3).

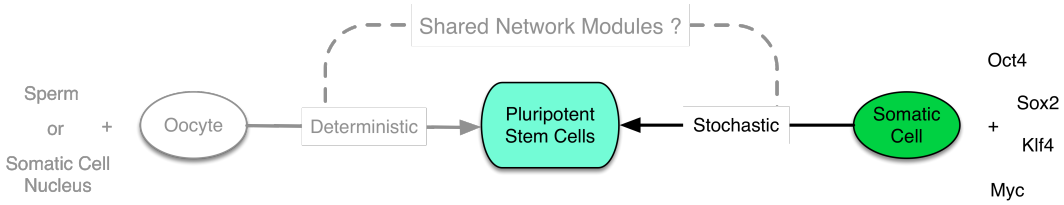


Fig. 4.3 Gene network analysis of iPS cell generation.

4.2.1 Exploratory Data Analysis

The data set analysed in this chapter (Chantzoura et al., 2015) was produced using a mouse secondary reprogramming system for the analysis of the sequential stages of reprogramming. A secondary reprogramming system consists of somatic cells extracted from an animal which was generated from primary iPSCs produced with inducible transgene vectors. This system permits the elimination of variation of integration sites of the reprogramming vector in the genome making the second round of reprogramming more homogeneous and more efficient. The aim of the original study was to compare the reprogramming routes driven by two commonly used polycistronic cassettes constructs (MKOS and OKMS) which expresses the four Yamanaka factors separated by 2A peptides but in different

order (Chantzoura et al., 2015). This two polycistronic cassettes differ mainly in their relative expression levels of Klf4 which is a result of *Klf4* being full sequence length in MKOS and in OKMS is missing the coding sequence for the first 9 amino terminal amino acids (Kim et al., 2015). This study provided RNA-seq gene expression data from sorted cells at different stages during reprogramming experiments using either MKOS or OKMS reprogramming cassettes (Chantzoura et al., 2015). Table 4.1 lists the utilised samples, cell type information, marker profiles used for cell sorting and reprogramming cassette utilised.

The authors have previously characterised intermediate cell stages based on the expression profile of a number of markers (O'Malley et al., 2013a). The reprogramming intermediate cell populations (2NG- are CD44 -ve, ICAM1 -ve, Nanog-GFP -ve and Ecadherin -ve or +ve, 3NG- are CD44 -ve, ICAM1 +ve, Nanog-GFP -ve and Ecadherin -ve or +ve, 3NG+ are CD44 -ve, ICAM1 +ve, Nanog-GFP +ve and E-cadherin +ve) were sorted at day 10 of reprogramming (Table 4.1). The data set also contains other cells which were used as references for somatic and pluripotent cells of different types. MKOS/OKMS iPSCs are established iPSC clones. TNG is an embryonic stem cell line carrying a Nanog-GFP reporter published in (Chambers et al., 2003). From this line TNG MKOS and OKMS embryonic stem cells were generated after targeting the *Sp3* locus with the MKOS or the OKMS cassette respectively. E14 is a reference embryonic stem cell line and MEF are mouse embryonic fibroblasts either wild type or generated from TNG MKOS or OKMS ESCs derived embryos. D6 is the D6s4B5 iPSC line published in (O'Malley et al., 2013a).

Table 4.1 Chantzoura *et al.* RNA-sequencing data set

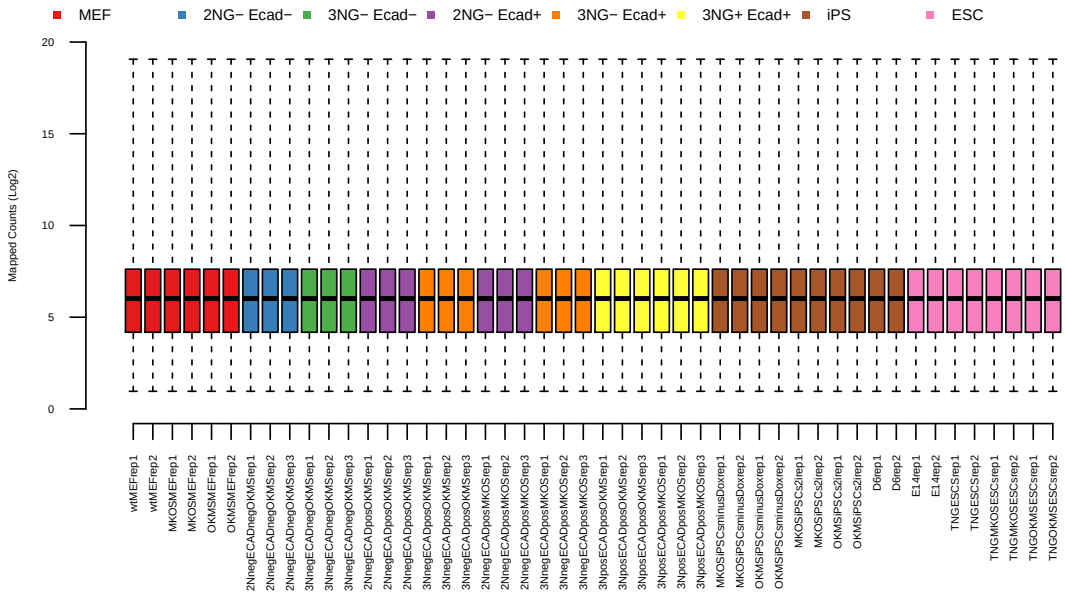
Sample Name	Cell Type	Markers	Cassette
wtMEFrep1	MEF	Nanog-GFP -ve Ecadherin -ve	NONE
wtMEFrep2	MEF	Nanog-GFP -ve Ecadherin -ve	NONE
MKOSMEFrep1	MEF	Nanog-GFP -ve Ecadherin -ve	MKOS
MKOSMEFrep2	MEF	Nanog-GFP -ve Ecadherin -ve	MKOS
OKMSMEFrep1	MEF	Nanog-GFP -ve Ecadherin -ve	OKMS
OKMSMEFrep2	MEF	Nanog-GFP -ve Ecadherin -ve	OKMS
2NnegECADnegOKMSrep1	2NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
2NnegECADnegOKMSrep2	2NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
2NnegECADnegOKMSrep3	2NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
3NnegECADnegOKMSrep1	3NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
3NnegECADnegOKMSrep2	3NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
3NnegECADnegOKMSrep3	3NG- Ecad-	Nanog-GFP -ve Ecadherin -ve	OKMS
2NnegECADposOKMSrep1	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
2NnegECADposOKMSrep2	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
2NnegECADposOKMSrep3	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
3NnegECADposOKMSrep1	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
3NnegECADposOKMSrep2	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
3NnegECADposOKMSrep3	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	OKMS
2NnegECADposMKOSrep1	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
2NnegECADposMKOSrep2	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
2NnegECADposMKOSrep3	2NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
3NnegECADposMKOSrep1	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
3NnegECADposMKOSrep2	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
3NnegECADposMKOSrep3	3NG- Ecad+	Nanog-GFP +ve Ecadherin -ve	MKOS
3NposECADposOKMSrep1	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	OKMS
3NposECADposOKMSrep2	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	OKMS
3NposECADposOKMSrep3	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	OKMS
3NposECADposMKOSrep1	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	MKOS
3NposECADposMKOSrep2	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	MKOS
3NposECADposMKOSrep3	3NG+ Ecad+	Nanog-GFP +ve Ecadherin +ve	MKOS
MKOSiPSCsminusDoxrep1	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
MKOSiPSCsminusDoxrep2	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
OKMSiPSCsminusDoxrep1	iPS	Nanog-GFP +ve Ecadherin +ve	OKMS
OKMSiPSCsminusDoxrep2	iPS	Nanog-GFP +ve Ecadherin +ve	OKMS
MKOSiPSCs2irep1	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
MKOSiPSCs2irep2	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
OKMSiPSCs2irep1	iPS	Nanog-GFP +ve Ecadherin +ve	OKMS
OKMSiPSCs2irep2	iPS	Nanog-GFP +ve Ecadherin +ve	OKMS
D6rep1	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
D6rep2	iPS	Nanog-GFP +ve Ecadherin +ve	MKOS
E14rep1	ESC	Nanog-GFP +ve Ecadherin +ve	NONE
E14rep2	ESC	Nanog-GFP +ve Ecadherin +ve	NONE
TNGESCSrep1	ESC	Nanog-GFP +ve Ecadherin +ve	NONE
TNGESCSrep2	ESC	Nanog-GFP +ve Ecadherin +ve	NONE
TNGMKOSESCsrep1	ESC	Nanog-GFP +ve Ecadherin +ve	MKOS
TNGMKOSESCsrep2	ESC	Nanog-GFP +ve Ecadherin +ve	MKOS
TNGOKMSESCsrep1	ESC	Nanog-GFP +ve Ecadherin +ve	OKMS
TNGOKMSESCsrep2	ESC	Nanog-GFP +ve Ecadherin +ve	OKMS

The RNA-Seq data set for Chantzoura et al. (2015) was downloaded from the Gene Expression Omnibus (GEO) repository from the provided accession number GSE70816. The raw counts data can be found as supplementary data within this submission by the file labelled as "GSE70816 Chantzoura2015 Expression ReadCounts.txt".

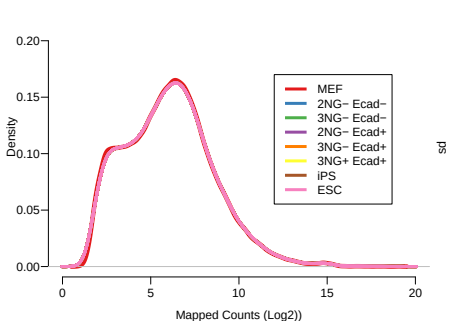
Data Normalisation

The data set comprises 48 samples and 25922 features/genes. Upon initial analysis of the data followed by tests with different normalisation approaches the samples were normalized by regularised log transformation using the DESeq2 package (Love et al., 2014) and removed the genes with counts of less than 10 in 20% of the samples resulting in 13995 genes.

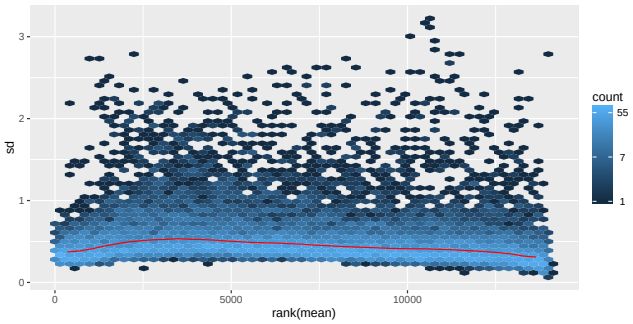
We can see in Figures 4.4a and 4.4b that the global read distribution for all the samples after normalisation is fairly equivalent indicating high quality of all samples. We can observe in the density plot (Figure 4.4b) that all the samples have a high proportion of undetected and or low expressed genes, a commonly expected feature of RNA-Seq data. One of the properties of the regularised log transformation is the shrinkage of the variance of low expressed genes resulting in a more even variability across the ranges of level of expression as we can observe in Figure 4.4c. This reduction of expression noise improves the quality of downstream analysis such as clustering and dimension reduction techniques such as principal component analysis.



(a) Box plot



(b) Density plot



(c) Gene expression standard deviation

Fig. 4.4 **Chantzoura data set mapped reads statistics.** (a) Box plot: Distribution of the RNA-seq data sets representing the expression levels (regularised log transformed and quantile normalized) for all 13995 genes considered. The sample groups are color coded as displayed on the legend above the plot; (b) Density plot: Log-intensity distribution of all libraries superposed and color coded as per legend insert; (c) Gene expression standard deviation vs. ranked mean: The color intensity of scatter plot reflects the number of genes at a given mean rank as per legend. The red line depicts the running median estimator (window-width 10%).

Unsupervised Hierarchical Clustering

The unsupervised hierarchical clustering of the samples by their Euclidean distances further indicate the high quality of the data set (Figure 4.5). We can observe that most cell types form well defined distinct clusters apart from the initial inter-

mediary pre-iPS cells. The intermediate later stage iPS cells (3NG+ Ecad+) cluster well with established iPS cells and the earlier stage cells form two groups that cluster well based on the expression of the surface marker E-Cadherin. Interestingly, the E-Cadherin +ve intermediate pre-iPS cells 2NG- Ecad+ and 3NG- Ecad+ cluster perfectly into two groups distinguished by the reprogramming cassette used (MKOS and OKMS) instead of the expression of the surface marker ICAM1 which was one of the markers used to sort these two intermediary reprogramming cell populations. The fact that the different reprogramming cassettes have the property of producing distinctly related reprogramming intermediate cells might confer this analysis an advantage since we have an opportunity to characterise the genetic network of more intermediate stages, possibly producing an even higher resolution analysis than we would expect if this phenomenon was not observed.

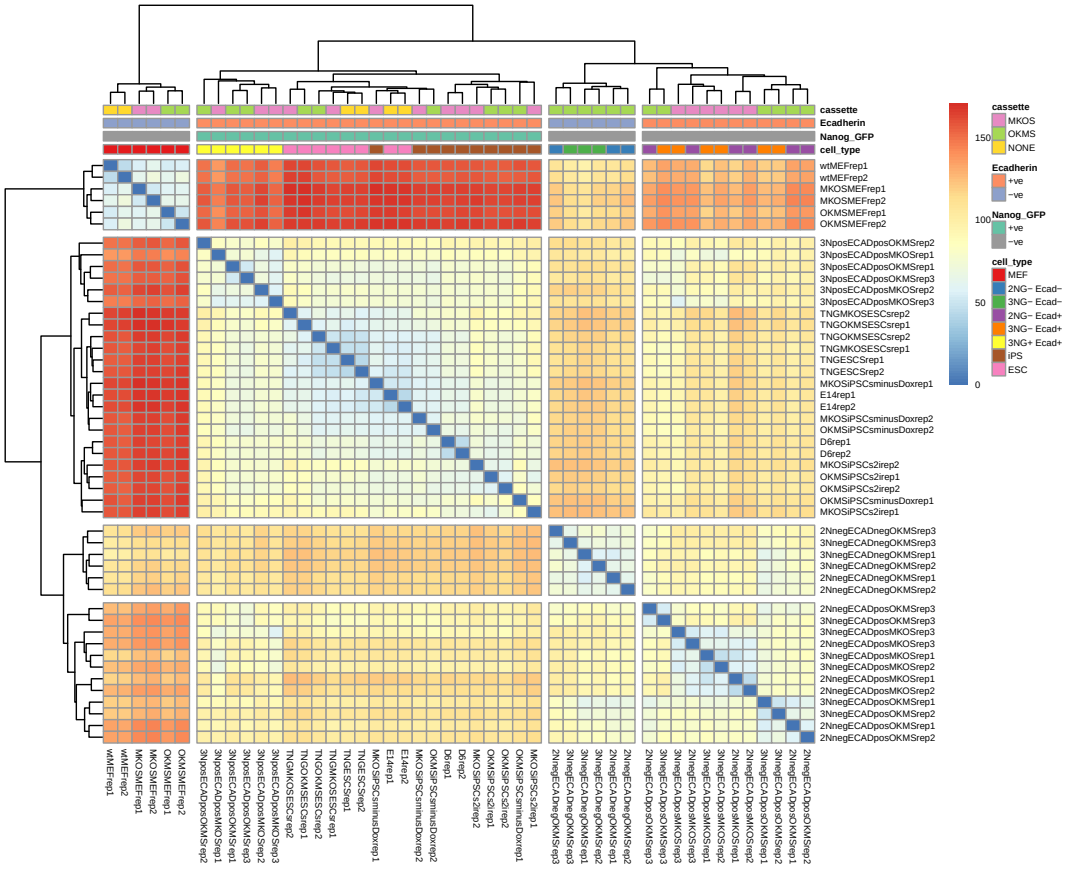


Fig. 4.5 **Chantzoura data set sample clustering:** The heatmap represents the clustered pair-wise Euclidean distances between all samples based on their regularised log transformed expression of all all 13995 genes analysed. The euclidean distances range from 0 (dark blue) to just over 150 (red) as depicted in the legends next to the plot. The surface marker E-Cadherin, the reporter gene Nanog-GFP, the type of reprogramming cassette used (MKOS, OKMS or none) as well as the sample groups MEF, 2NG- Ecad-, 2NG- Ecad+, 3NG- Ecad-, 3NG- Ecad+, 3NG+ Ecad+, iPSCs and ESCs were also color coded to assist in the sample clustering visualisation.

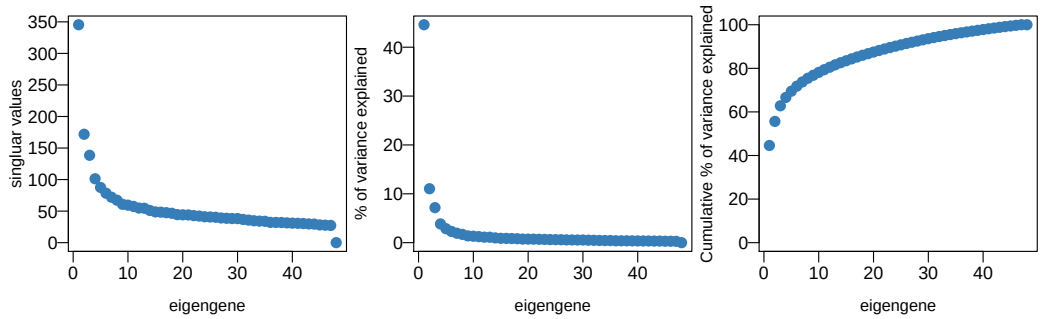
Principal Component Analysis

This high-throughput gene expression data set was further processed in order to reduce it's dimensions by singular value decomposition transformation and was used for performing principal component analysis. I was able to verify that the variance observed can be well represented by the first few principal components. The first principal component has by far the highest singular value over all (Figure 4.6a) and explains over 41% of the variance between samples (Figure 4.6b). Furthermore the combination of the first 3 principal components is able to explain around 60% of the variance (Figure 4.6a, 4.6b and 4.6c).

This analysis gives us another perspective of the relatedness of the samples observed with the hierarchical clustering and further reinforces the suitability of the sorted intermediate reprogrammed cells for the analysis of sequential steps in reprogramming towards pluripotency. The principal component analysis plots shows a remarkable clear route from somatic cell state (MEFs) moving towards the pluripotent cell state (ES and iPS cells) . The order of the cell stages observed are MEFs, 2NG- Ecad-, 3NG- Ecad-, 2NG- Ecad+, 3NG- Ecad+, 3NG+, Ecad+, iPSCs and ESCs (Figures 4.7a and 4.7c).

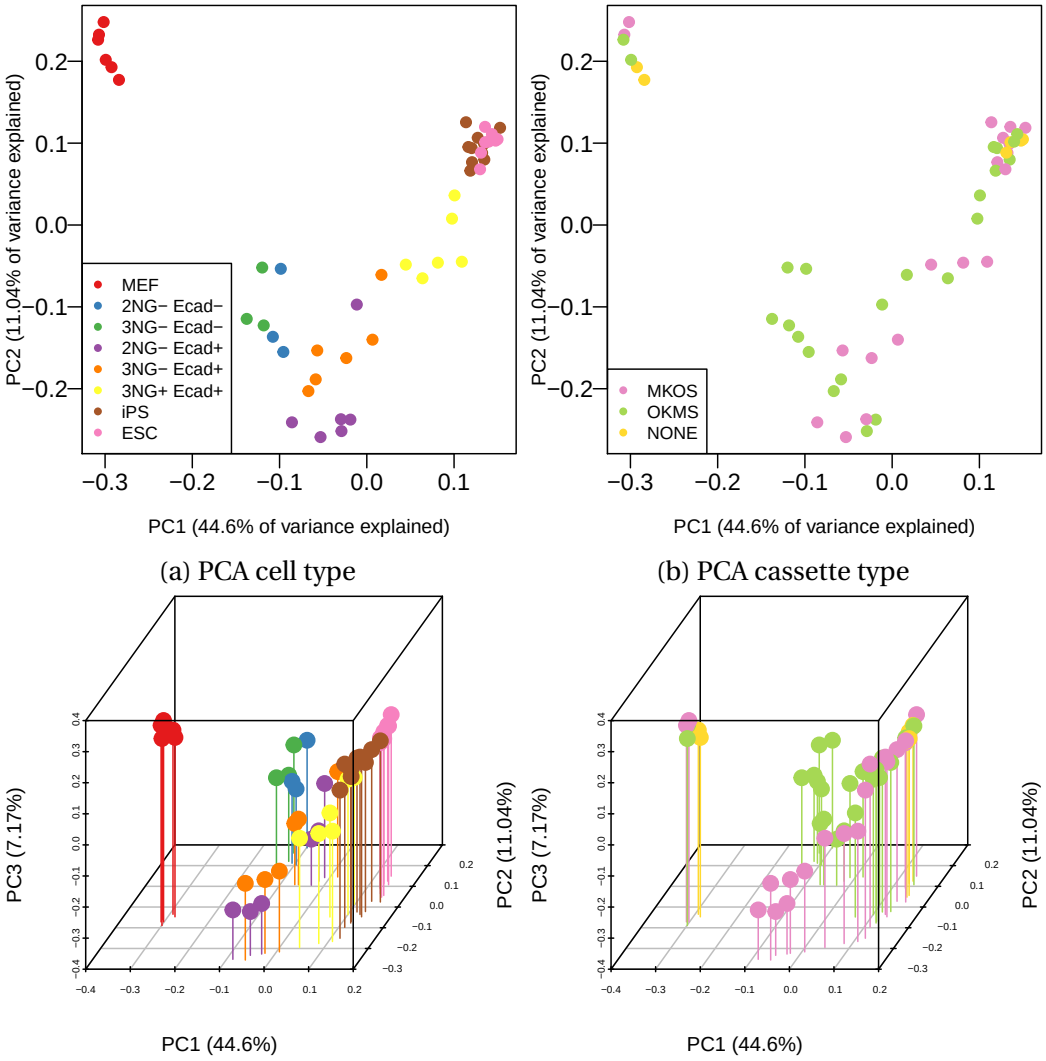
Moreover, when we analyse the sequential stages of reprogramming focusing on the type of cassette construct utilised we can confirm the findings from Chantzoura et al. (2015) that the earlier intermediate stages 2NG- Ecad+ and 3NG- Ecad+ produced by the OKMS cassette seem to be trapped in an earlier stage when compared to the equivalent cell populations produced with the MKOS cassette (Figure

4.7b and 4.7d). This observation further highlights the added resolution of this dataset.



(a) Singular values (b) Variance explained (%) (c) Cumulative variance

Fig. 4.6 **Chantzoura data set singular value decomposition**



(a) PCA cell type

(b) PCA cassette type

(c) 3D PCA cell type

(d) 3D PCA cassette type

Fig. 4.7 **Chantzoura data set principal component analysis:** Principal components (eigengenes) were calculated via singular value decomposition and plotted the first two (a and b) or three principal components (c and d) to assist in visualisation of the sample groupings. The plots were both coloured based on the sample groups and reprogramming cassette used as indicated in the figure legends.

4.2.2 Weighted Gene Co-Expression Network Analysis

One of the key features of weighted gene co-expression analysis (WGCNA) is the preservation of continuous data in the classification of network module membership (Zhang and Horvath, 2005). Traditional methods employ a hard threshold for such classification which leads to discrete outputs. The choice for a hard threshold is arbitrary and incur a loss of information especially in the boundaries of the chosen threshold.

WGCNA uses a soft-threshold in the form of a power over the adjacency measurements which itself is derived from the correlation of measurements. This strategy leads to a "fuzzy" classification value which lies continuously between 0 and 1.

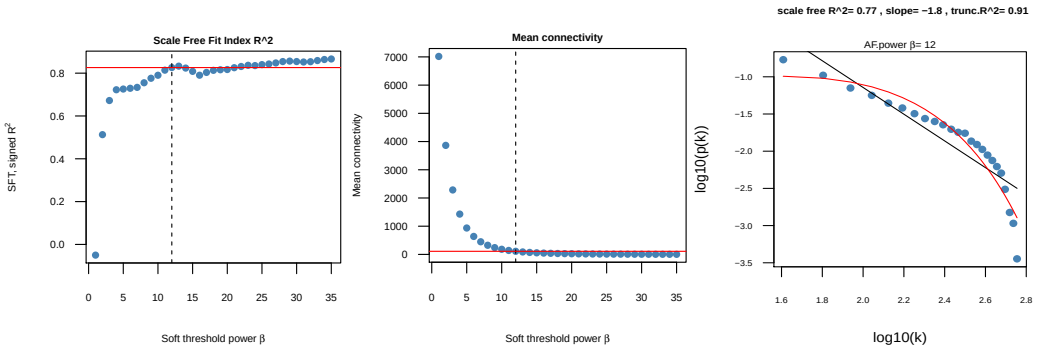
Scale Free Topology Fitting

Most biological networks display a power-law distribution (scale-free topology), which is characterised by having a small number of highly connected nodes (i.e. hub genes) that link the rest of the less connected nodes to the system (Barabasi and Albert, 1999). Using the WGCNA package (Langfelder and Horvath, 2008) implemented in R I have performed tests for selecting the soft threshold power to try satisfy a scale free topology and at the same time produce the most meaningful biological network modules.

Two measures of agreement of scale free topology are the scale free fit index R^2 and the overall mean connectivity (κ). When R^2 is close to 1 and the overall mean κ is

relatively low the network is considered to be approximately scale free. Therefore I selected the value 12 for the soft threshold as it is the point where R^2 gets close to saturation and the mean connectivity is close to the minimum (Figure 4.8a and 4.8b).

Another diagnostic of network scale free topology is done by plotting log(p(κ)) versus log(κ). We can see in Figure 4.8c that the chosen power (β = 12) leads to an approximate scale free topology with an R^2 of 0.77 and slope -1.8. The black line shows the linear model fit for scale-free topology. We can observe that highest κ values cause the network to satisfy the scale free topology only approximately and might be better modelled by the exponentially truncated power law as fitted by the red line in Figure 4.8c (Csányi and Szendrői, 2004).



(a) Soft threshold selection (b) Mean connectivity (c) Scale free fitting

Fig. 4.8 **Selection of power β for the Chantzoura data set:** A power β is selected at the point where the scale free fit index R^2 is close to saturation (a) and the mean connectivity is close to the minimum (b). The power β of 12 was chosen as the soft threshold value which causes the network topology to approximate a scale free topology the most (c).

Construction of a Weighted Gene Co-Expression Network

I constructed a signed weighted gene co-expression network following the guidelines from the authors of the method (Zhang and Horvath, 2005). In order to prioritise only positive gene expression correlations and to obtain a network robust to outliers, the adjacency matrices were constructed using an algorithm that takes into account the correlation coefficient sign derived from bi-weight mid-correlations. I initially created a matrix of pairwise correlations (Pearson correlation) between all pairs of genes across the measured samples ($cor(x_i, x_j)$) followed by the creation of an adjacency matrix ($A_{i,j}$) by raising the co-expression coefficients to the selected soft threshold power $\beta = 12$ ($A_{i,j} = |0.5 + 0.5xcor(x_i, x_j)|^{12}$). The adjacency matrix ($A_{i,j}$) was then used for the calculation of the topological overlap matrix (TOM), which is a robust and biologically meaningful measure of network interconnectedness. Network interconnectedness represents the strength of the co-expression relationship of two genes with respect to all other genes in the network. The grouping of genes with highly similar co-expression relationship was done by average linkage hierarchical clustering on the topological overlap. The hierarchical clustering tree was cut using the Dynamic Hybrid Tree Cut algorithm (Langfelder et al., 2007). The expression profile of each module was summarised by representing it as the first principal component termed here as module eigengene.

Figure 4.9 displays the resulting hierarchical cluster tree of all 13995 expressed genes in the dataset and the color bands below is a visual representation of the

module assignments (first coloured band) and the gene significance (GS) measure (red and blue indicates positive and negative correlation with analysed cell types respectively). The strongest clustering driver was the reprogramming status and the two biggest modules, turquoise and blue, correspond to mouse embryonic fibroblasts (MEF) cells and cells in pluripotent state respectively. The black, green, red, brown and pink modules correlate with the intermediate reprogramming cell sub-populations 2NG- Ecad+ and 3NG- Ecad+, while the magenta and yellow modules correlate with 2NG- Ecad- and 3NG- Ecad- which are E-Cadherin -ve. These intermediate modules may be interesting for further investigation since they may reveal important reprogramming mechanistic landmarks towards successful reprogramming to pluripotency.

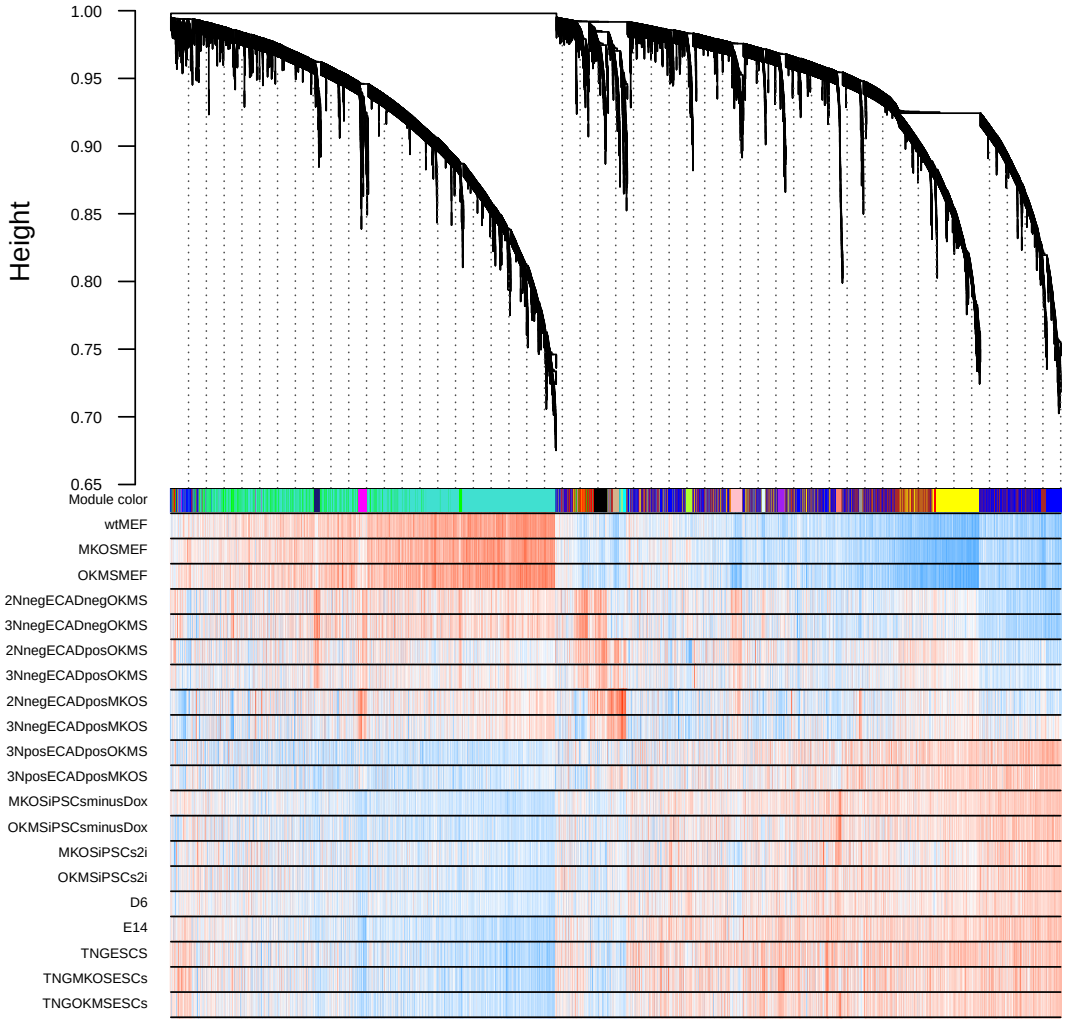


Fig. 4.9 **Network module definition for the Chantzoura data set.** Gene dendrogram obtained from the reprogramming dataset of Chantzoura et al. 2015. This tree was defined by hierarchical clustering the dissimilarity based on consensus topological overlap measures using average linkage of the 13995 genes analysed. The color coded modules were assigned based on the dynamic hybrid branch cutting method represented as the first color band below the the dendrogram. The several other color bands represent the gene significance measures to all the cell samples analysed: “red” indicates a high positive correlation with the a cell type and “blue”, high negative correlation. Note that the two major network modules “turquoise” and “blue” strongly correlate with MEFs and Nanog +ve cells respectively and the other modules are associated with the intermediate stages of reprogramming.

Merging Highly Similar Network Modules

I could verify that some defined module eigengenes have very similar expression profiles and those ones with the most similarity should be merged together. The module eigengenes were used for performing a hierarchical clustering and an arbitrary height was selected at a point were module eigengenes that fall below should be merged as illustrated in Figure 4.10

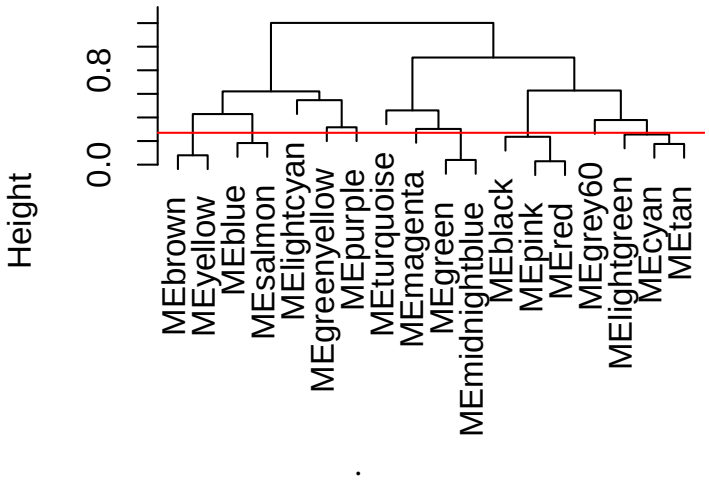


Fig. 4.10 **Module merging height for the Chantzoura data set.** Hierarchical clustering of module eigengenes. The red line indicates the arbitrary height of 0.27 selected for merging very similar module eigengenes.

The function *mergeCloseModules* from the WGCNA package was used for merging and reassigning the final module eigengenes. The top panel of Figure 4.11 shows a new hierarchical clustering of the resulting merged eigengenes and the heatmap in the bottom panel illustrate the the pairwise correlations of such modules. The chosen merging threshold height was considered conservative as there were still some module eigengenes which were significantly correlated to each other. The green and magenta modules are associated with early reprogramming stages and the brown and blue modules associated with the latest stages of pluripotency

acquisition. The two modules from these stages, which have a pair wise correlation larger than 0.70, as seen in the heatmap from Figure 4.11, could potentially still be merged together but were left as they were as an attempt to recover the most biological information from these two important stages of reprogramming.

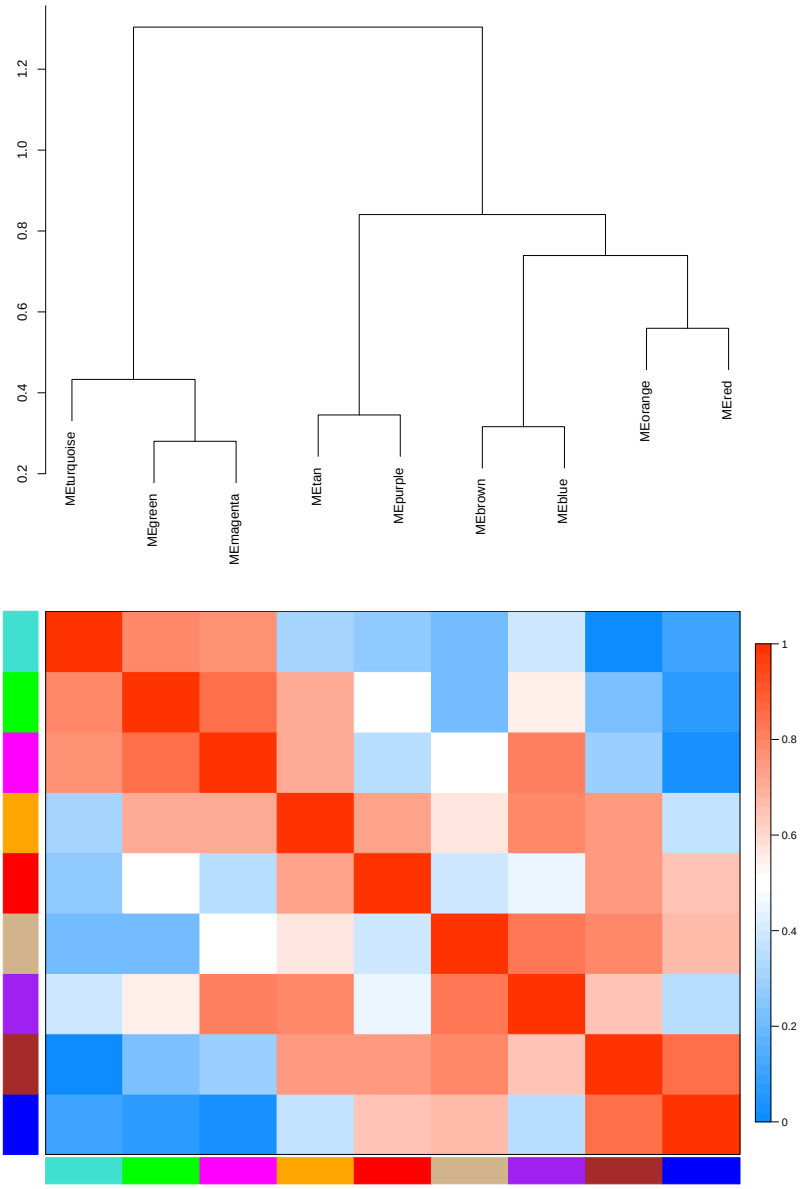


Fig. 4.11 **Module eigengenes network for the Chantzoura data set.** The top panel shows a hierarchical clustering dendrogram of the eigengenes based on dissimilarity. The bottom panel shows the eigengene network, which is defined by a signed weighted correlation network varying from 0 (no correlation) to 1 (highest possible correlation) as displayed in the color legend.

After merging the highly similar module eigengenes we get an improved picture of the representation of cell types (Figure 4.12). The larger modules turquoise, brown and blue remained almost unchanged representing MEFs and pluripotent cell types respectively. The intermediate sub-populations 2NG- Ecad+ and 3NG- Ecad+ are now represented by the tan and purple modules. The green, magenta and orange modules represents the 2N- Ecad- and 3N- Ecad- sub-populations.

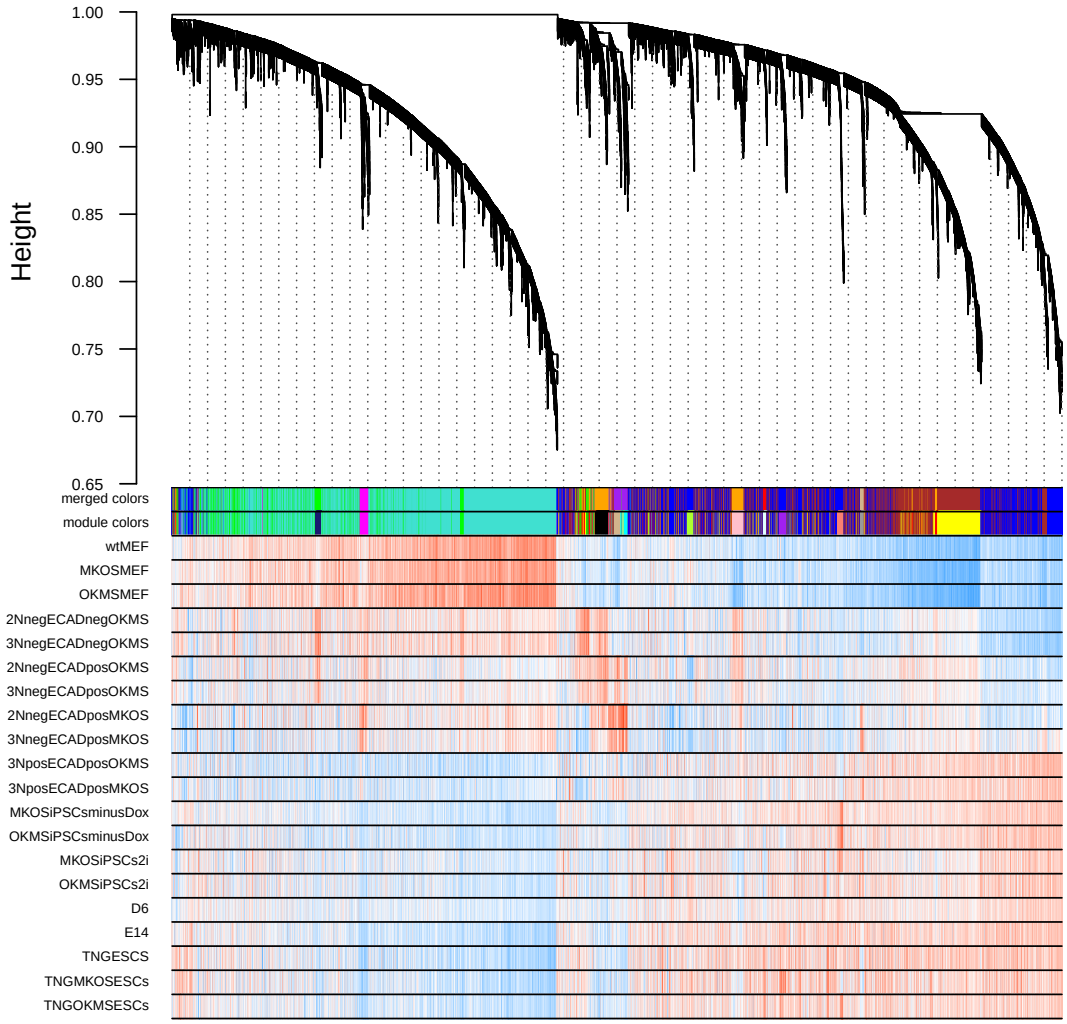


Fig. 4.12 **Network module definition after merges for the Chantzoura data set.** This tree was defined by hierarchical clustering the dissimilarity based on consensus topological overlap measures using average linkage of the 13995 genes analysed as shown before in Figure 4.9 but this time the highly correlated modules were merged. The color coded modules were assigned based on the dynamic hybrid branch cutting method represented as the first color band below the the dendrogram. Note that now there are two module color bands corresponding to the original and the merged module assignments. The several other color bands (blue and red) represent the gene significance measures to all the cell samples analysed: “red” indicates a high positive correlation with the a cell type and “blue”, high negative correlation.

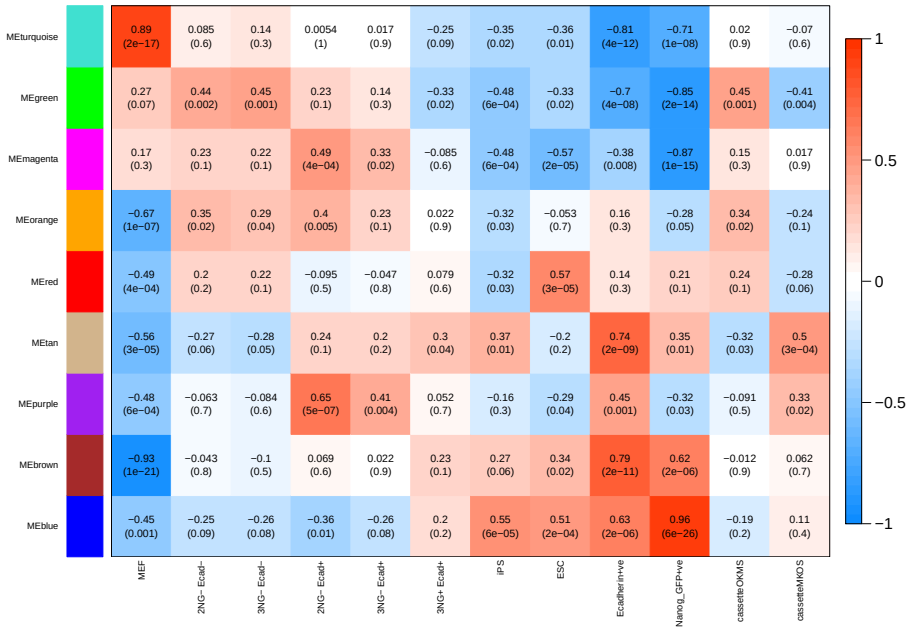
Relating Module Eigengenes to Cellular Traits

The next step was to measure the association significance of the defined module eigengenes and cellular traits. The module eigengenes are optimal representations of the gene expression profiles of each module, therefore their correlation measures with different traits provide a useful measure of association significance (Horvath, 2011) .

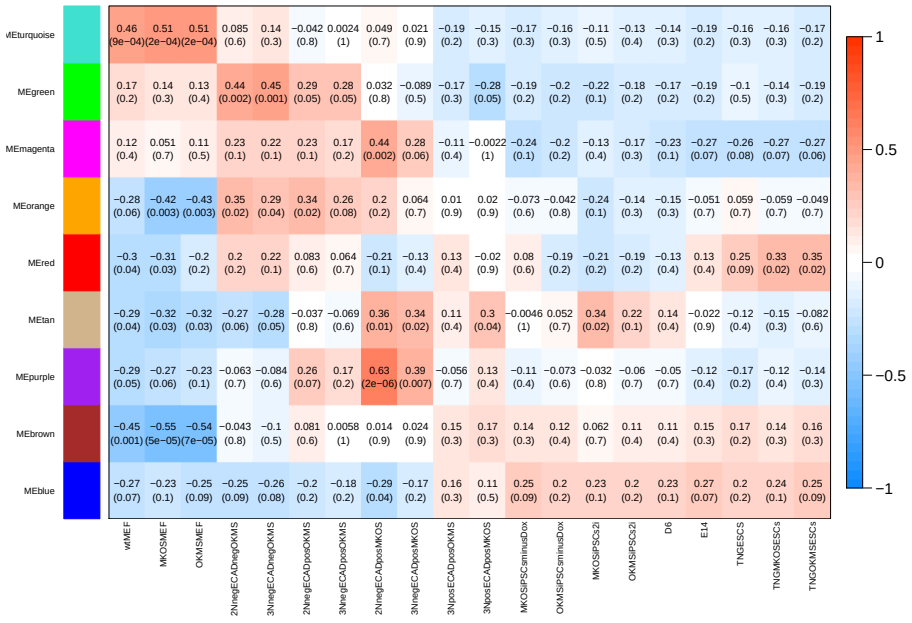
The heatmaps in Figure 4.13a and 4.13b summarises this association analysis. The cell color palette between red, white and blue represent positive neutral and negative correlations respectively. The pairwise correlation coefficients for each module-trait are displayed inside the cells with the significance values (student asymptotic p-values) between brackets. The correlation of the module eigengene and the gene co-expression profile provides a quantitative measure of module membership (MM) which is useful for quantitatively measuring the expression similarity of all transcripts in the dataset to every module.

One of the main purposes for the publication by Chantzoura et al. (2015) was to compare the properties of the two reprogramming cassettes analysed (MKOS and OKMS). It was reported that the MKOS cassette is associated with a significantly faster reprogramming kinetics and that a great proportion of cells treated with OKMS reprogramming cassette appear to get "stuck" at intermediate reprogramming stages. The current reanalysis of Chantzoura et al. (2015) data set by WGCNA further confirms these findings. Figures 4.14a and 4.14b illustrate the measurements of association (average gene correlation) of each of the defined module

eigengenes with the two reprogramming cassettes. We can see that the reprogramming cassette MKOS is positively associated with the blue and brown modules (Late pluripotency genes) while the reprogramming cassette OKMS shows a negative association with these modules. Moreover the reprogramming cassette OKMS is highly associated with modules from the earliest stages of reprogramming green, magenta and orange while MKOS is negatively associated with the green and orange modules and significantly less associated with the magenta module.

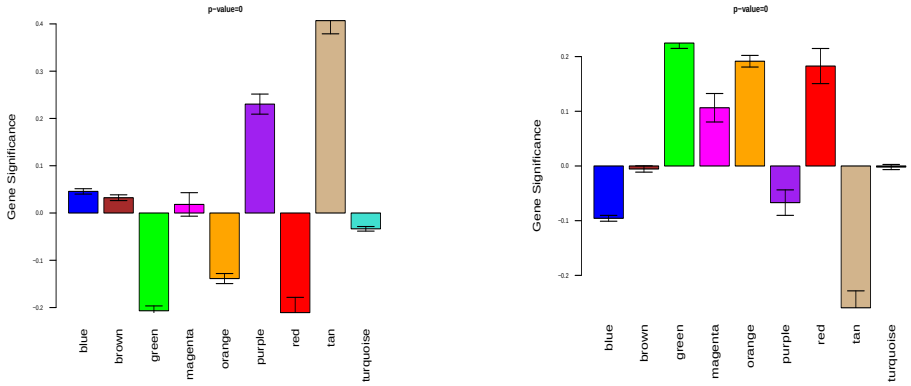


(a) Module eigengene association with cell type



(b) Module eigengene association with sample type

Fig. 4.13 **Chantzoura data set module eigengene associations with phenotypes.** Network module association to cell types and cell markers: The heatmaps in Figures 4.13a and 4.13b represent a high level representation of the identified network modules and the highly correlated modules to relevant trait provide an efficient mean of prioritisation of modules for further analysis. The module eigengene values were correlated with each cell type analysed as well as the markers Ecadherin and Nanog-GFP and the two alternative reprogramming cassettes MKOS and OKMS. The color coded heatmaps represent the level of correlation between the module eigengenes (rows) and cell type or feature (columns). As shown by the color legend guides on the right, red represent strong positive correlation, white stands for no correlation and blue represent strong negative correlation. Each correlation coefficient value is printed in the corresponding table cell along with the association significance level between brackets. Again we can see that the turquoise and blue module are very significantly highly correlated with MEF cells (R=0.84, p = 1e-13) and Nanog-GFP +ve cells (R=0.83, p = 3e-13).



(a) MKOS cassette (b) OKMS cassette

Fig. 4.14 **Module significance to reprogramming cassette type:** alternative visualisation of the level of association (mean gene significance) across modules and the reprogramming cassettes MKOS 4.14a and OKMS 4.14b. The higher the mean gene significance in a module, the more significantly related the module is to the feature of interest. We can see that the two reprogramming cassettes are associated with close to mutually exclusive network modules. The last two columns of Figure 4.13a represent similar information.

4.2.3 Network Modules Based Gene Set Enrichment Analysis

The associations of the various cell types along the reprogramming route with the final network module eigengenes provide an attractive approach for ranking gene sets for performing gene set enrichment analysis. The selection of genes with high module membership values (high association with a module) and high gene significance (high association) for a cell trait/type permits us to select network hub genes which have the highest probabilities of being the essential drivers of biological functions most prominent in the cell traits/types in question.

The Figure panels 4.16a to 4.24 illustrate such an approach for each of the identified network modules for this reprogramming data set. The sub-figure a in each panel is formed by a gene expression heat map (top) and the corresponding module eigengene expression which gives a summary representation of the expression of the genes in that module. The number of genes in each modules is indicated between parentheses on the top of each heatmap. These expression visualisations allow a precise appreciation the dynamics of gene expression profiles through the various stages of reprogramming. The sub-figure b in each panel illustrates the correlation of the module membership from the genes in each module as function of their gene significance for the best correlated cell trait/type. In order to prioritise the most important hub genes a threshold was set for module membership and gene significance values. The genes with the highest values for both measures within such threshold were selected as input for performing gene set enrichment analysis using the online tool Enrichr (Chen et al., 2013a)

and the results are summarised as bar plots of the enrichment levels of the most significantly enriched gene ontology (GO) terms.

Figure 4.15 summarises the results from all Figures 4.16a to 4.24. Based on the WGCNA results the observed gene co-expression network modules during reprogramming fall roughly into three categories with regards to their expression dynamics: a) Somatic related module expression goes gradually down at different speed (Turquoise and Green); b) Intermediate stages related module expression that goes up in the middle of the process but is down regulated when pluripotency is reached (Orange, Magenta, Purple, Red and Brown); c) Pluripotency related modules that goes up in the final stages of reprogramming (Brown and Blue).

The associated biological processes enriched in each network module gives us an insight of the events occurring on the route to pluripotency. The transition from somatic cells MEF to the first intermediate reprogramming cell population 2NG-ECAD- is marked by gradual but pronounced decrease of genes associated with extracellular matrix organisation, and cell migration. Interestingly the 2NG- ECAD- population significantly acquires the expression of genes related to non-coding RNA processing and RNA splicing and this module is gradually down regulated before the acquisition of pluripotency. The red module is also associated with RNA splicing and has transient up-regulation in 2NG- ECAD-, 3NG+ ECAD+ and ESCs, however its expression is down-regulated in iPSCs. The magenta module is enriched for genes associated with the activation of immune response and their expression gets gradually up-regulated from 2NG- ECAD-, peaking its expression

when *E-Cadherin* expression is activated (2NG- ECAD+) but is sharply down-regulated upon pluripotency acquisition marked by Nanog expression (3NG+ ECAD+). The purple module also has a transient expression associated with the activation of *E-Cadherin* expression which marks the mesenchymal to epithelial transition (MET) stage of reprogramming and then is down-regulated upon Nanog expression (3NG+ ECAD+). The Tan module expression is also transiently up-regulated around MET phase and is associated with amino acid bio-synthesis. After MET (ECAD+) the brown module gets up-regulated and its expression is maintained steady until full pluripotency. The brown module is associated with cell division regulation, DNA repair and recombination. After Nanog expression is switched on (3NG+ ECAD+) the cells are considered to have completed the reprogramming to pluripotency and this period is marked by the expression activation of the large blue network module which the most significant hub genes are associated with meiosis and epigenetic regulation.

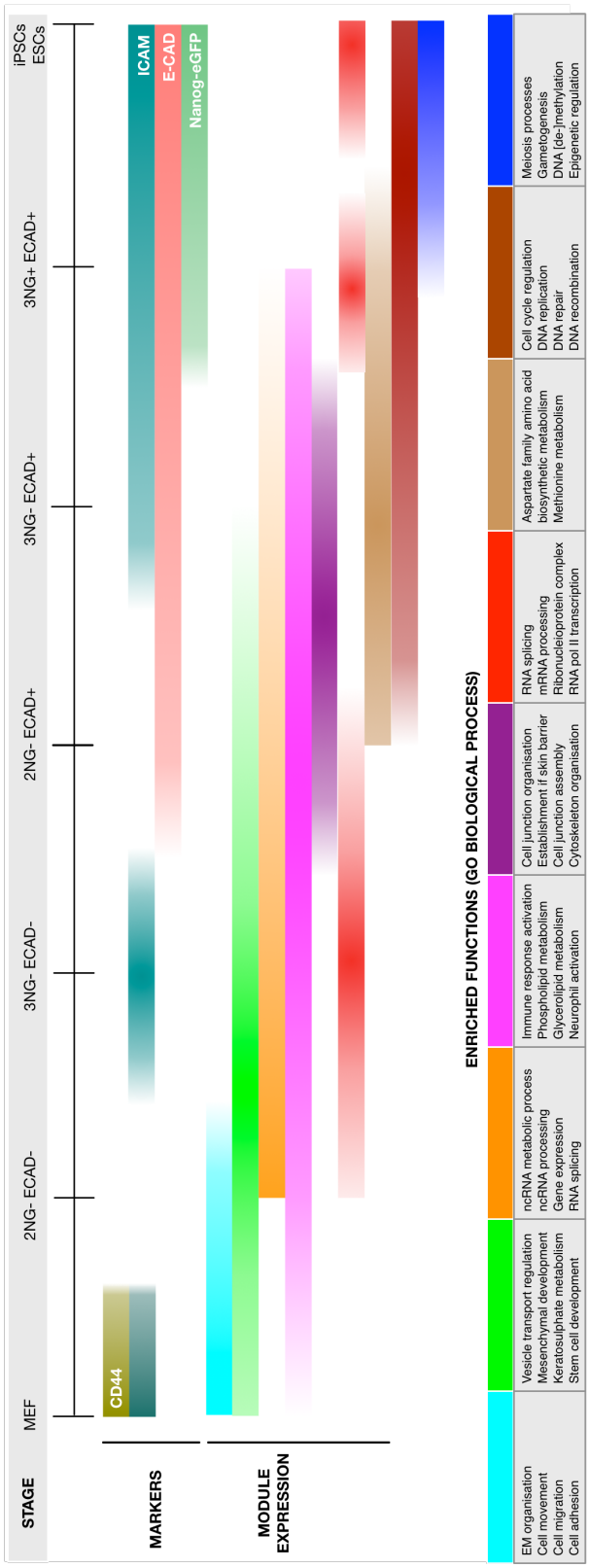


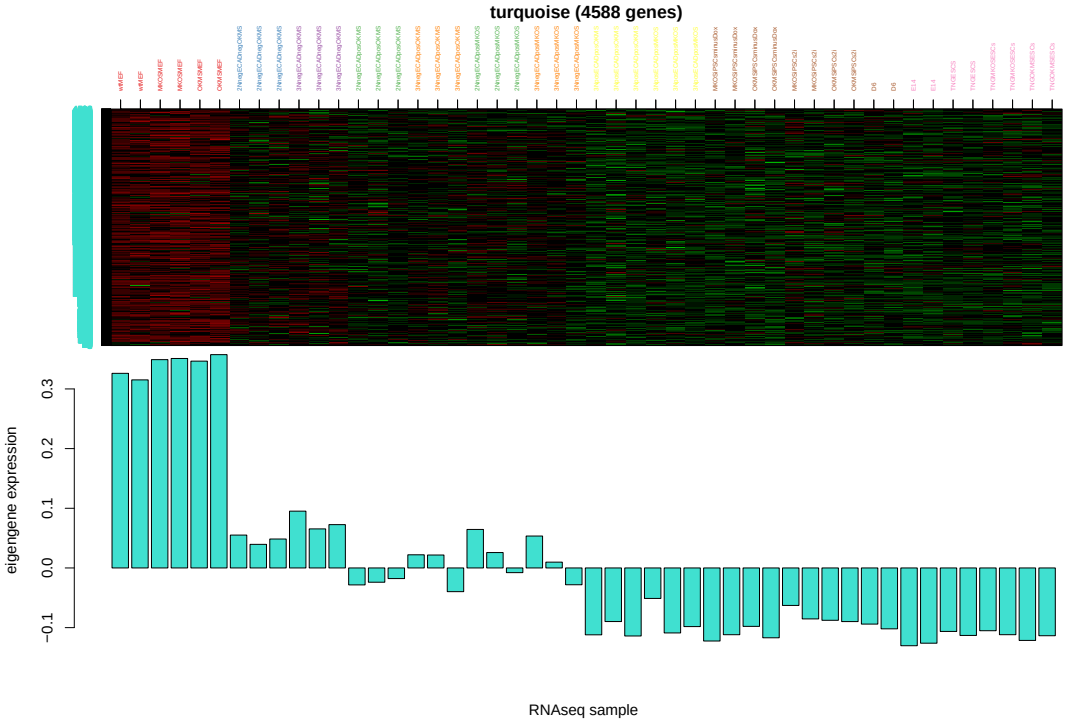
Fig. 4.15 **Dynamics of Biological Processes in iPS cells reprogramming:** The stages of reprogramming are labelled in the grey bar on the top of the figure. The markers used for the cell sorting and the network eigengenes expressions are represented as horizontal color bars with the color intensity correlating with its expression levels. The most significantly enriched functions are summarised on the color coded boxes on the bottom of the figure.

4.2.4 Conclusion

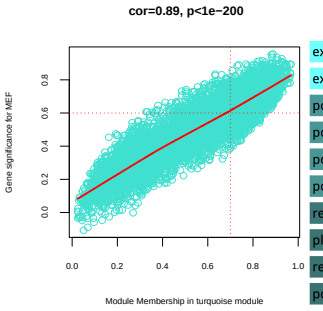
A similar approach to study gene co-expression network dynamics during human fibroblast reprogramming has been recently reported(Paul et al., 2016). However this published study had several technical issues which compromised the relevance of their findings. Firstly, the authors were unable to fully reprogram any of the assessed somatic cells into self-renewing iPSC cells, which failed to significantly express several key pluripotency regulators such as *NANOG*, *ESRRB*, *TBX3*, *TET1* and *TET2* and failed to form teratomas in mice suggesting a partial reprogramming at most. Secondly their microarray gene expression analysis was performed in the whole bulk cell population at different time points after adenoviral delivery only up to 3.5 days. The reprogramming process is very heterogeneous and asynchronous therefore the result of gene analysis of bulk cell populations are inherently noisy and imprecise. Moreover the expression profiles observed during the experiment's time-course suggests that the transient short pulse expression of the reprogramming factors produced by the adenoviral vector used in the study was insufficient to permanently change fate of the somatic cells.

The analyses presented in this and the next chapter successfully address all these limitations. The RNA-seq gene expression analysis was performed on sorted cell populations based on robust markers able to distinguish several stages from the starting somatic cells throughout the several intermediate reprogrammed stages up to fully reprogrammed iPS cells, resulting in reliable gene expression correlations with the various cell states. This analysis has revealed interesting

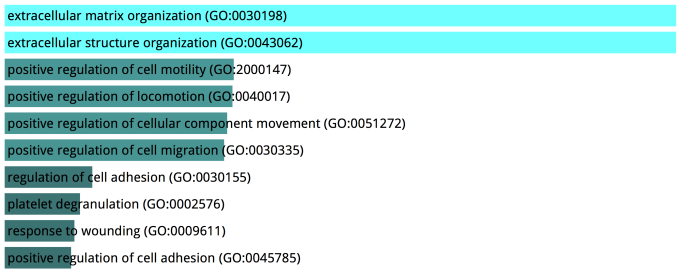
biological functions during the process of reprogramming and its reproducibility will be addressed in the next chapter using a separate data set. The modules that are shown to co-occur in different analyses have a high chance of being important components for the mechanisms of reprogramming and may offer the possibility for the identification of process bottlenecks which limits the efficiency of iPSC generation.



(a) Turquoise module eigengene expression

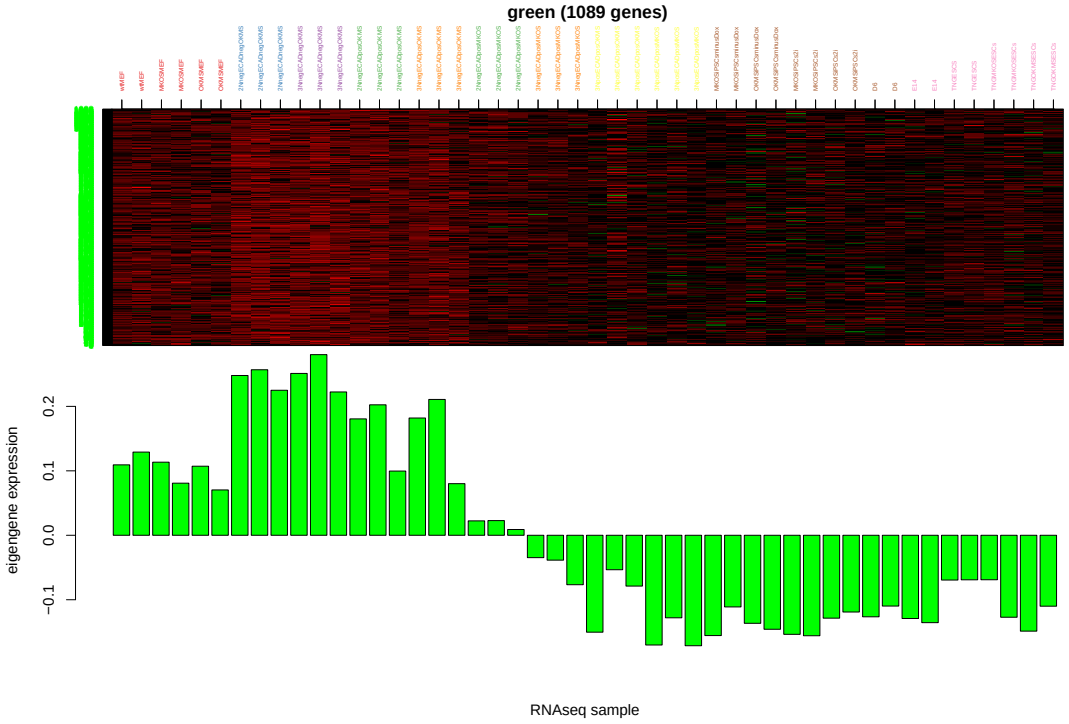


(b) Cutoff thresholds

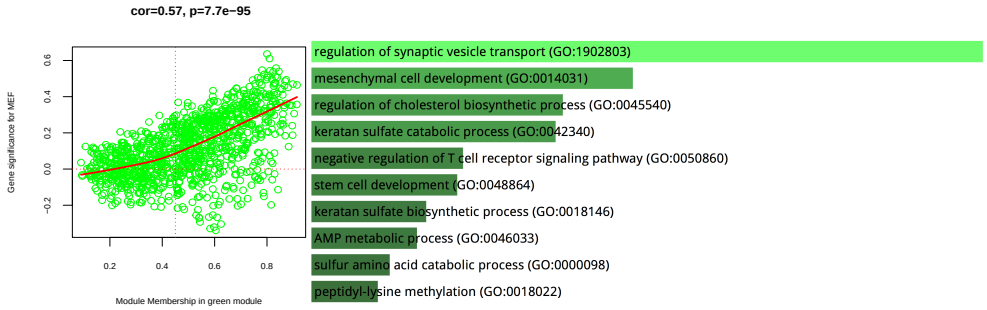


(c) Turquoise module Gene Ontology

Fig. 4.16 **Turquoise module characterisation:** The turquoise module is composed of 4588 genes whose eigenegene is significantly associated with MEF cells (Figure 4.16b: R = 0.89, p < 1e-200). Figure 4.16a: heatmap of turquoise module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the turquoise eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.16b: Selection of hub genes with high module membership values (MM >= 0.7) and gene significance for MEF cells (GS >= 0.6). Each spot in the scatter plot correspond to a gene and the dotted red gates represent the arbitrary thresholds for filtering the intra-modular hub genes. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.16c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.

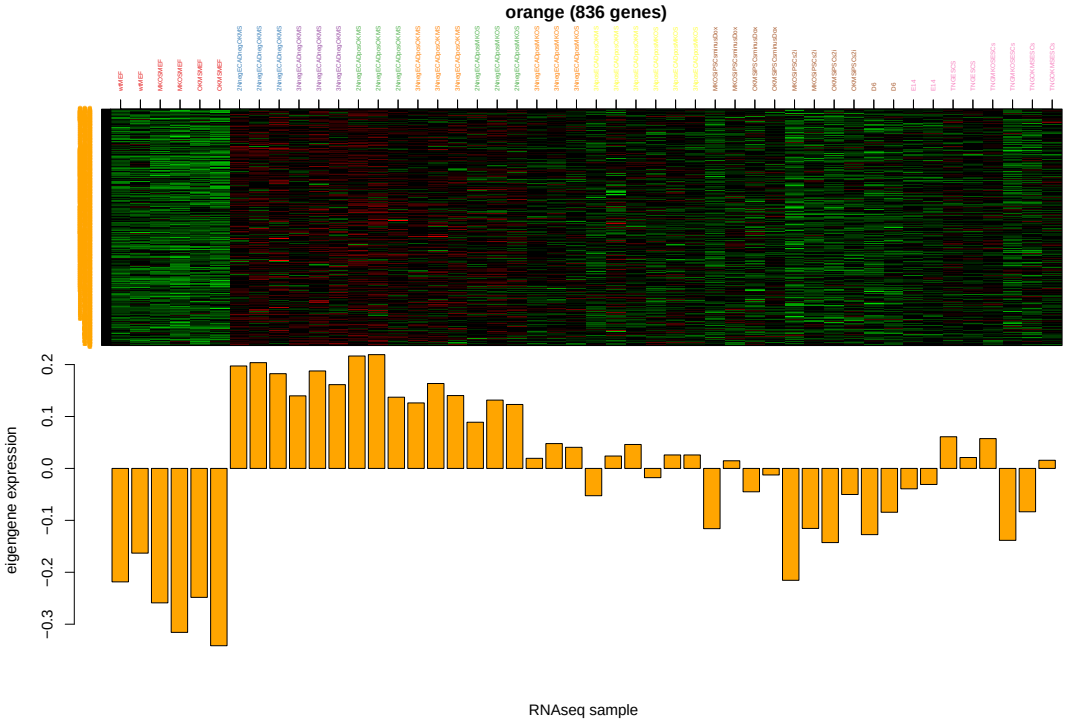


(a) Green module eigengene expression

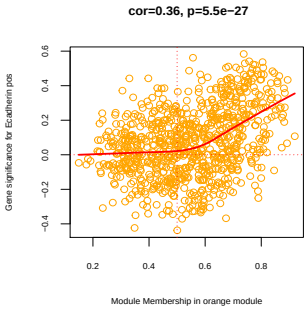


(b) Cutoff thresholds (c) Module Gene Ontology

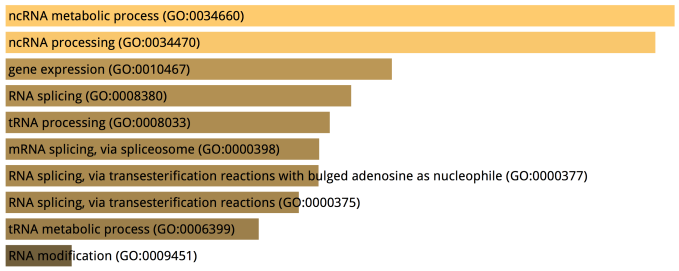
Fig. 4.17 **Green module characterisation:** The green module is composed of 1089 genes whose eigenegene is significantly associated with MEF cells (Figure 4.17b: R = 0.57, p = 7.7e-95). Figure 4.17a: heatmap of green module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the green eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.17b: Selection of hub genes with high module membership values (MM >= 0.42) and gene significance for MEF cells (GS >= 0). Each spot in the scatter plot correspond to a gene and the dotted red gates represent the arbitrary thresholds for filtering the intra-modular hub genes. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.17c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.



(a) Orange module eigengene expression

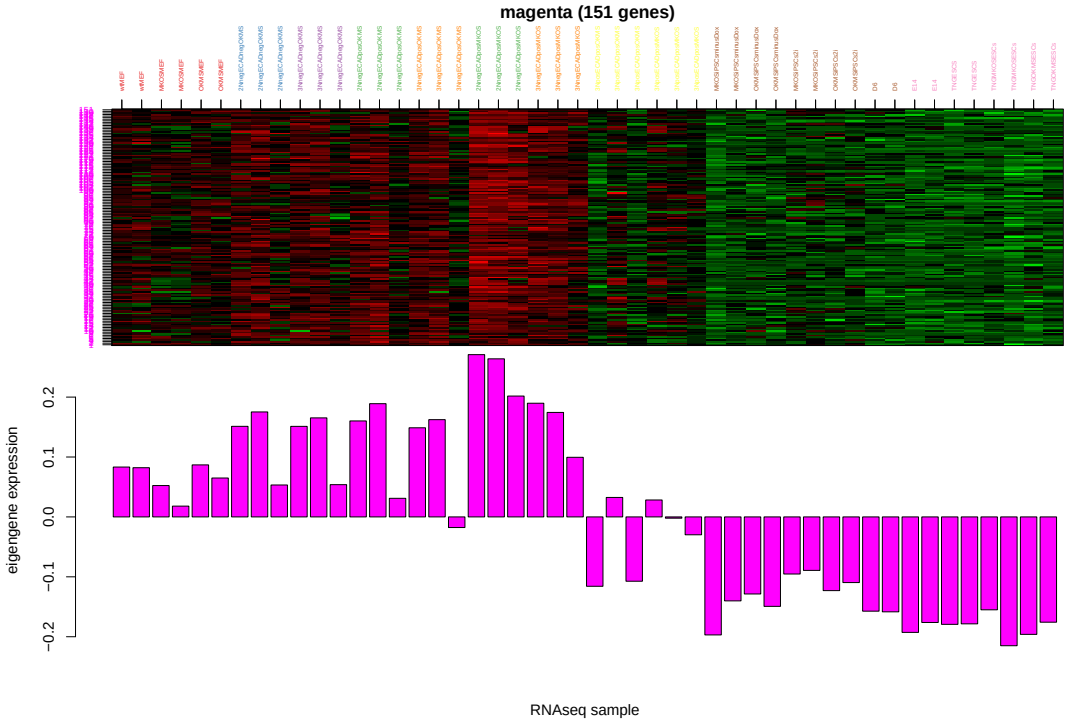


(b) Cutoff thresholds

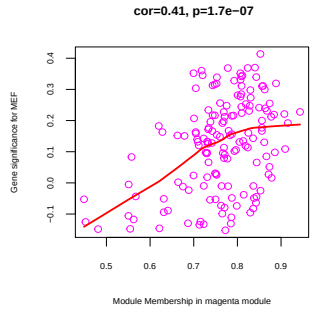


(c) Module Gene Ontology

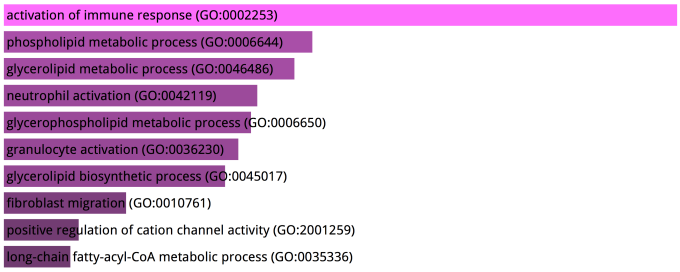
Fig. 4.18 **Orange module characterisation:** The orange module is composed of 836 genes whose eigenegene is significantly associated with Ecadehrin positive cells (Figure 4.18b: R = 0.36, p = 5.5e-27). Figure 4.18a: heatmap of orange module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the orange eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.18b: Selection of hub genes with high module membership values (MM >= 0.5) and gene significance for Ecadehrin positive cells (GS >= 0). Each spot in the scatter plot correspond to a gene and the dotted red gates represent the arbitrary thresholds for filtering the intra-modular hub genes. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.18c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.



(a) Magenta module eigengene expression

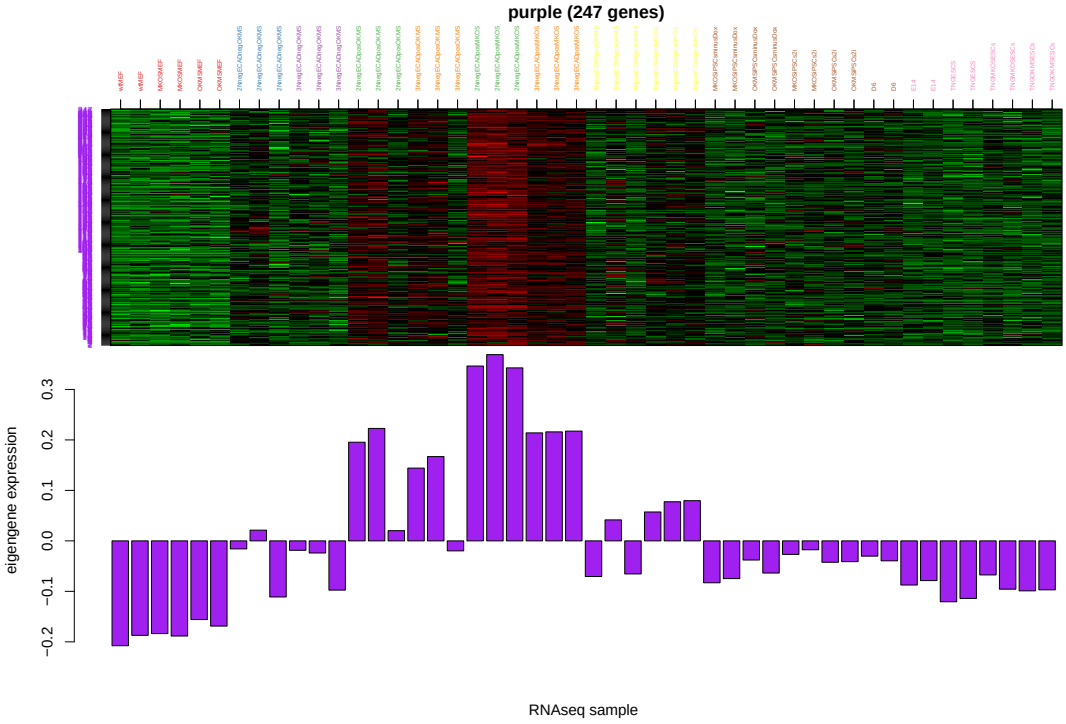


(b) Cutoff thresholds

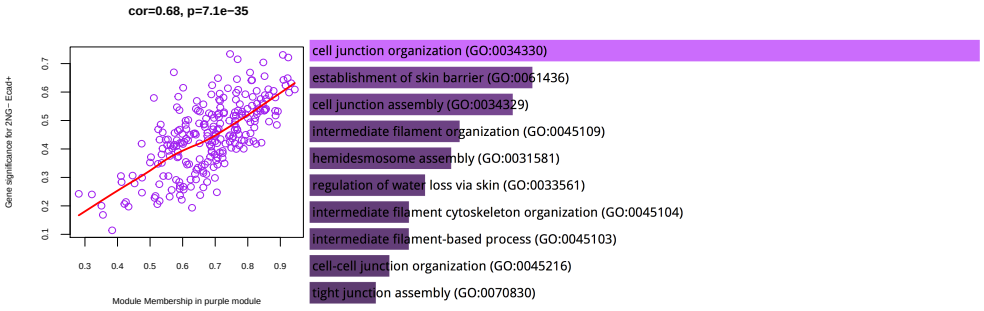


(c) Magenta module Gene Ontology

Fig. 4.19 **Magenta module characterisation:** The magenta module is composed of 151 genes whose eigenegene is significantly associated with MEF cells (Figure 4.19b: R = 0.41, p < 1.7e-07). Figure 4.19a: heatmap of magenta module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the magenta eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.19b: Each spot in the scatter plot correspond to a gene. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.19c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.

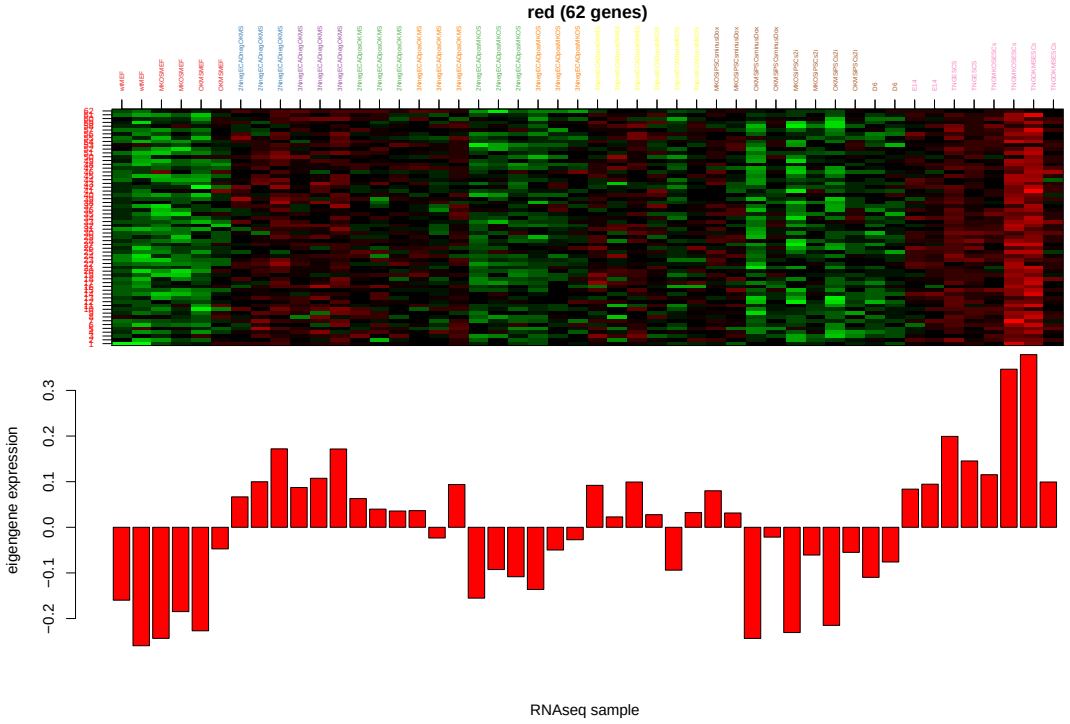


(a) Purple module eigengene expression

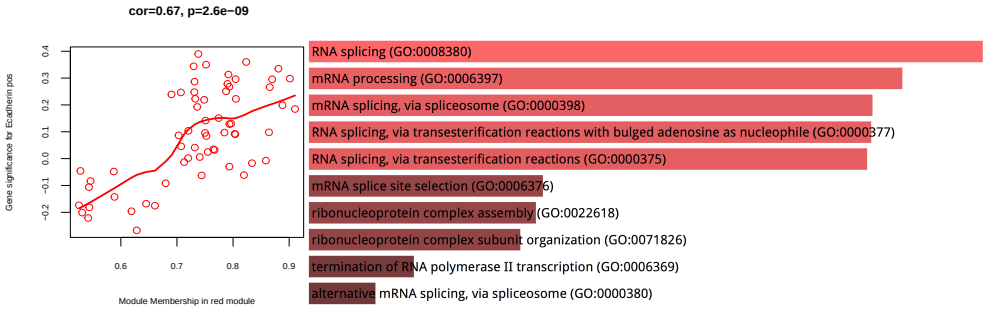


(b) Cutoff thresholds (c) Module Gene Ontology

Fig. 4.20 **Purple module characterisation:** The purple module is composed of 247 genes whose eigenegene is significantly associated with 2NG- Ecad+ cells (Figure 4.20b: R = 0.68, p < 7.1e-35). Figure 4.20a: heatmap of purple module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the purple eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.20b: Each spot in the scatter plot correspond to a gene. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.20c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.

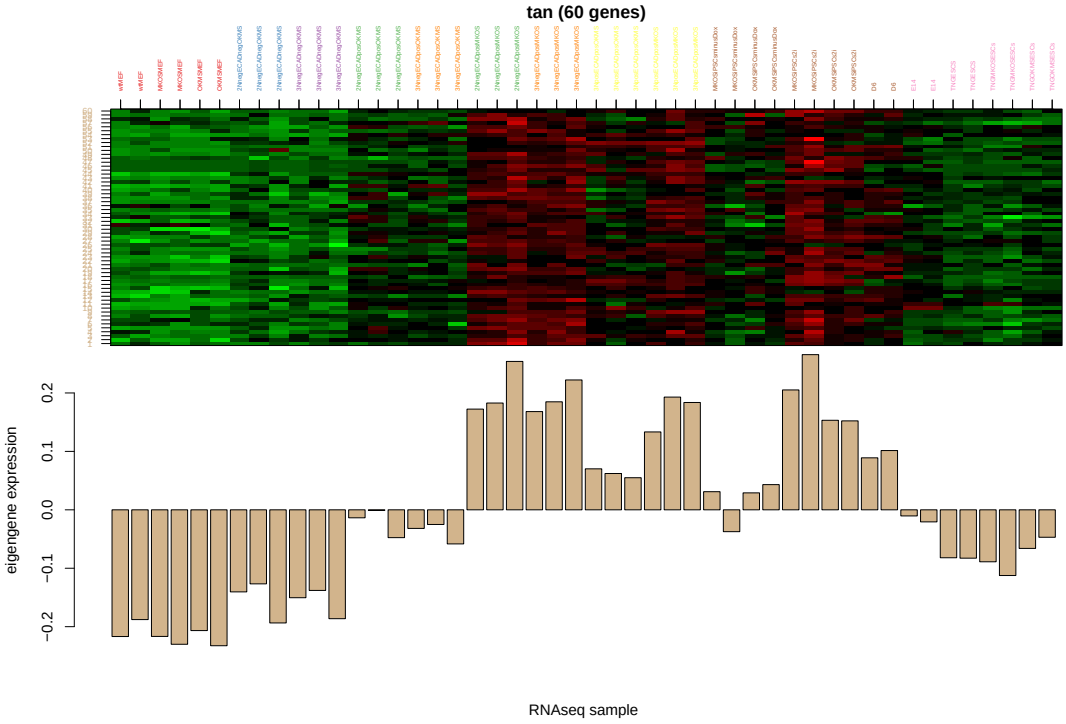


(a) Red module eigengene expression

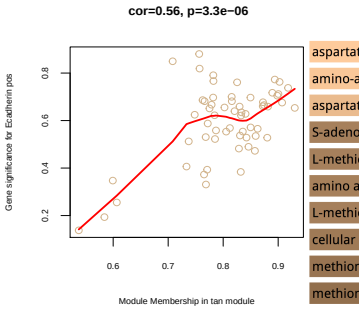


(b) Cutoff thresholds (c) Red module Gene Ontology

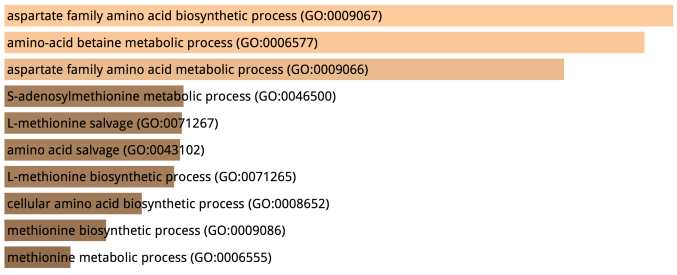
Fig. 4.21 **Red module characterisation:** The red module is composed of 62 genes whose eigenegene is significantly associated with Ecadehirin positive cells (Figure 4.21b: R = 0.67, p < 2.6e-09). Figure 4.21a: heatmap of red module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the red eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.21b: Each spot in the scatter plot correspond to a gene. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.21c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.



(a) Tan module eigengene expression

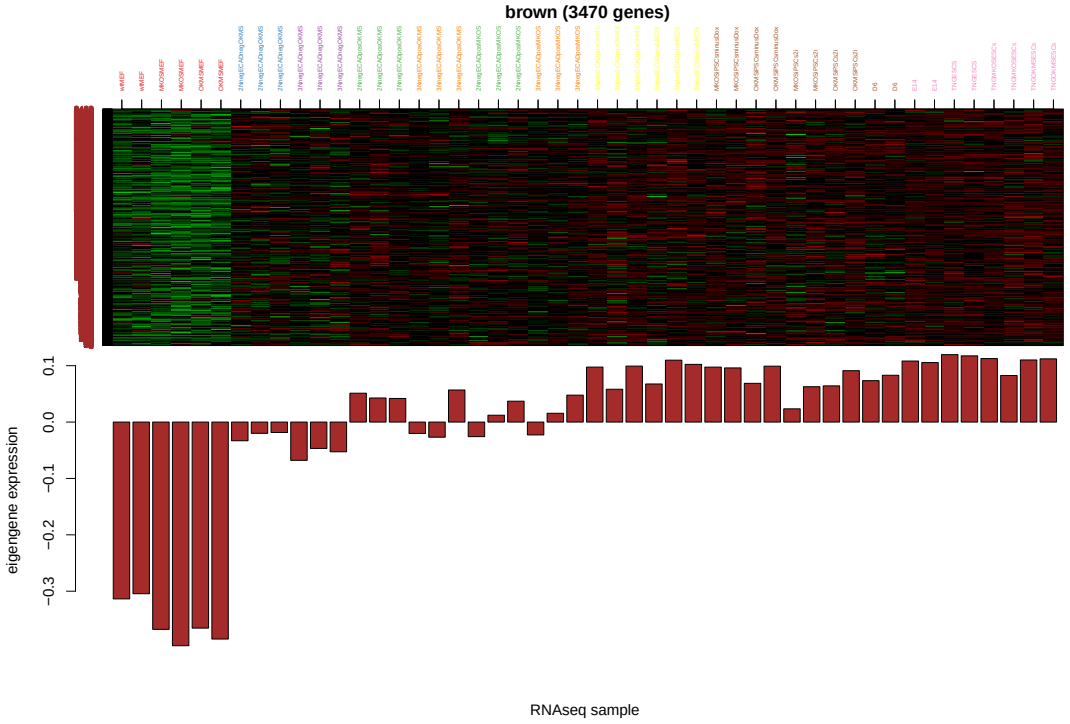


(b) Cutoff thresholds

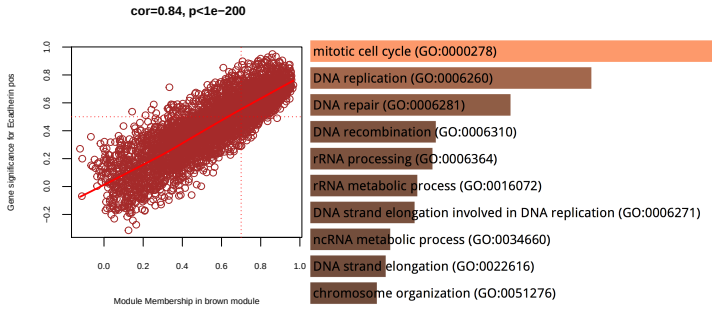


(c) Tan module Gene Ontology

Fig. 4.22 **Tan module characterisation:** The tan module is composed of 60 genes whose eigenegene is significantly associated with Ecadehrin positive cells (Figure 4.22b: R = 0.56, p < 3.3e-06). Figure 4.22a: heatmap of tan module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the tan eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.22b: Each spot in the scatter plot correspond to a gene. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.16c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.

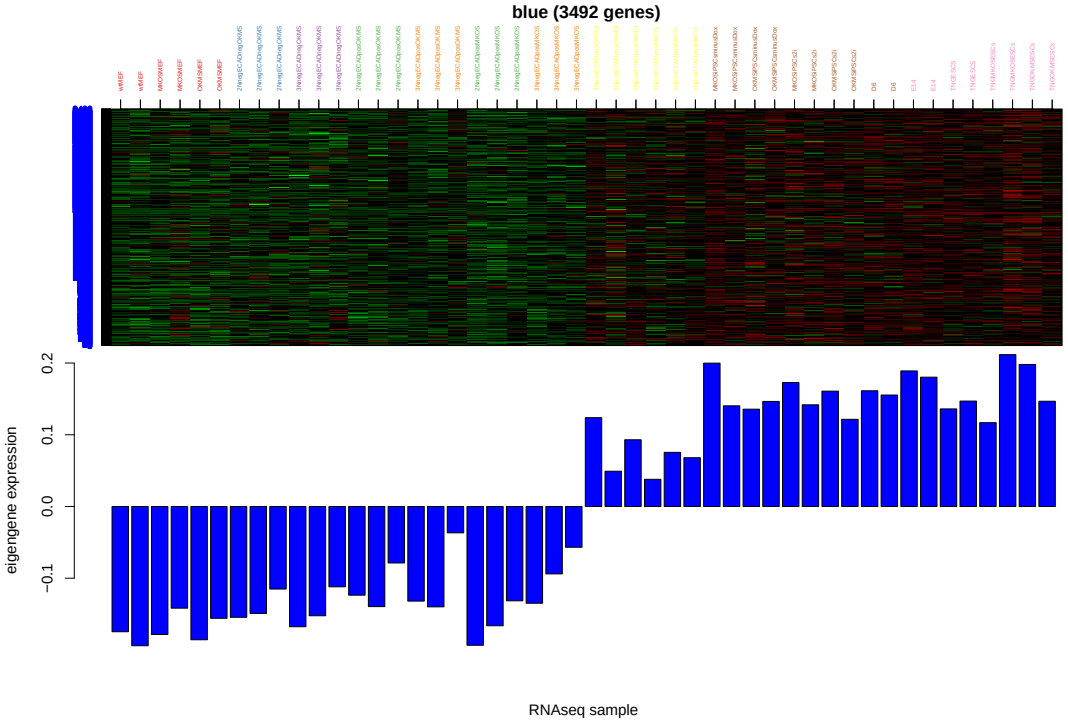


(a) Brown module eigengene expression

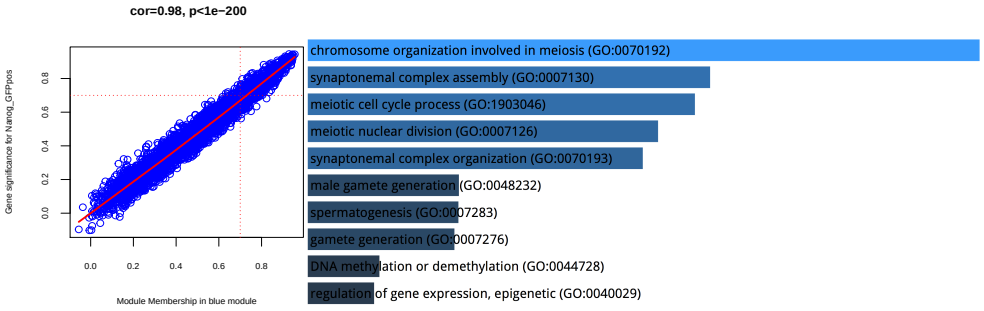


(b) Cutoff thresholds (c) Brown module Gene Ontology

Fig. 4.23 **Brown module characterisation:** The brown module is composed of 3470 genes whose eigenegene is significantly associated with Ecadehrin positive cells (Figure 4.23b: R = 0.84, p < 1e-200). Figure 4.23a: heatmap of brown module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the brown eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.23b: Selection of hub genes with high module membership values (MM >= 0.7) and gene significance for Ecadehrin positive cells (GS >= 0.5). Each spot in the scatter plot correspond to a gene and the dotted red gates represent the arbitrary thresholds for filtering the intra-modular hub genes. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.23c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.



(a) Blue module eigengene expression



(b) Cutoff thresholds (c) Blue module Gene Ontology

Fig. 4.24 **Blue module characterisation:** The blue module is composed of 3492 genes whose eigenegene is significantly associated with Nanog-GFP positive cells (Figure 4.24b: R = 0.98, p < 1e-200). Figure 4.24a: heatmap of blue module gene expression with the rows corresponding to genes and the columns corresponding to the 48 cell samples. Red and green color codes represent the scaled gene expression values as over-expression and under-expression respectively. The bar plots below the heatmap represent the expression levels of the blue eigenvector (eigengene) across the 48 samples with each bar corresponding to a principal component of the eigenvector. The value of an eigengene is high when most genes in the module are over-expressed in a given sample (red in the heatmap) or the value is low when under-expressed (green in the heatmap). Figure 4.24b: Selection of hub genes with high module membership values (MM >= 0.7) and gene significance for Nanog-GFP positive cells (GS >= 0.7). Each spot in the scatter plot correspond to a gene and the dotted red gates represent the arbitrary thresholds for filtering the intra-modular hub genes. The red line correspond to a loess regression fit and the correlation coefficient and p-value is printed above the plot. Figure 4.24c: The filtered hub genes were used for gene ontology function enrichment analysis and the results are represented as a bar plot. The enriched GO term is used as bar labels and the length of the bars represent the enrichment level and the color brightness of the bar correspond to the significance of enrichment with grey being non significant.

Chapter 5

Reproducibility of Reprogramming Gene Networks

5.1 Introduction

This chapter presents the re-analysis of the RNA-Seq data set produced by O'Malley et al. (2013b), which developed a high-resolution analysis system for the study of reprogramming somatic cells to pluripotency. This original study was the precursor to the work developed by Chantzoura et al. (2015) using similar experimental strategies (Figure 5.1). Therefore this data set was chosen for the purpose of assessing the reproducibility of the gene network analysis performed in the previous chapter and to further characterise the biological processes associated with the various intermediate stages of reprogramming somatic cells to pluripotency. A better understanding of such consensus processes should enable the identifica-

tion of facilitator features and possible bottlenecks for reprogramming which could be manipulated for improving iPS generation efficiency.

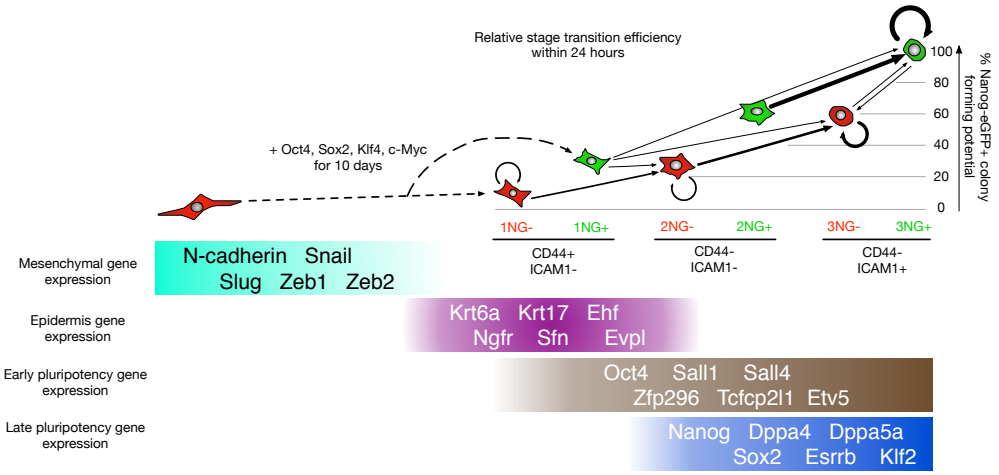


Fig. 5.1 Experimental strategy for the analysis of iPS cell generation stages (adapted from O'Malley et al. (2013b)).

5.2 Results and Discussion

5.2.1 Data set manipulation

The data set from the study performed by O'Malley et al. (2013b) was acquired as raw FASTQ RNA-sequencing data which were stored by the authors in the ArrayExpress public repository, hosted by the European Bioinformatics Institute under the accession number E-MTAB-1654 and ERP002573 (Secondary accession at European Nucleotide Archive – ENA). Table 5.1 lists the utilized samples, cell type information, marker profiles used for cell sorting and their respective sequencing run identification numbers. The sequenced samples are composed of

mouse embryonic fibroblasts, ES and iPS cells and 6 intermediate reprogramming cell populations types, all with at least 2 biological replicates per group. The iPS and intermediate cells were generated using a doxycycline inducible secondary reprogramming system and the different intermediate reprogramming cell populations were sorted based on a combination of two cell surface markers (CD44 and ICAM1) and a Nanog expression reporter (eGFP knock in). The nomenclature used for identifying the different intermediate stages of reprogramming are the following: the number 1, 2 and 3 in the start of the name refers to the cell surface profiles CD44 +ve/ICAM1 –ve, CD44 -ve/ICAM1 –ve and CD44 -ve/ICAM1 +ve respectively. N- and N+ refers to Nanog-GFP negative and positive and r1-3 indicate the number of the biological replicate. All intermediate reprogramming cell populations were sorted after 10 days of reprogramming factors (OKMS) induction except for the Day15 3N+ samples, which were sorted 5 days later.

The single-end read sequencing FASTQ files were directly uploaded to the Galaxy main instance server for mapping to mouse genome. The quality of the reads was first inspected using FastQC v0.11.2 (Galaxy Version 0.65) read quality reports tool. The sequencing read lengths ranged between thirty-three and thirty-nine base pairs with the total number of sequences varying between around ten and forty million reads per library which all had the mean sequencing quality per base above the Phred score of thirty. Figure 5.2 illustrates the per base quality score distribution of a representative run (ERR268441). The reads were then mapped to the mouse genome build mm10/GRCm38 using the gapped-read aligner tool TopHat v2.0.14 (Galaxy version 0.9) with the default settings for single-end RNA-

sequencing data. The overall read mapping rate was all around 95 percent of which around 20 percent had multiple alignments and the mean per base quality score were maintained above the Phred score of thirty as illustrated by Figure 5.3.

After confirming quality, BAM files were downloaded and used for downstream analyses using R statistical software. The list of BAM files was managed with the Bioconductor package *Rsamtools* for preparing them for mapping to annotated mouse genes. For this purpose the gene annotation reference provided by the package *TxDb.Mmusculus.UCSC.mm10.knownGene* was used by the *GenomicFeatures* package function *exonsBy* for grouping the non-overlapping exons of all transcripts per gene. A gene count table was then generated using the function *summarizeOverlaps* using the settings mode as *union* and *singleEnd* as TRUE resulting in a Bioconductor *RangedSummarizedExperiment* object which was used throughout the rest of the analyses to help manage the multiple parts of this dataset. The gene count table can be found in the attached media (omalley_gene_counts.csv). The samples metadata was collected using the package *ArrayExpress* and further processed for the creation of some relevant variables such as cell types, presence of cell sorting makers CD44 and ICAM and Nanog-GFP reporter gene expression. The obtained metadata was then stored in the *colData* slot of a *RangedSummarizedExperiment* object designed for storing sample phenotypic data and other metadata. A spreadsheet containing the full metadata for this data set can be found in the attached media (omalley_metadata.csv).

Table 5.1 O'Malley *et al.* RNA-sequencing data set

Run	Sample Name	Cell Type	FACS Markers
ERR268464	MEFr1	MEF cell	-
ERR268447	MEFr2	MEF cell	-
ERR268465	1N-r1	intermediate population	CD44 +ve, ICAM1 –ve, Nanog-eGFP -ve
ERR268443	1N-r2	intermediate population	CD44 +ve, ICAM1 –ve, Nanog-eGFP -ve
ERR268459	1N-r3	intermediate population	CD44 +ve, ICAM1 –ve, Nanog-eGFP -ve
ERR268445	1N+r1	intermediate population	CD44 +ve, ICAM1 –ve, Nanog-eGFP +ve
ERR268442	1N+r2	intermediate population	CD44 +ve, ICAM1 –ve, Nanog-eGFP +ve
ERR268457	2N-r1	intermediate population	CD44 -ve, ICAM1 –ve, Nanog-eGFP -ve
ERR268455	2N-r2	intermediate population	CD44 -ve, ICAM1 –ve, Nanog-eGFP -ve
ERR268460	2N+r1	intermediate population	CD44 -ve, ICAM1 –ve, Nanog-eGFP +ve
ERR268450	2N+r2	intermediate population	CD44 -ve, ICAM1 –ve, Nanog-eGFP +ve
ERR268446	2N+r3	intermediate population	CD44 -ve, ICAM1 –ve, Nanog-eGFP +ve
ERR268461	3N-r1	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP -ve
ERR268454	3N-r2	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP -ve
ERR268448	3N-r3	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP -ve
ERR268456	3N+r1	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP +ve
ERR268452	3N+r2	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP +ve
ERR268444	3N+r3	intermediate population	CD44 -ve, ICAM1 +ve, Nanog-eGFP +ve
ERR268449	Day15 3N+r1	intermediate population	CD44 -ve, ICAM1 +ve , Nanog-eGFP +ve
ERR268453	Day15 3N+r2	intermediate population	CD44 -ve, ICAM1 +ve , Nanog-eGFP +ve
ERR268441	iPSCr1	iPSC	-
ERR268463	iPSCr2	iPSC	-
ERR268451	E14r1	ESC	-
ERR268458	E14r2	ESC	-
ERR268462	E14r3	ESC	-

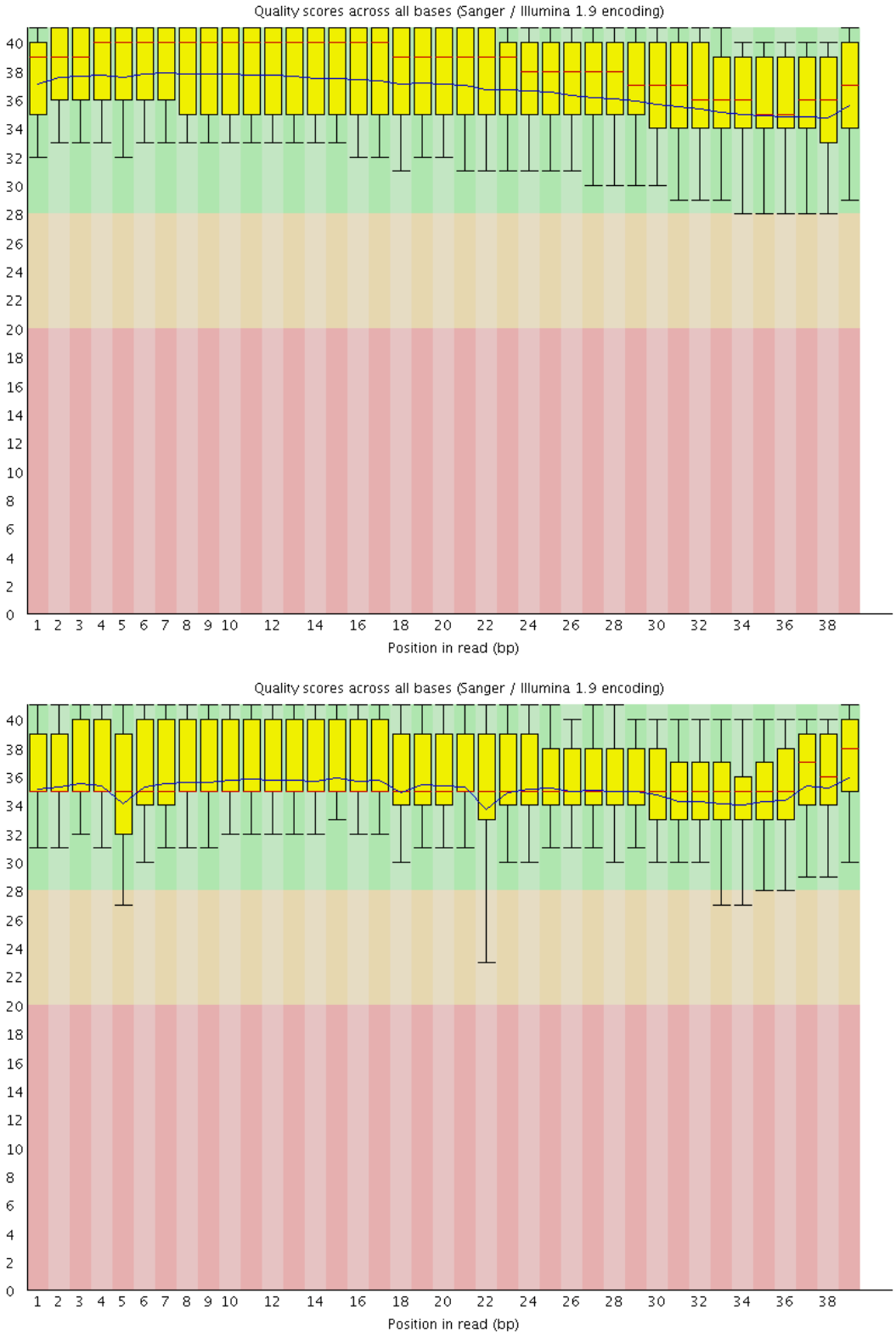


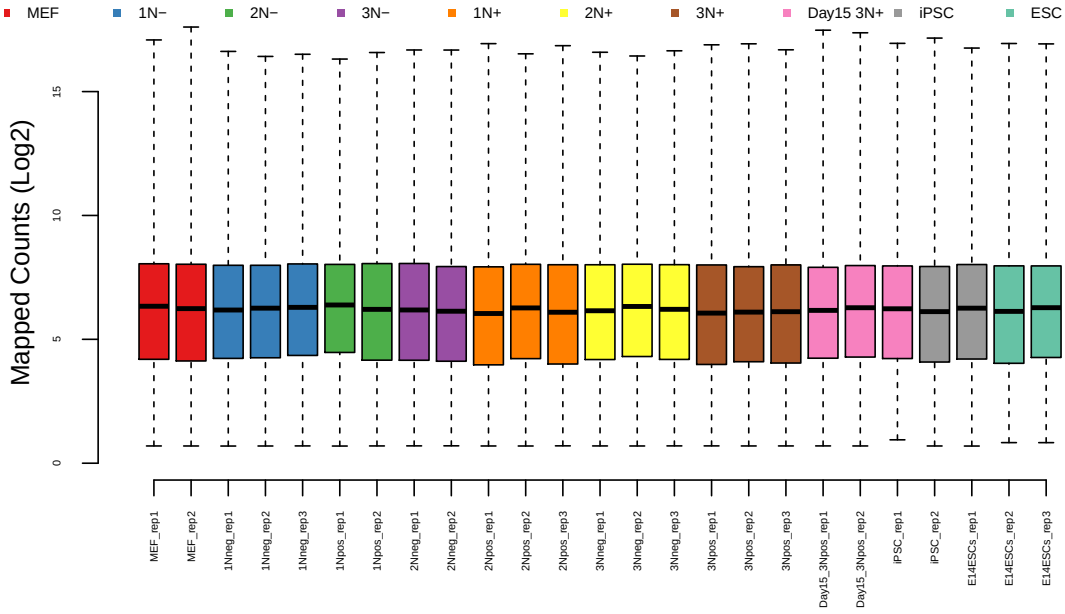
Fig. 5.3 FastQC per base quality score distribution of mapped ERR268441 BAM file

5.2.2 Exploratory Data Analysis

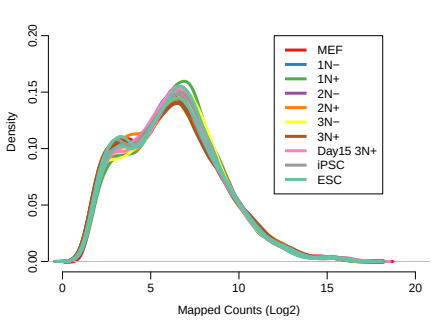
To further determine the quality and get acquainted with the properties of the data set produced by O'Malley et al. (2013b). I performed a series of exploratory data analysis techniques. This step is important for investigating for potential outlier libraries and to confirm grouping of samples. To help reduce expression noise, genes which had an expression count below 10 in eighty percent of samples were filtered out of the data set resulting in 13430 genes to be analysed. Regularized logarithm transformation was applied to the gene count data set using the package *DESeq2* as a mean for stabilizing the data variance as this is a robust method to size factor variation and prepare the data for further analysis such as clustering of samples and other downstream machine learning applications. The transformed dataset can be found in the attached media (omalley_rlog_expression.csv)

Some descriptive statistical properties of the transformed dataset are illustrated in Figure 5.4. The box plots in Figure 5.4a depict the distribution of the RNA-seq data sets representing the expression levels (regularised log transformed) for all 13430 genes. We can observe that the distribution of all samples is fairly similar to each other with regards to both sample groups and biological replicates. This property can be further confirmed by the density plot in figure 5.4b indicating that the data set is well normalized and no outlier library was detected. The scatter plot in Figure 5.4c was used to evaluate the correlation of the expression mean and standard deviation and indicates that there is no obvious dependence

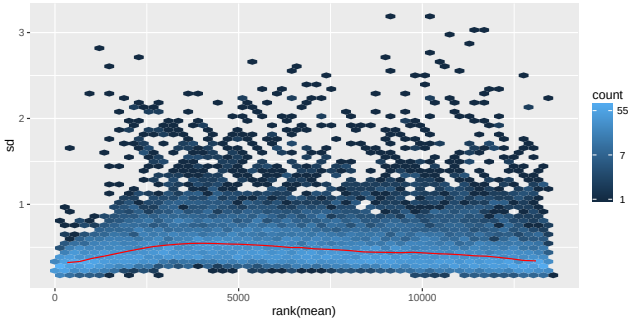
between these two properties. The horizontal red line highlights that the data set is approximately homoscedastic.



(a) Box plot



(b) Density plot



(c) Standard deviation vs. ranked mean

Fig. 5.4 O'Malley data set mapped reads statistics. (a) Box plot: Distribution of the RNAseq data sets representing the expression levels (regularized log transformed) for all 13430 genes. The sample groups are color coded as displayed on the legend above the plot; (b) Density plot: Log-density distribution of all libraries superposed and color coded as per legend insert; (c) Gene expression standard deviation vs. ranked mean: The color intensity of scatter plot reflects the expression level as per legend. The red line depicts the running median estimator (window-width 10%)

Unsupervised Hierarchical Clustering

The relationship between samples as well as the coherence of cell surface markers and the expression of Nanog-GFP gene reporter was investigated by performing sample hierarchical clustering. Figure 5.5 shows a heatmap representing the pairwise Euclidean distances between all samples based on regularised log transformed expression of all 13430 genes analysed. We can see that the MEF, and the pluripotent cell samples (ESCs, iPSCs and Day15 3N+) clearly form discrete groups with perfect agreement for the grouping of biological replicates. As for the intermediate reprogramming population samples the clusterings were not so much clear cut with the biological replicate samples for 3N+ being the only sample group that clustered with each other perfectly.

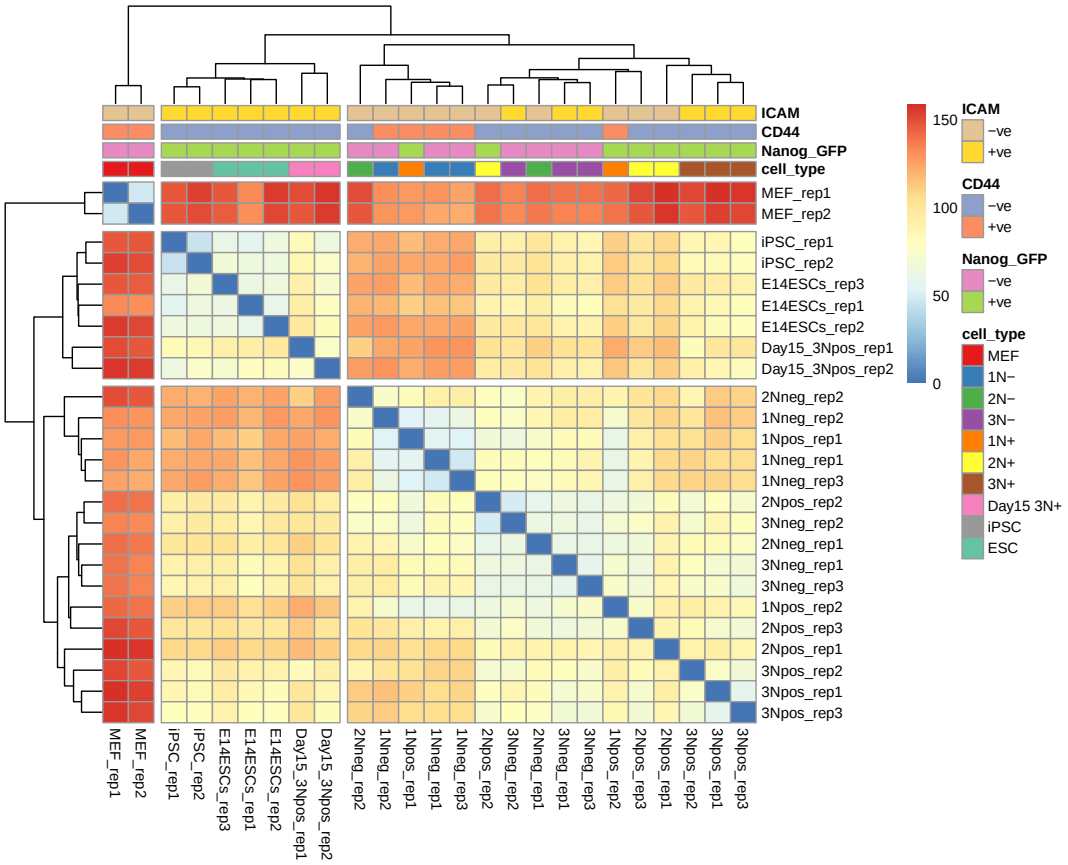
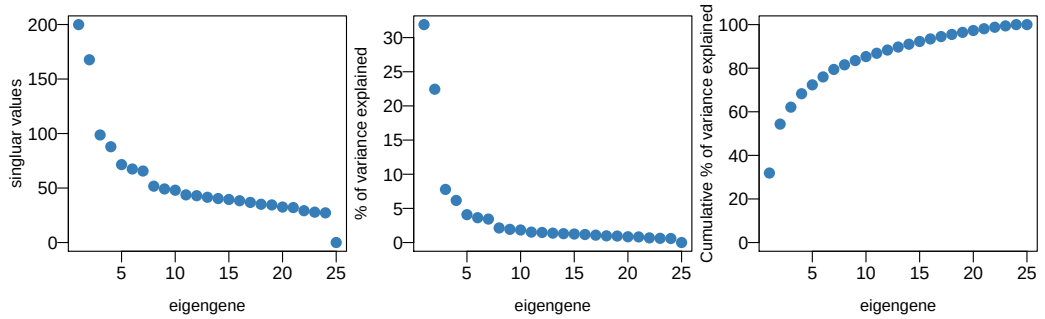


Fig. 5.5 O'Malley data set sample clustering. The heatmap represents the clustered pairwise euclidean distances between all samples based on their regularised log transformed expression of all all 13430 genes analysed. The euclidean distances range from 0 (dark blue) to 150 (red) as depicted in the legends next to the plot. The two surface markers ICAM1 and CD44, the reporter gene Nanog-GFP as well as the sample groups MEF, 1N-, 1N+, 2N-, 2N+, 3N-, 3N+, Day15 3N+, iPSCs and ESCs were also color coded to assist in the sample clustering visualisation.

Principal Component Analysis

The regularised log transformed expression data set was scaled by subtracting the expression of each gene from all samples by their means, transposed and then used to perform a singular value decomposition of the data matrix. Figure 5.6 illustrates the analysis. Figure 5.6a displays singular values for each principal component (eigengene), Figure 5.6b shows the proportion of explained variance per eigengene and Figure 5.6c illustrates the cumulative explained variance. The first two principal components explain around 32 and 23 percent of the data set variance respectively with a cumulative variance explained of around 55 percent which was enough for separating well most of the sample groups as shown by the principal component analysis plot in Figure 5.7. Similar to the hierarchical clustering results the MEF and pluripotent samples formed very distinct tight groups and the intermediate reprogramming samples were also easily distinguishable groups, different to the hierarchical clustering. Most importantly the separation and positioning of the sample groups in the plot formed a clear path coinciding with experimental set up based on the markers used with the following order: MEF, 1N-, 1N+, 2N-, 2N+, 3N-, 3N+, Day15 3N+, iPSCs and ESCs. This observation gives confidence both in the quality of the data set as well as in the possibility of characterising the expression profile of each of these groups of samples based on their eigengene profiles as used in WGCNA.



(a) Singular values (b) Variance explained (%) (c) Cumulative variance

Fig. 5.6 O'Malley data set singular value decomposition

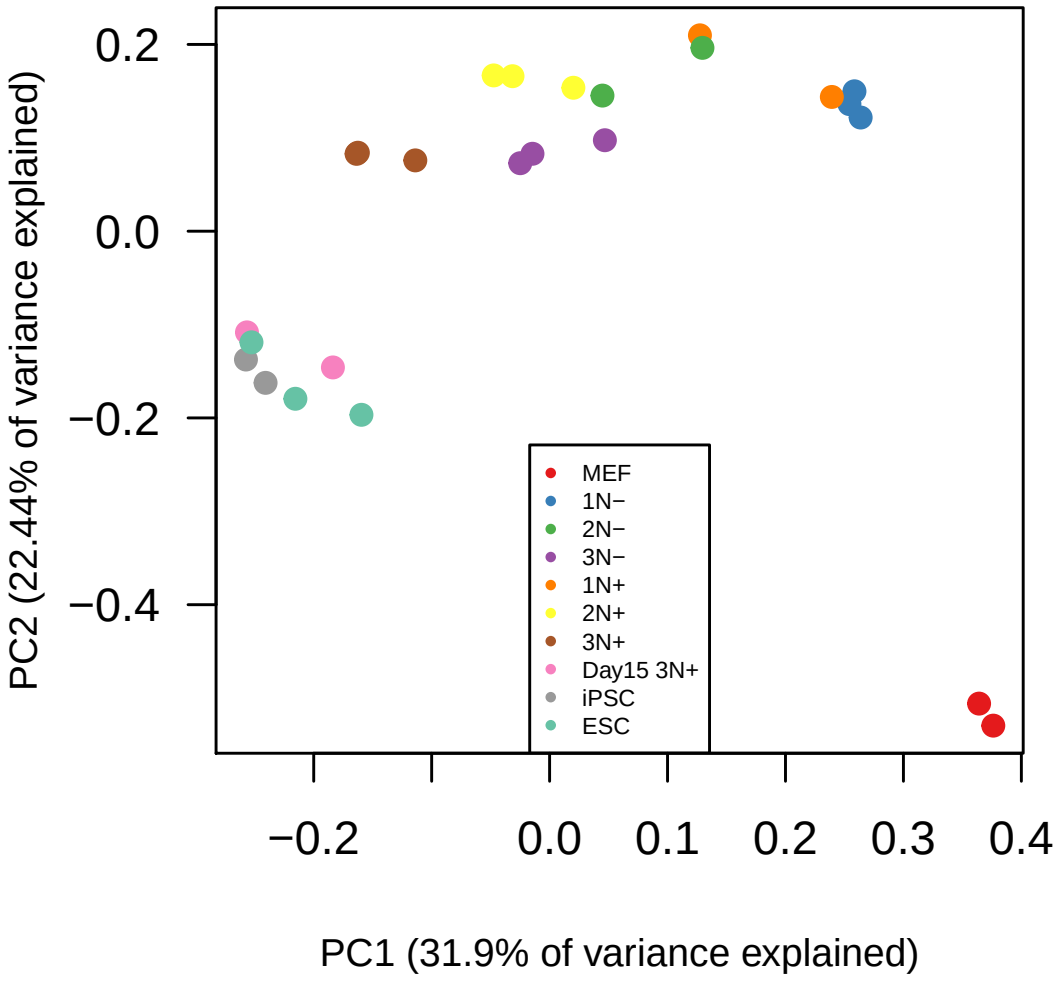
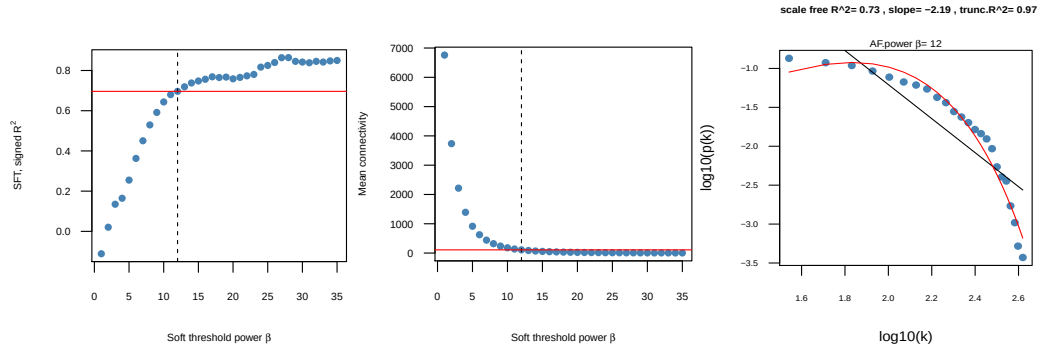


Fig. 5.7 Principal component analysis of O'Malley data set. Principal components (eigengenes) were calculated via singular value decomposition and the first two principal components were plotted with the samples coloured as per legend.

5.2.3 Weighted Gene Co-Expression Network Analysis

Scale Free Topology Fitting

In order to build a weighted gene co-expression network I selected the lowest possible β value which is used to transform the network adjacency matrix with the purpose of adjusting the network so it approximately satisfy the scale-free topology criterion. Figure 5.8 illustrates the analysis of the effect of power β values ranging from 1 to 35 on network properties such as connectivity and scale-free fit index. A spreadsheet with the resulting network properties as a function of the tested power thresholds can be found in the attached media (omalley_power_threshold.csv). The function *pickSoftThreshold* from the WGCNA package was used for calculating the weighted network by transforming the data set into a co-expression adjacency matrix and raising its values by the range of β values to be tested. The plot in Figure 5.8a shows that the power β value of 12 corresponds to a scale free connectivity index of around 0.7 and after that it begins to saturate. The plot in Figure 5.8b shows the mean connectivity is a decreasing function of the power β and reaches nearly 0 mean connectivity value when the soft-threshold is set to 12. To further evaluate how well the network satisfies the scale-free topology criterion when the power β is set to 12, the log10 of the connectivities (log10(κ)) was plotted as a function of the log10 of the frequency of the connectivities (log10(p(κ))) as shown in Figure 5.8c. The red line shows that the network connectivity model fits very well with the exponentially truncated power law model which is an alternative model of scale-free topology (Csányi and Szendrői, 2004).



(a) Soft threshold selection (b) Mean connectivity (c) Scale free fitting

Fig. 5.8 Selection of power β for the O'Malley data set

Construction of a Weighted Gene Co-Expression Network

A signed weighted gene co-expression network was constructed following the guidelines from the authors of the WGCNA method (Zhang and Horvath, 2005) as described in the previous chapter (Section 4.2.2). Figure 5.9 illustrates the resulting network module definition for the 13430 genes analysed. As previously described in Section 4.2.2 the resulting module eigengenes with similar expression profiles were merged together. Figure 5.10 shows the hierarchical clustering of the module eigengenes with the horizontal red line representing the arbitrary tree cutting height chosen for merging similar modules falling below that threshold. The reassigned final modules are illustrated in Figure 5.11. The module merging produced more well defined gene clusters which have clear differential associations with the different cell types utilised in this data set. Some of the final modules still had a relatively high correlation to each other as illustrated in the correlation heatmap in Figure 5.12. A spreadsheet containing the module assignment (mergedColors) per gene can be found in the attached media (omalley_module_assignment.csv).

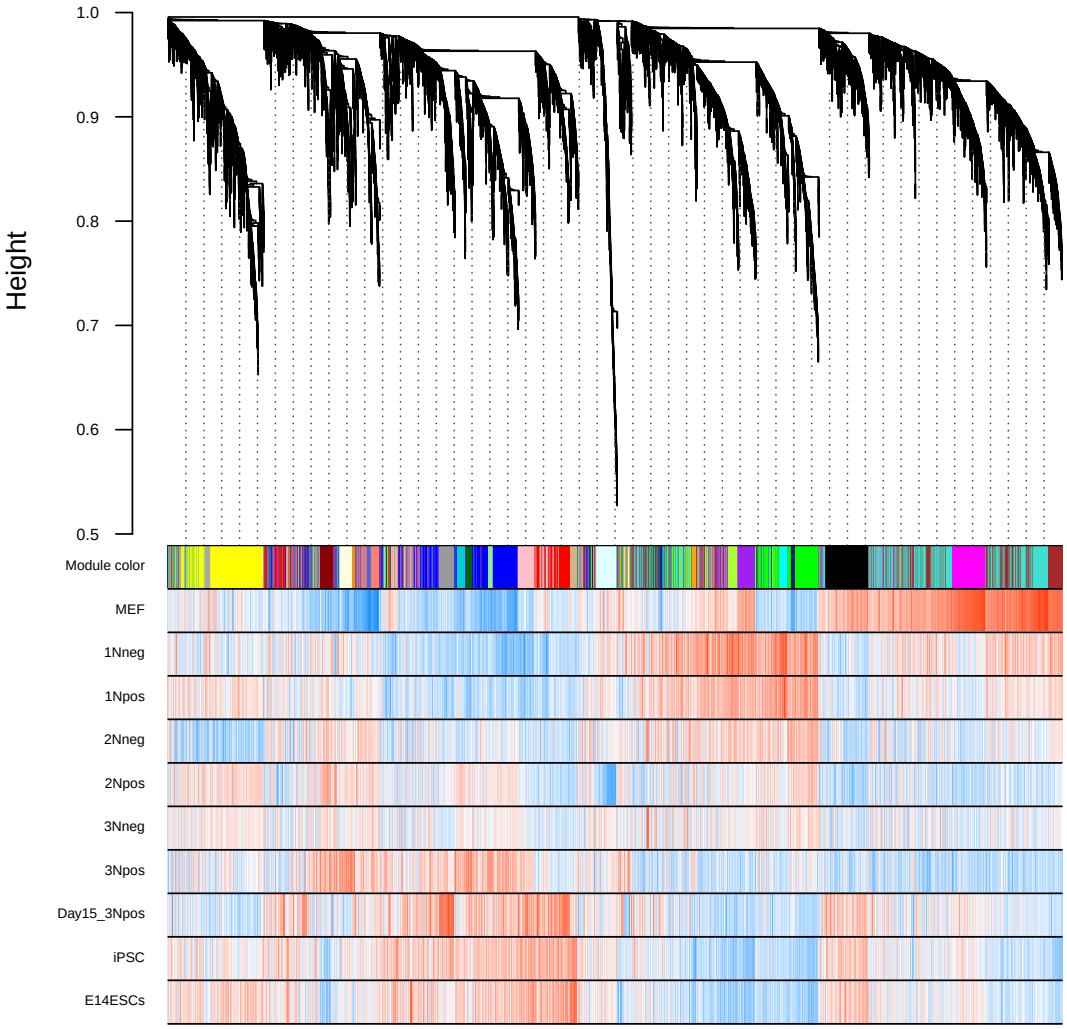


Fig. 5.9 Network module definition for the O'Malley data set. Gene dendrogram obtained from the reprogramming dataset of O'Malley et al. 2013. This tree was defined by hierarchical clustering the dissimilarity based on consensus Topological Overlap measures using average linkage of the 13430 genes analysed. The color coded modules were assigned based on the dynamic hybrid branch cutting method represented as the first color band below the dendrogram. The other color bands represent the gene significance measures to all the cell samples analysed: "red" indicates a high positive correlation with the a cell type and "blue", high negative correlation.

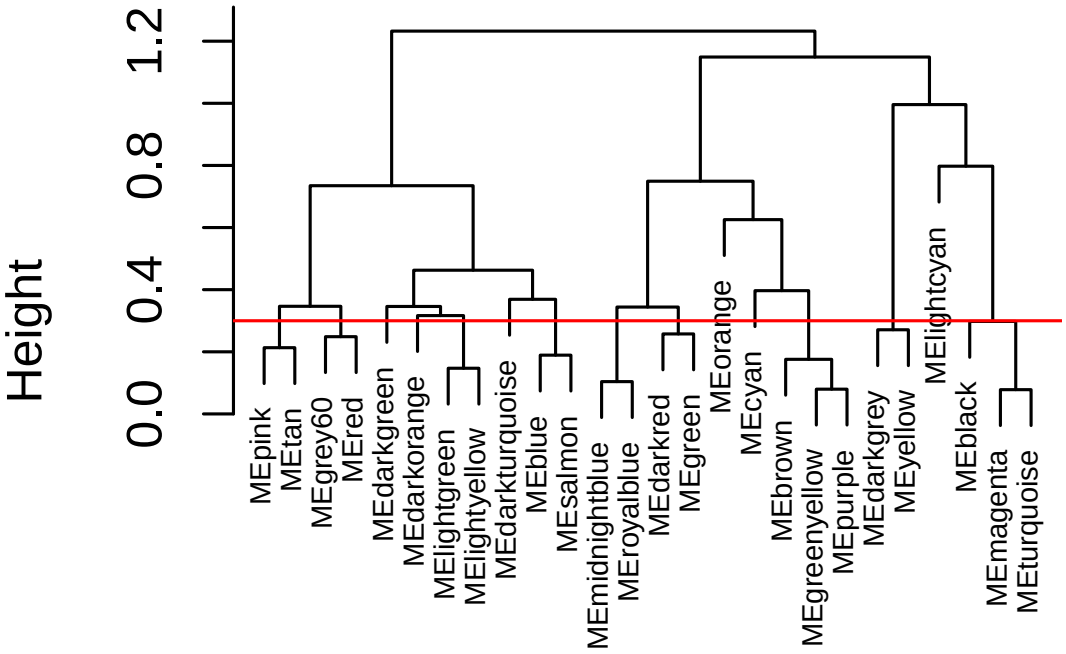


Fig. 5.10 **Module merging height for the O'Malley data set.** Hierarchical clustering of module eigengenes. The red line indicates the arbitrary height of 0.3 selected for merging very similar module eigengenes.

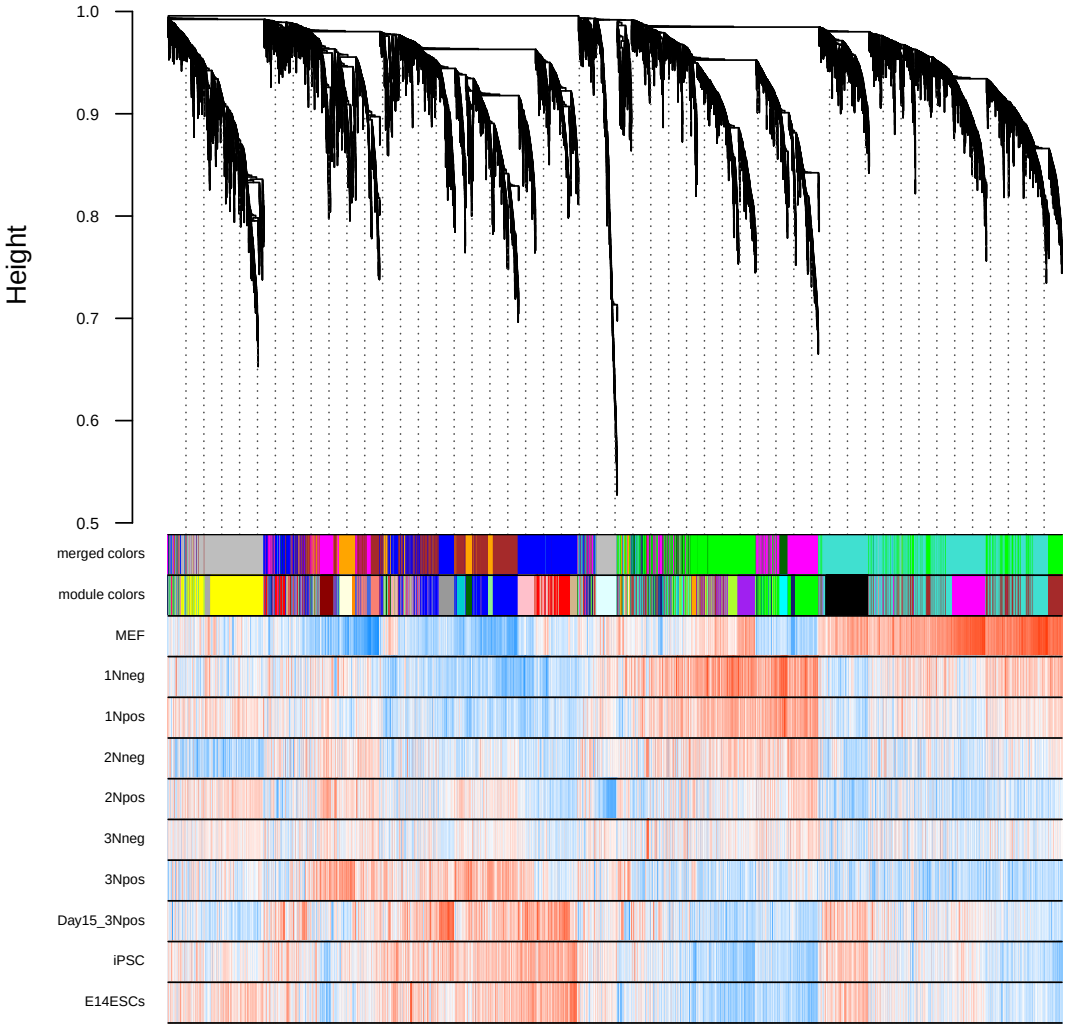


Fig. 5.11 **Network module definition after merges for the O'Malley data set.** This tree was defined by hierarchical clustering the dissimilarity based on consensus Topological Overlap measures using average linkage of the 13430 genes analysed as shown before in figure 5.9 but this time the highly correlated modules were merged. The color coded modules were assigned based on the dynamic hybrid branch cutting method represented as the first color band below the the dendrogram. The several other color bands (blue and red) represent the gene significance measures to all the cell samples analysed: "red" indicates a high positive correlation with the a cell type and "blue", high negative correlation.

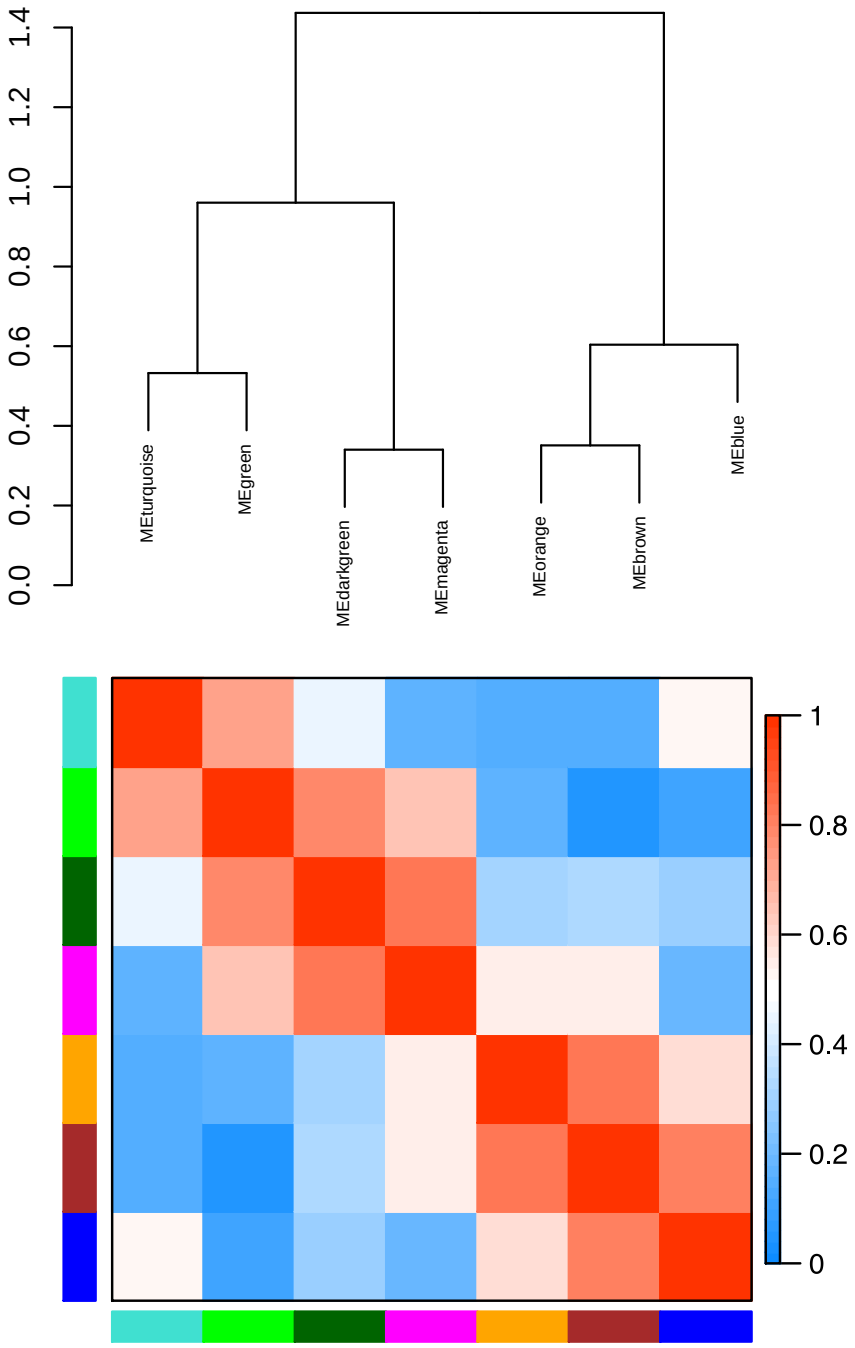


Fig. 5.12 **Module eigengene similarities and correlations for the O'Malley data set.** The top panel shows a hierarchical clustering dendrogram of the eigengenes based on dissimilarity. The bottom panel shows the eigengene network, which is defined by a signed weighted correlation network varying from 0 (no correlation) to 1 (highest possible correlation) as displayed in the color legend.

Relating Module Eigengenes to Cellular Traits

The profile of the defined module eigengenes had differential correlation with the different cellular traits analysed as illustrated in Figure 5.13. The values displayed in each cell in the heatmap refer to the pairwise correlation coefficients for each module eigengene-cellular trait which is color coded as pallet from red (positive correlation) to blue (negative correlation) as per figure legend. The value between parentheses correspond to the significance of such correlation (student asymptotic p-values). Based on these results we can see that the turquoise module represent well the gene co-expression network almost exclusive to MEF somatic cells (R=0.89, p = 2e-9) while the green, dark green, magenta and orange modules correspond the gradual transition towards pluripotency. The brown and blue modules on the other hand represent the final stage of pluripotency consolidation with the respective correlation coefficients of 0.65 and 0.56 for Nanog GFP +ve cells. These results are very similar to what was observed with the data set from (Chantzoura et al., 2015) analysed in the previous chapter (Figure: 4.13).

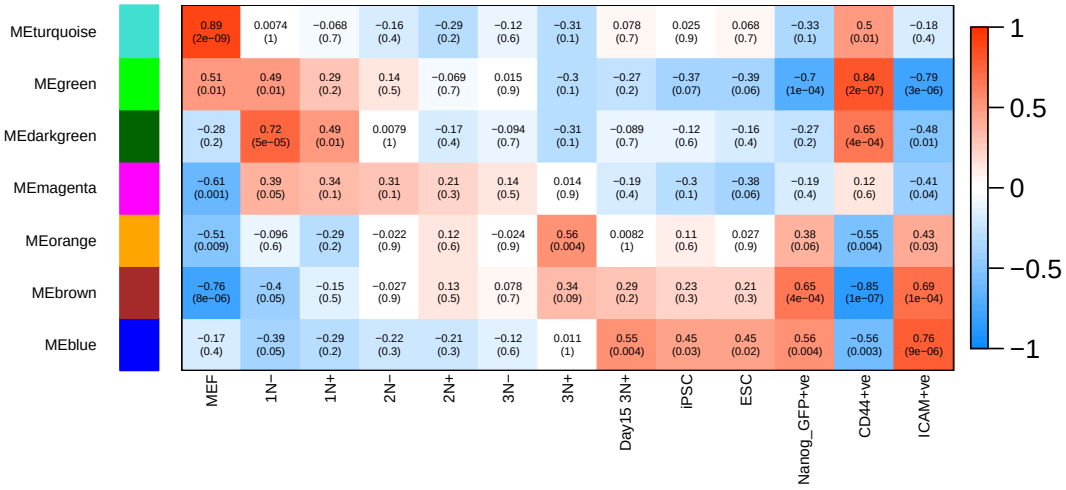


Fig. 5.13 **O'Malley data set module eigengene associations with cell types and cell markers:** This heatmap represents a high level representation of the identified network modules and the highly correlated modules to relevant cellular traits provide an efficient mean of prioritisation of modules for further analysis. The module eigengene values were correlated with each cell type analysed as well as the markers CD44, ICAM and Nanog-GFP. The color coded heatmaps represent the level of correlation between the module eigengenes (rows) and cell type or feature (columns). As shown by the color legend guides on the right, red represent strong positive correlation, white stands for no correlation and blue represent strong negative correlation. Each correlation coefficient value is printed in the corresponding table cell along with the association significance level between brackets. We can see that the turquoise and blue module are very significantly highly correlated with MEF cells (R=0.89, p = 2e-9) and Nanog-GFP +ve cells (R=0.56, p = 0.004).

5.2.4 Network Based Gene Set Enrichment Analysis

The figure panels 5.15 to 5.21 illustrate the process used in the characterisation of each of the defined modules, selection of hub genes and the use of those as input for gene set enrichment analysis.

The sub-figure a (Figures 5.15a to 5.21a) at the top of each panel is formed by a a gene expression heat map with all genes from each module and an eigengene expression bar plot which summarises the expression of all the genes along the different analysed cell types. The number of genes is shown between parentheses next to the module name. The selection of the hub genes was done after the visualisation of the correlation of all gene significance values with their corresponding module membership values (Figures 5.15b to 5.21b). The trait chosen for gene significance analysis was the one with the highest association with that module as viewed in Figure: 5.13. Arbitrary thresholds for module membership and gene significance are chosen appropriately in order to shortlist the potential hub genes with the highest values for both measures. These short listed hub genes were then used as input for gene set enrichment analysis using the online tool Enrichr (Chen et al., 2013a) and the results are summarised as bar plots in sub-figures 5.15c to 5.21c. Supplementary spreadsheets for the shortlisted hub genes for each module containing their gene symbols and module membership values can be found in the attached media (geneInfo_turquoise_om.csv to geneInfo_blue_om). These files can be directly uploaded into the Enrichr tool (http://amp.pharm.mssm.edu/Enrichr/) to reproduce the gene set enrichment

analysis displayed here. The sub-figures 5.15c to 5.21c shows the Enrichr results for Gene Ontology - Biological Processes 2017.

Figure 5.14 summarises the results of all Figures 5.15 to 5.21. We can verify that the associated Biological Processes enriched in the network modules produced for this data set are very similar to the ones obtained with the data set in the previous chapter. The modules associated with the initial phases of reprogramming (turquoise and green) are most enriched for genes associated with extracellular matrix organisation whose expression goes down as reprogramming starts. The modules associated with intermediate stages of reprogramming (dark green and magenta) are enriched for genes involved in skin development as previously observed with the (Chantzoura et al., 2015) data set. The orange module has transient expression associated with later stage activation of Nanog expression and its genes are enriched for genes involved in gene expression regulation and post-transcriptional processing. The brown and blue modules are associated with the establishment and maintenance of pluripotency at late reprogramming and is mostly enriched for functions such as stem cell maintenance, cell division regulation, cell fate specification and germ cell development.

These results are shown to be highly reproducible when compared to the the ones obtained with the previous data set (Chantzoura et al., 2015) and analysed in the previous chapter. One of the biggest advantages that WGCNA offers in comparison to standard differential gene expression (DGE) is that it preserves the gene expression profile structure of correlated genes instead of focusing only in

fold change and significance levels as traditionaly done. Standard differential gene expression relies on hard thresholds (fold change and significance) which leads to a certain loss of information which does not occur in WGCNA which produces continuous values leading to an increase in resolution as a consequence of the soft threshold regime employed. The following sections in this chapter analyses the results produced by WGCNA and DGE and utilise the fundamental network concepts unique to network analysis to further evaluate the reproducibility of module assigned between the two reprogramming data sets analysed so far.

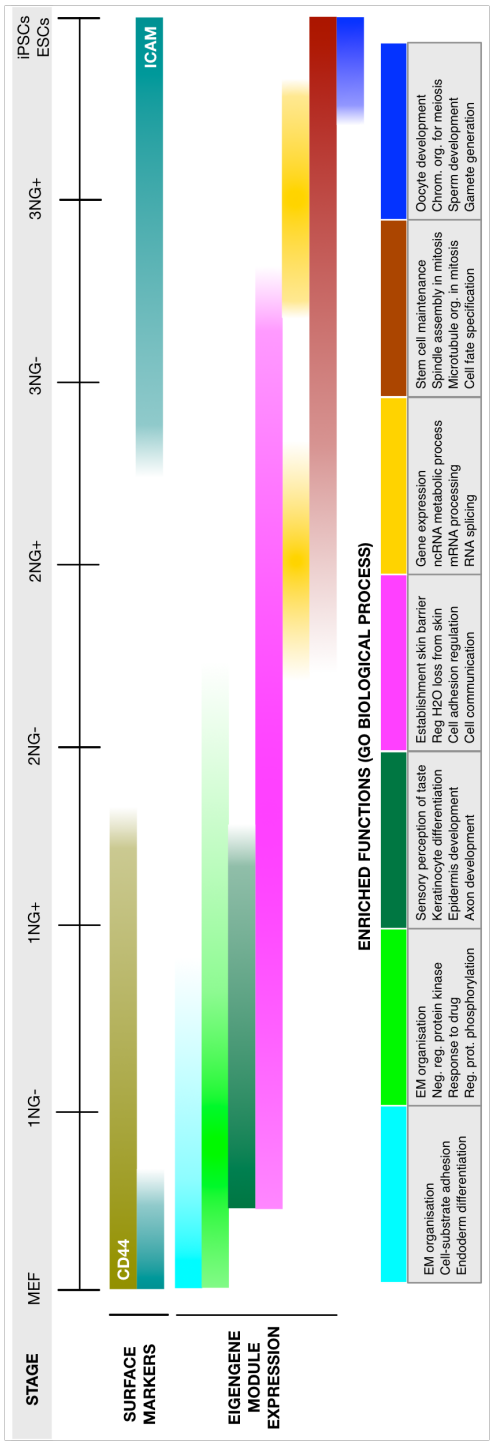
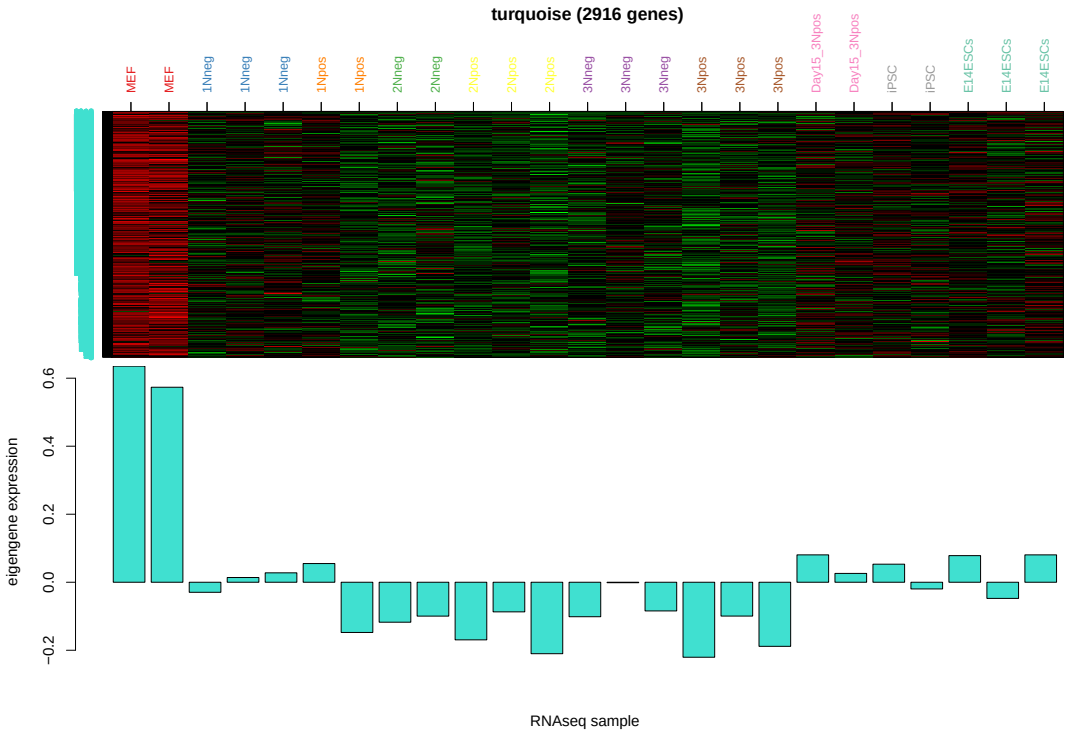
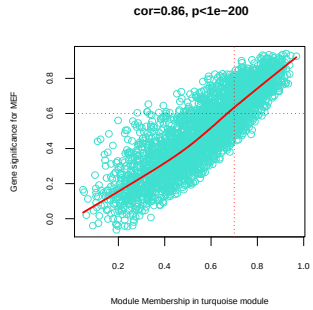


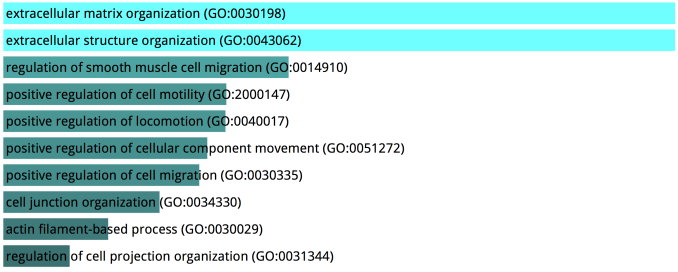
Fig. 5.14 **Dynamics of Biological Processes in iPS cells reprogramming:** The stages of reprogramming are labelled in the grey bar on the top of the figure. The markers used for the cell sorting and the network eigengenes expressions are represented as horizontal color bars with the color intensity correlating with its expression levels. The most significantly enriched functions are summarised on the color coded boxes on the bottom of the figure.



(a) Turquoise module eigengene expression

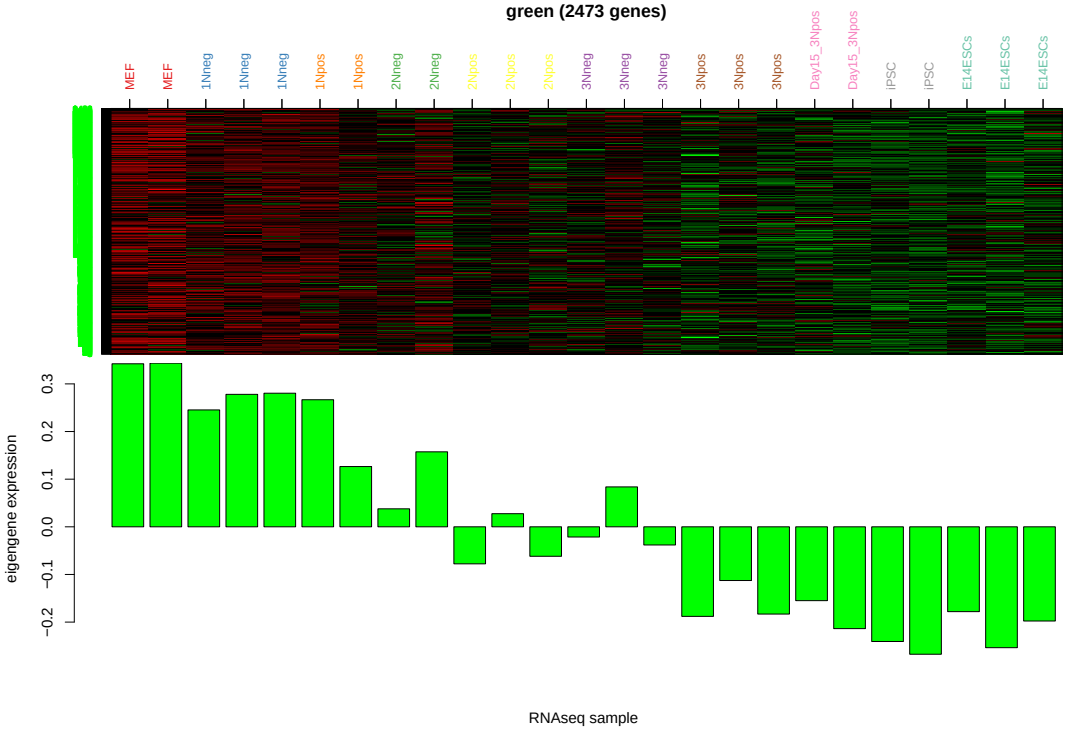


(b) Cutoff thresholds

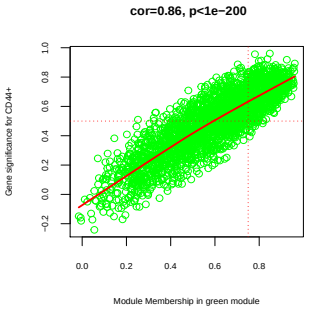


(c) Turquoise module Gene Ontology

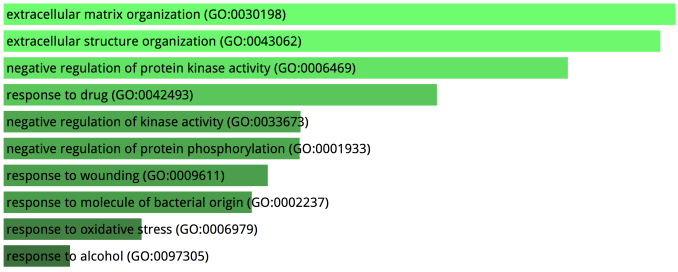
Fig. 5.15 Turquoise module characterisation



(a) Green module eigengene expression

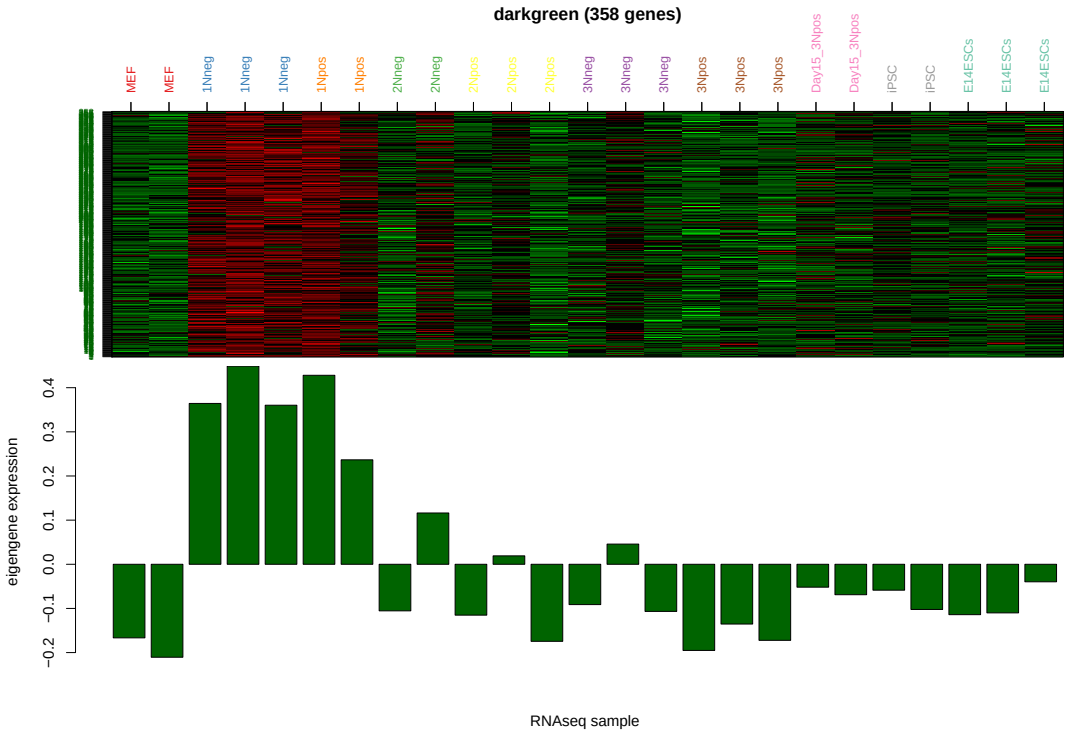


(b) Cutoff thresholds

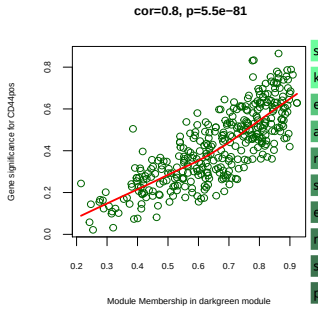


(c) Module Gene Ontology

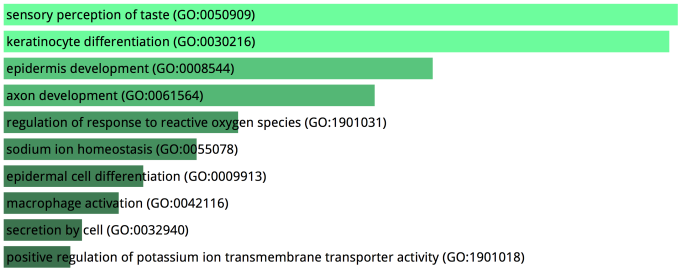
Fig. 5.16 Green module characterisation



(a) Dark green module eigengene expression

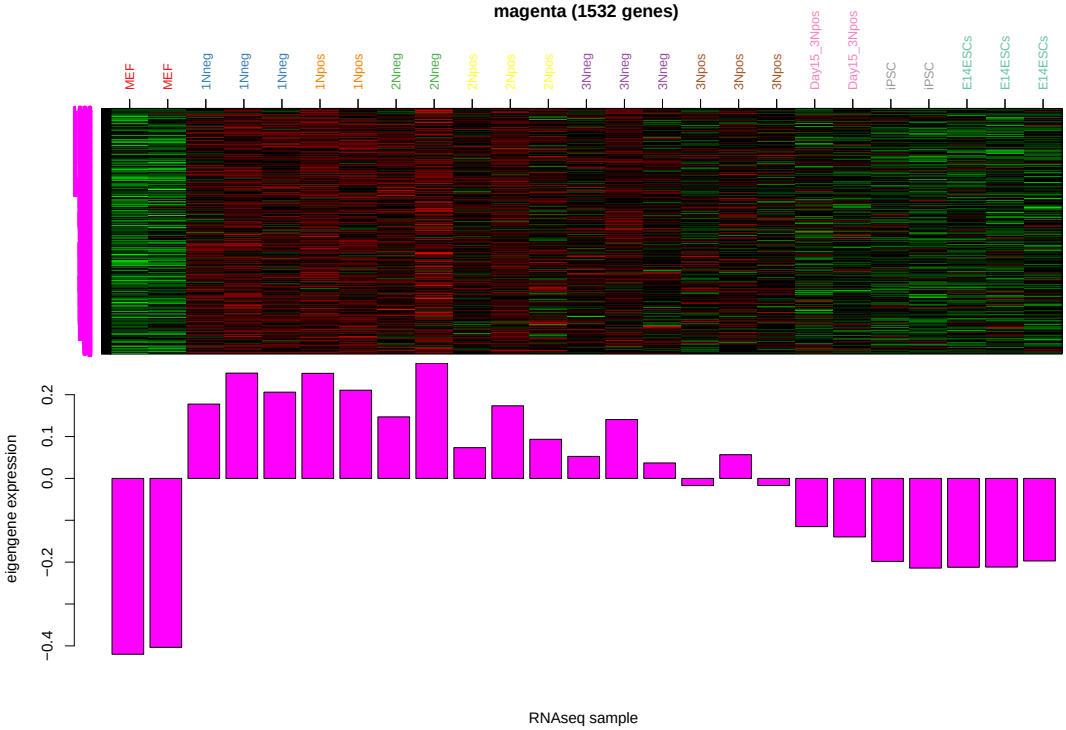


(b) Cutoff thresholds

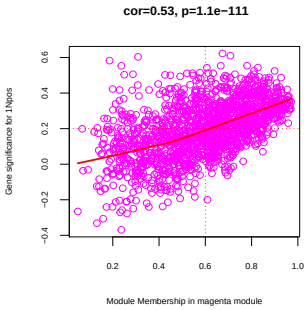


(c) Module Gene Ontology

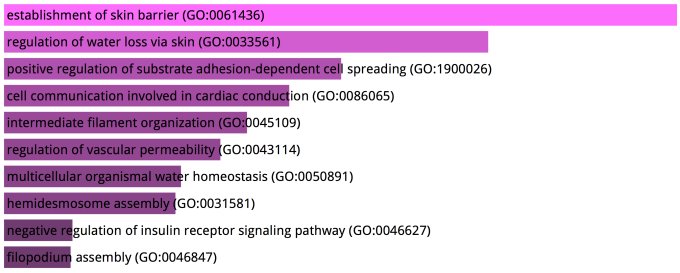
Fig. 5.17 Green module characterisation



(a) Magenta module eigengene expression

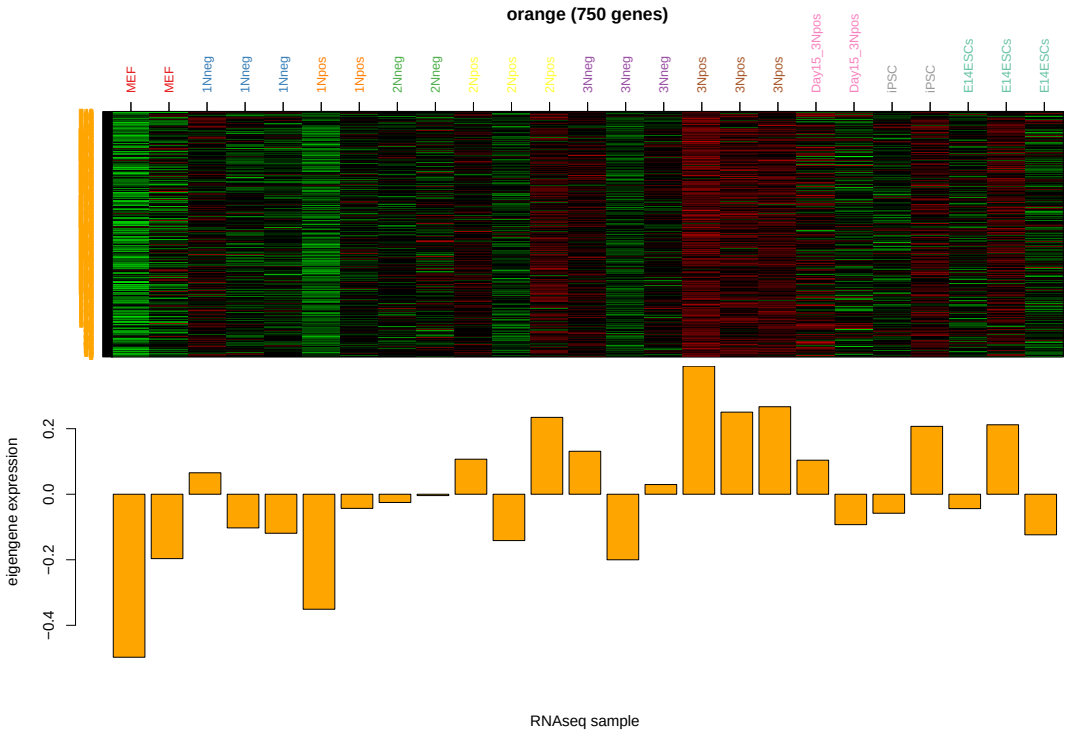


(b) Cutoff thresholds

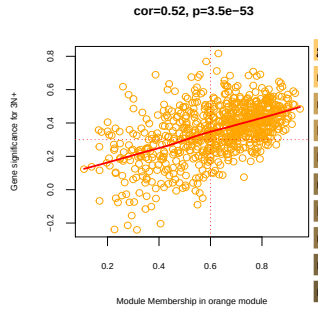


(c) Magenta module Gene Ontology

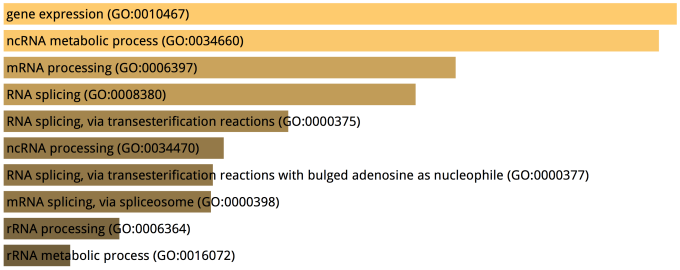
Fig. 5.18 Magenta module characterisation



(a) Orange module eigengene expression

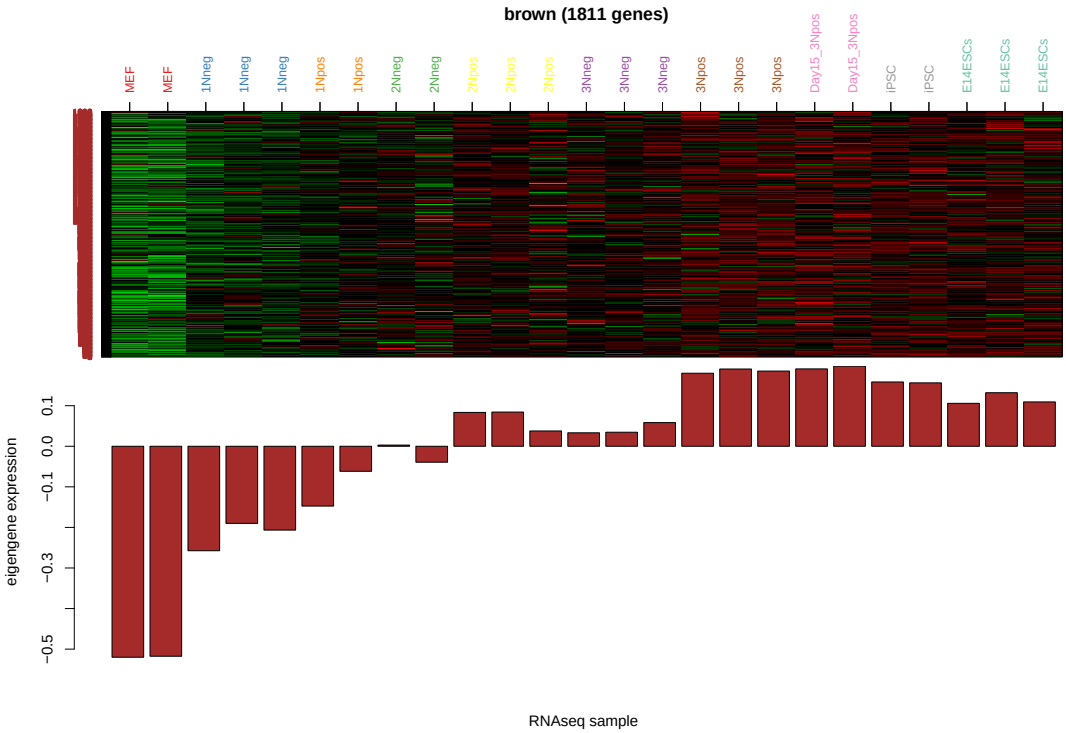


(b) Cutoff thresholds

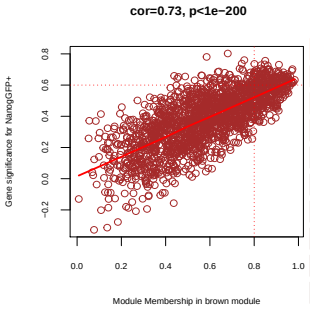


(c) Module Gene Ontology

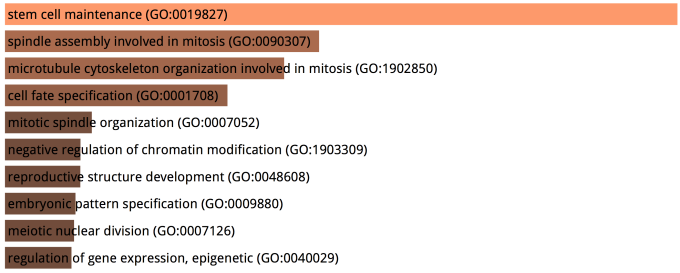
Fig. 5.19 Orange module characterisation



(a) Brown module eigengene expression

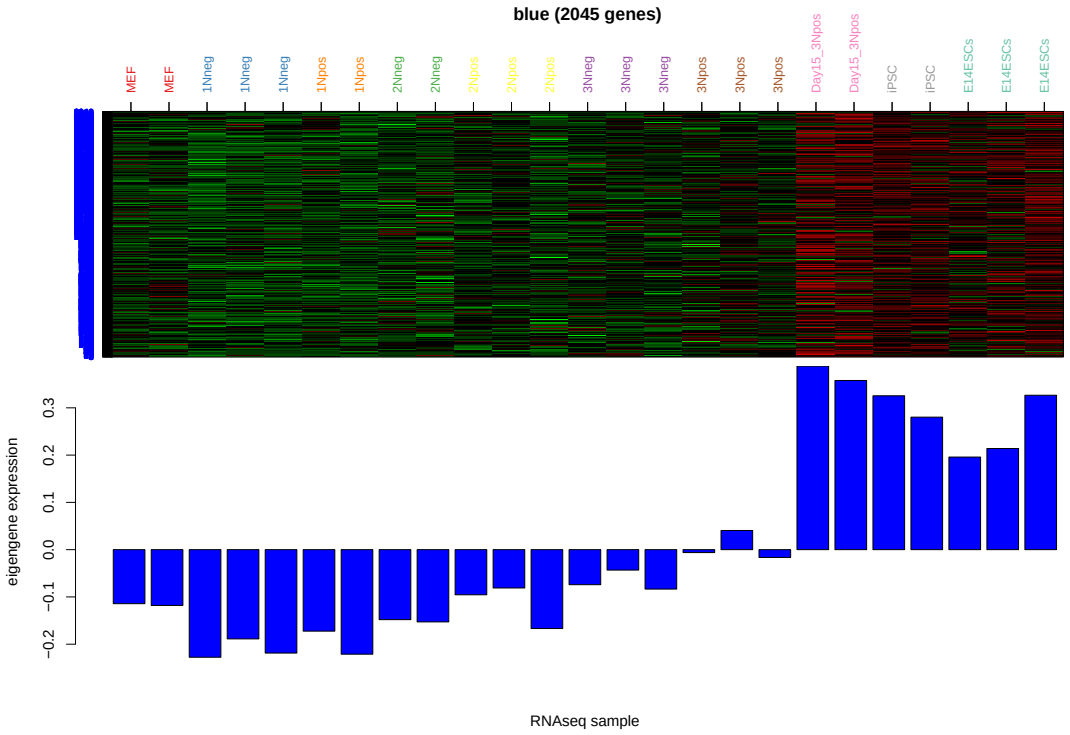


(b) Cutoff thresholds

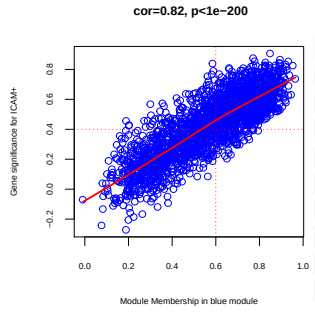


(c) Brown module Gene Ontology

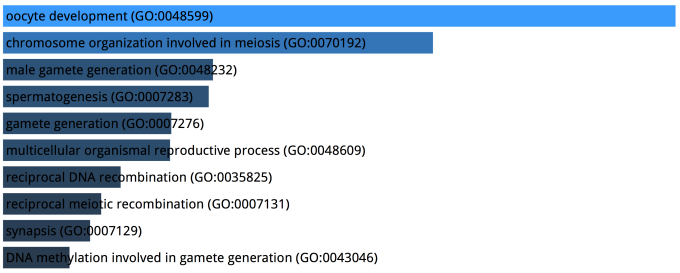
Fig. 5.20 Brown module characterisation



(a) Blue module eigengene expression



(b) Cutoff thresholds



(c) Blue module Gene Ontology

Fig. 5.21 Blue module characterisation

5.2.5 Increased Resolution of Network Based Analysis

Some of the same biological processes uncovered with the current WGCNA approach were also observed in the original analyses of Chantzoura et al. (2015) and O'Malley et al. (2013b) data sets which focused on DGE 5.1. However, the relationship among each gene is inevitably lost when only gene expression change differences are considered. The Figure 5.22 illustrates a contingency table for the cross tabulation of all the network genes and modules used for WGCNA in the current work and the differentially expressed genes (DEG) and their assigned clusters which were defined in the original work (O'Malley et al., 2013b). The values printed in each cell are the numbers of overlapping genes and their associated Fisher exact test p-values. To assist the visualisation, the cells are also coloured based on their significance levels (-log(p-value)) with a gradient of red tones as displayed by the legend. The WGCNA modules with their gene numbers are printed with their respective colours on the x axis ordered by their occurrence during reprogramming from MEF to iPS. The original DEG clusters are printed on the y axis. The DEGs were identified using edgeR and DESeq Bioconductor packages by selecting genes that had a log fold change of 1.5 or higher and FDR of 0.05 or lower with both packages and the clusters were classified by K-means clustering. The grey WGCNA module and DEG clusters are composed of genes not associated with any module or cluster respectively. We can observe that most DEG clusters have the majority of their genes significantly assigned to at least two WGCNA modules indicating that the latter approach was able to further separate

the DEG clusters with higher resolution. Moreover the number of unassigned genes in the DGE approach is far greater than for WGCNA with nearly 8 times more genes being ignored (10836 and 1416 respectively).

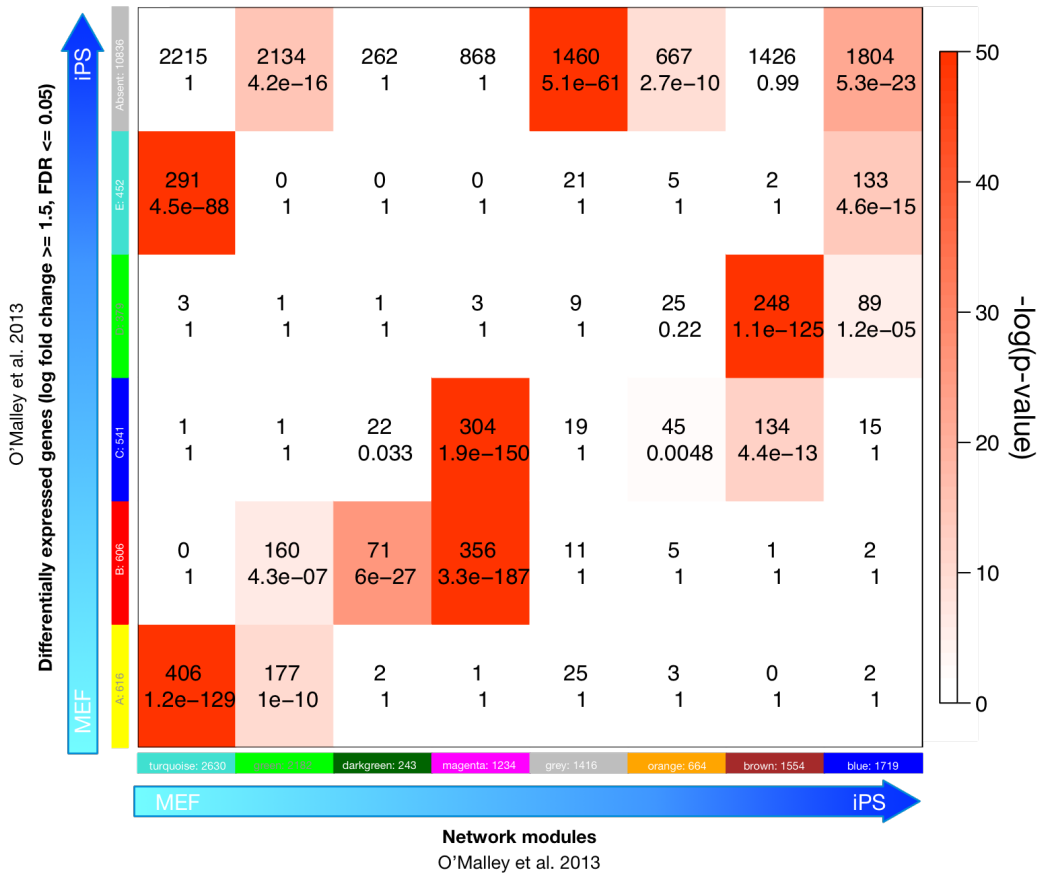


Fig. 5.22 **Contingency table DEGs vs Network modules:** Cross Tabulation of all the network genes and modules used for WGCNA in the current work and differentially expressed genes (DEGs) and their assigned clusters which were defined in the original work (O'Malley et al., 2013b). Each row and column are labelled by the corresponding module/cluster color and the total number of genes in the module/cluster. In the table, numbers give counts of genes overlapping row cluster and column module. The table is color-coded by log(p), the Fisher's exact test p value, according to the color legend on the right.

5.2.6 Network Module Preservation in Reprogramming

The comparison of networks is a powerful approach which allows the identification of differences and similarities in gene regulation between conditions and between data sets, regardless of whether these were produced with different platforms such as micro-arrays and RNA-Seq or even from different species. In this section I compare the networks from two data sets created to study the mechanisms of reprogramming somatic cells to pluripotency and the identification of network modules preserved in both studies gives us a higher confidence on the reproducibility of these results.

We have already seen a good agreement in module eigengene association with similar stages of reprogramming as well as the results of gene set enrichment analysis which produced enrichment to similar GO biological functions. This agreement indicates a good overlap of genes among supposedly equivalent network modules in the two data sets. This overlap can indeed be confirmed via standard cross-tabulation based statistics as illustrated in Figure 5.23 . This figure shows that there are highly significant overlaps of genes in most modules as inferred by Fisher exact test. Values within each cell correspond to the number of overlapping genes and their significance as hypergeometric p-values in the pairwise comparisons between modules.

Despite this remarkable agreement between data sets this approach using contingency tables has the disadvantage that it is ill suited for making the strong statement that a given module is not preserved. It may be argued that the module

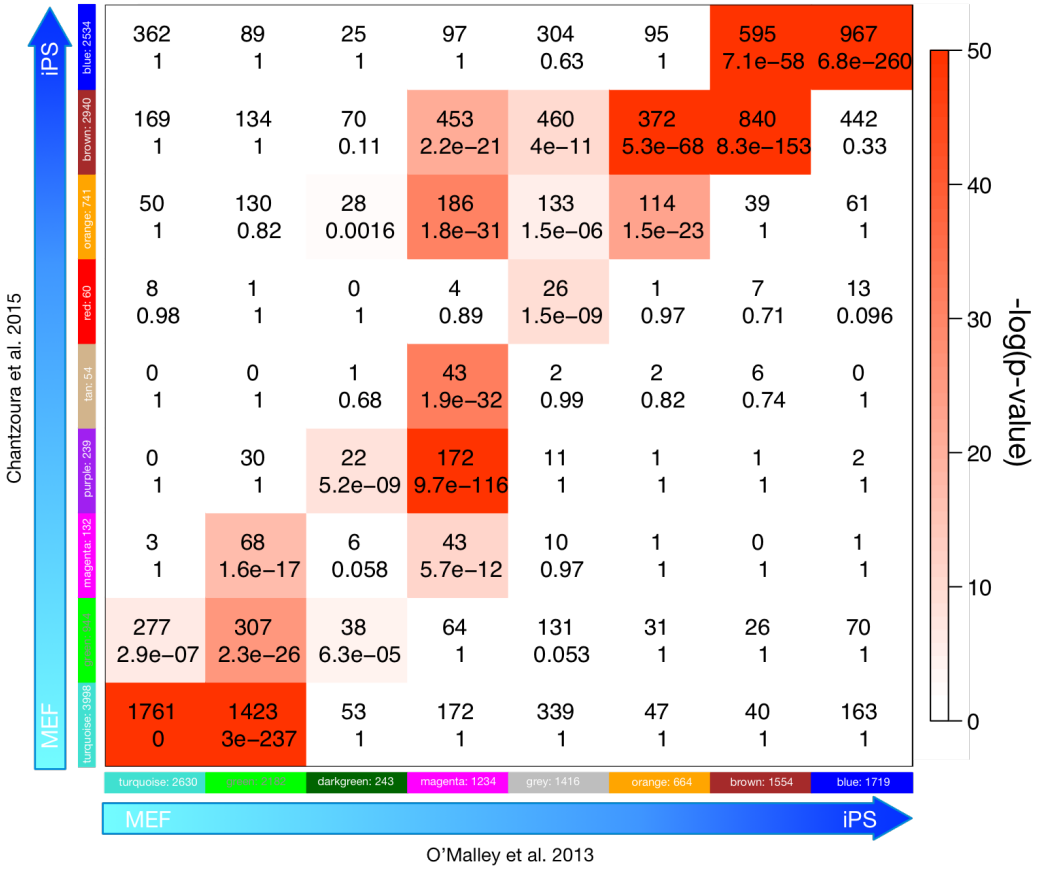
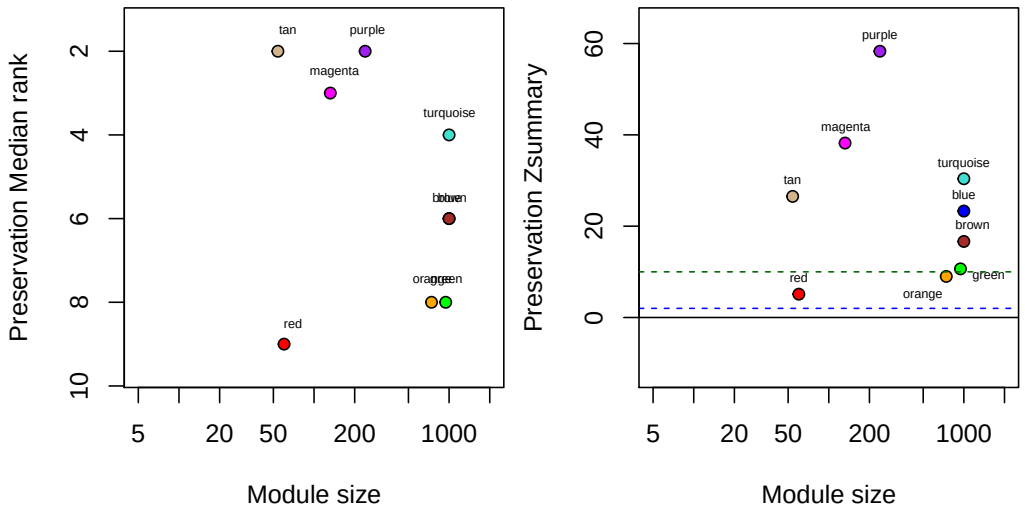


Fig. 5.23 **Chantzoura and O'Malley module genes cross tabulation:** Cross Tabulation based approach for module preservation analysis between Chantzoura (rows) and O'Malley (columns) network modules. Each row and column are labelled by the corresponding module color and the total number of genes in the module. In the table, numbers give counts of genes in the intersection of the corresponding row and column module. The table is color-coded by log(p), the Fisher's exact test p value, according to the color legend on the right.

memberships are dependent on the method used for cutting hierarchical clusters and module preservation could possibly be achieved by appropriate cluster classification tweaks. The analysis of module preservation based on inherent network properties such as gene connectivities and module densities offers a more robust method to more confidently quantify the reproducibility of module assignments.

The WGCNA package provides an algorithm implementation which takes into consideration over 20 network statistics which are then summarised into two main complementary composite module preservation statistics, Zsummary and MedianRank measures taking as input the module assignment in the reference data set and the adjacency matrices from the reference and test data sets. The Zsummary is obtained by a computationally intensive permutation test of each network statistics and then aggregated. The significance threshold of such composite Z score is evaluated as values below 2 (Zsummary < 2) indicate no preservation, values between 2 and 10 (2 < Zsummary < 10) indicate weak to moderate evidence for module preservation and values above 10 (Zsummary > 10) indicate strong evidence of module preservation. One undesirable property of the Zsummary statistic is that it depends on the module size (number of genes). Therefore the MedianRank module preservation contributes to the analysis as it ranks the module preservation independently of the module sizes providing a relative module preservation measure. However it does not provide a significance threshold. Figure 5.24 illustrates such an analysis when the Chantzoura data set was used as a reference and the O'Malley data set the test. Based on this analysis we can verify

that the red and orange modules are moderately preserved and all other modules are highly preserved (Figure: 5.24b) .



(a) Median Rank module preservation (b) Zsummary module preservation

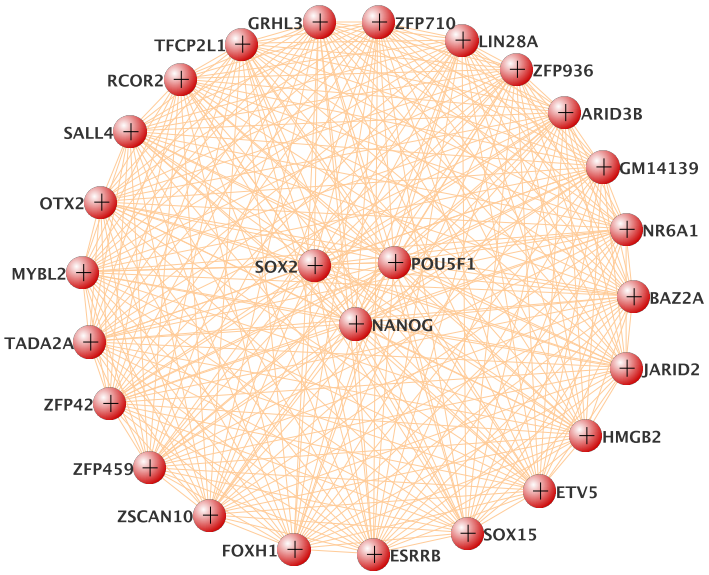
Fig. 5.24 Chantzoura and O'Malley data sets module preservation statistics

Interestingly the top three modules with the highest MedianRank values are the tan, purple and magenta modules, all associated with transient intermediate stages of reprogramming (Figure: 5.24a). The highly preserved magenta, purple and tan modules from Chantzoura networks overlaps mostly with the magenta module in the O'Malley network. As previously described, the hub genes from these modules are enriched for genes associated with gene expression regulation in epidermis development (Purple), immune response activation and lipid metabolism (Magenta) and biosynthetic metabolism (Tan). No direct association of the genes in magenta, purple and tan modules have been reported so far. It would be interesting to assess if these network modules are also preserved in other

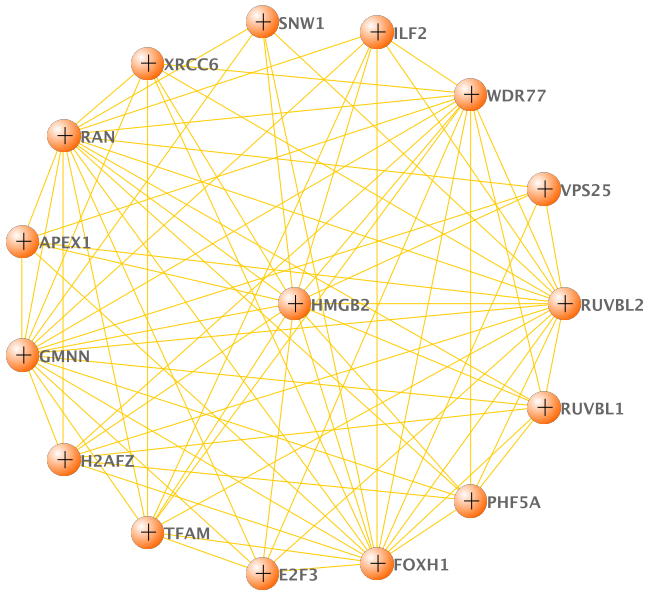
reprogramming data sets or these are more associated to the experimental setup from O'Malley and Chantzoura experiments which are very similar to each other.

The analysis of the gene co-expression network from the pluripotency (brown) module reveals the gene interactions of the well known core pluripotency regulatory network with *Oct4* (aka *Pou5f1*), *Sox2* and *Nanog* being some of the most connected genes (Figure: 5.25a). Upon inspection of the network structures of intermediate reprogramming stages and the established pluripotent brown modules I observed a potentially interesting network within the orange module. This module is highly associated with RNA splicing machinery genes however if I subset the hub genes from this module which are either genes for transcription factors, transcription co-factors, chromatin associated proteins or chromatin remodellers it reveals a sub-network having the high-mobility-group protein B2 (*Hmgb2*) gene as the most connected gene which is highly associated *H2af.z* and the genes for two chromatin remodelling proteins *Ruvbl1* and *Ruvbl2* (Figure: 5.25b), which are thought to be associated with H2A.Z chromatin deposition machinery and histone acetyltransferase complex (Choi et al., 2009; Mizuguchi et al., 2004; Subramanian et al., 2013a). *Hmgb2* is a top hub gene in both the pluripotency (brown) module and the chromatin remodelling (orange) module and have been recently reported to be an important pluripotency regulator via association with Oct4 (Campbell and Rudnicki, 2013; Zhao et al., 2017). Therefore the combination of *Hmgb2*, *Ruvbl1*, *Ruvbl2*, *H2af.z* and other genes from the orange module may further improve the efficiency of reprogramming mediated by *H2af.z* over-expression

together with Yamanaka factors as observed in Chapter 3 and elsewhere (Wang et al., 2015).



(a) Pluripotency Modules (Brown)



(b) Chromatin Remodelling Module (Orange)

Fig. 5.25 Network structures of the pluripotency and intermediate reprogramming chromatin remodelling modules with the most preserved hub genes

5.2.7 Conclusion

The so called Yamanaka factors triggers a highly stochastic process during iPS cell generation, gradually accumulating characteristics which confer higher probability of successful reprogramming eventually progressing to a more deterministic phase accompanied by two waves of pluripotency gene expression. (Buganim et al., 2012; O'Malley et al., 2013b; Polo et al., 2012; Samavarchi-Tehrani et al., 2010). The reanalysis of the data sets from O'Malley et al. (2013b) and Chantzoura et al. (2015) by WGCNA and the evaluation of their gene co-expression network structure preservation permitted me to not only identify the gene expression signatures of the various intermediate reprogramming cell types but also provides a way to increase its resolution and a way of prioritising genes based on their co-expression interactions with other genes withing a network. Intramodular hub genes with the highest connectivities have higher probabilities of having key roles on the biological processes associated with such network and the selection of candidate reprogramming factors based on this properties provides an attractive means of increasing the chances finding molecules able to improve the efficiency of reprogramming.

The results presented in this and the previous chapters identified a sub-network of a module associated with potentially relevant chromatin remodelling functions. A core hub gene in this sub-network is *Hmgb2* which is also a top ranked hub gene in established a pluripotency network. *H2af.z* is another member of this sub-network, along with *Ruvbl1* and *Ruvbl2* which are chromatin remodelling

proteins with known association to H2A.Z chromatin deposition (Choi et al., 2009; Mizuguchi et al., 2004; Subramanian et al., 2013b). I have demonstrated in Chapter 3 that *H2af.z* over-expression is able to improve iPS cells generation efficiency when co-expressed with conventional Yamanaka factors. Given the network position of *H2af.z* in this chromatin remodelling sub-network, it is possible that its reprogramming boosting effect might be further enhanced by the forced expression of other network partner components such as *Ruvbl1*, *Ruvbl2*, *Hmgb2* and *Vps25* speeding up the transition to a fully fledged pluripotency network. The following chapter addresses the assessment of reprogramming module preservation occurring during early development from oocyte to morula, the natural route to pluripotency.

Chapter 6

Gene Expression Networks in Pre-Implantation Development and their Similarities to Reprogramming Networks

6.1 Introduction

Pre-implantation development provides insight into efficient genome reprogramming of differentiated cells (Figure 6.1). Following fertilisation the maternal factors deposited in the oocyte are known to be responsible for epigenetically resetting the paternal sperm DNA and chromatin across the zygote to prepare for whole-genome activation, also known as zygotic genome activation (ZGA). This transcriptional burst has a small wave at late pronuclear (1C) stage and peaks at the two-cell (2C) stage in mouse (Xue et al., 2013).

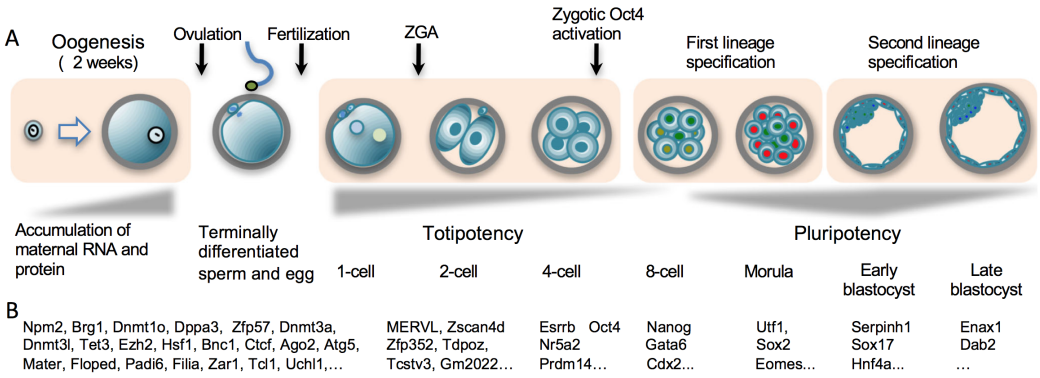


Fig. 6.1 **Mouse pre-implantation development:** (**A**) Mature oocytes are fertilised by sperm producing totipotent zygotes which gradually lose their totipotency with each cell division with blastomeres becoming gradually more specialised. The blastocyst is formed at embryonic day 3.5 and finally implants in the uterus at embryonic day 4.5. The stored maternal factors of the oocyte trigger zygotic genome activation (ZGA) at the 2 cell stage when there is a massive degradation of maternal factors and activation of a new embryonic gene-expression profile, followed by progressive activation of lineage specific genetic programs. (**B**) representative key genes expressed in each pre-implantation stage. Adapted from Wu et al. (2017)

The ZGA at 2C stage embryos is marked by the activation of transposable elements, such as the endogenous retrovirus MERVL, and a number of 2C exclusive

transcripts as those from *Zscan4* gene family, essential for the transition from 2C to 4C stages and subsequent embryo development (Falco et al., 2007). Interestingly it has been recently observed that totipotent 2C-like cells can transiently arise spontaneously in mouse ES cells cultures at very low frequency of less than 0.5% (Macfarlan et al., 2012) and the expression of 2C specific genes is negatively regulated by KDM1A also known as LSD1 (Macfarlan et al., 2011).

The zygote and individual blastomeres from 2C stage are fully totipotent at a single cell level, that is they are able to develop into a complete organism and also the extra-embryonic tissues. The genome activation of *Oct4* and *Cdx2* and their mutually antagonist expression marks the first lineage specification in the mouse (Niwa et al., 2005; Ralston and Rossant, 2005; Strumpf et al., 2005). While the blastomeres expressing *Cdx2* become committed to form the extra-embryonic trophectoderm, the zygotic activation of *Oct4* marks the beginning of the pluripotency network circuitry formation followed by the activation of a number of pluripotency master regulators, such as *Nanog* and *Sox2* at consecutive embryonic stages until pluripotent cells become confined to the inner cell mass (ICM) of the blastocyst from which ES cells can be readily derived (Nichols and Smith, 2012).

The pre-implantation development from zygote to the blastocyst stage, when pluripotency is fully established, occurs in a timely step-wise fashion (Figure 6.2) and the recent development of single cell RNA-Seq has brought us an opportunity to study the genetic circuits operating in such an important phase of early life at an unprecedented resolution. In this chapter I describe the re-analysis of such a data

set (Xue et al., 2013) with the purpose of exploring the possibility of finding shared gene expression networks between reprogramming and the natural embryonal route to pluripotency. The identification of shared network hub genes between these two alternative paths to pluripotency may allow us to better understand the mechanisms involved in such events and could possibly lead to more efficient reprogramming recipes.

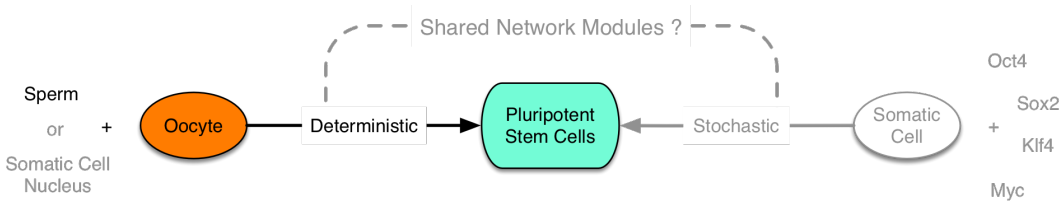


Fig. 6.2 **Natural route to pluripotency:** The asssembly of genetic networks towards pluripotency is deterministic during pre-implantation development.

6.2 Results and Discussion

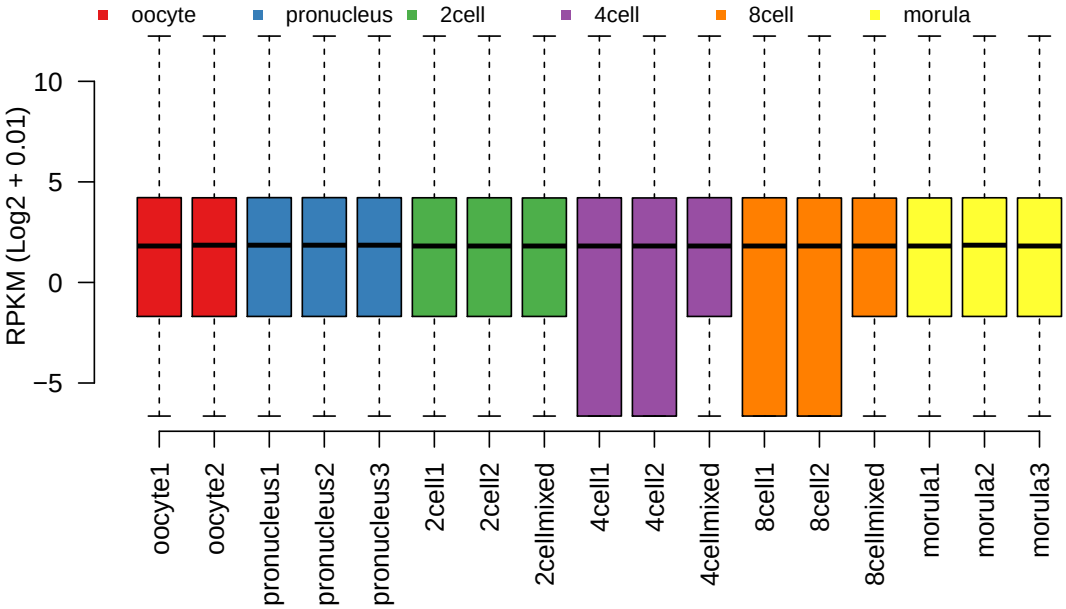
6.2.1 Data Set Manipulation

The publicly available data set analysed in this chapter was produced by Xue et al. (2013) and was downloaded as a pre-processed file (File can be found in the attached media as Xue/GSE44183_mouse_expression_mat.txt) from the NCBI Gene Expression Omnibus repository (Assession GSE44183) and converted to a Bioconductor *SummarizedExperiment* object to assist the analysis in R programming environment. The samples from this data set are from single cells from consecu-

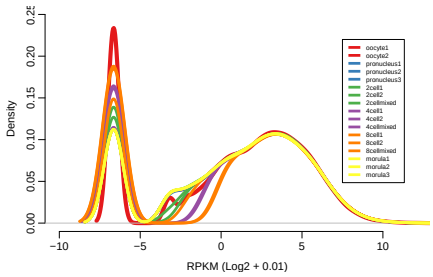
tive stages of pre-implantation mouse development (oocytes up to 8 cell stage) or single morula embryos. The two sample replicates for 2, 4 and 8 cell stages were also combined to produce a mixed 2, 4 and 8 cell stage third replicate to improve sequencing sensitivity. The pre-processed data set was produced by mapping the RNA-Seq raw reads to the mm9 mouse genome using default parameters in BWA and then transformed using the reads per kilobase per million (RPKM) metric. Genes with low expression in all stages (average RPKM < 0.1) were filtered out, followed by quantile normalisation.

6.2.2 Exploratory Data Analysis

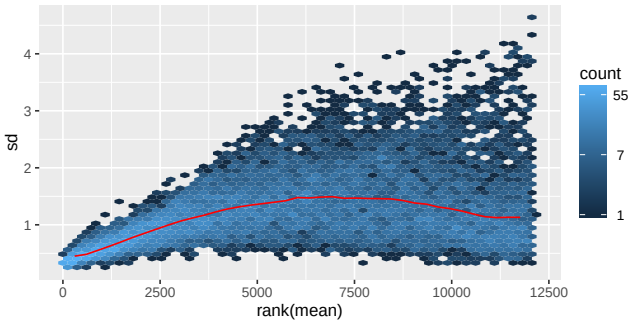
To determine the quality and properties of the data set produced by Xue et al. (2013) I performed exploratory data analysis by assessing the normalized data as well as the gene expression variance distributions across the range of expression levels. Figures 6.3a and 6.3b illustrate that the obtained data set was already well normalized with the single cell samples for 4 and 8 cell stages having some expression skew for genes with lower expression. Moreover there was no obvious dependence of the standard deviation on the mean expression (Figure 6.3c). The horizontal red line in Figure 6.3c highlights that the data set is approximately homoscedastic.



(a) Box plot



(b) Density plot



(c) Gene expression standard deviation

Fig. 6.3 **Xue data set mapped reads statistics:**(a) Box plot: Distribution of the RNA-seq data sets representing the expression levels (RPKM log transformed) for all 13879 genes in the data set. The sample groups are color coded as displayed on the legend above the plot; (b) Density plot: Log-density distribution of all libraries superposed and color coded as per legend insert; (c) Gene expression standard deviation vs. ranked mean: The color intensity of scatter plot reflects the expression level as per legend. The red line depicts the running median estimator (window-width 10%)

Unsupervised Hierarchical Clustering

In order to assess the sample relationships between groups as well as their replicates I performed unsupervised hierarchical clustering. Figure 6.4 indicates that the samples of this data set were of good quality with the different sample types

replicates forming coherently distinct clusters. The heatmap corresponds to the pairwise Pearson correlation between all samples based on the expression of all 13879 genes in the data set. We can see that the oocyte and pronuclear stage samples are very similar to each other, nevertheless they are clearly distinct reflecting the early minor wave ZGA described by Xue et al. (2013). The major ZGA marks the the formation of a 2-cell embryo in mouse development (Schultz, 1993; Xue et al., 2013) causing the 2C stage samples to cluster away from all other samples as it would be expected. Moreover the successive transition from 4 cell embryo to morula was also well captured by the hierarchical clustering in Figure 6.4.

Principal Component Analysis

Principal component analysis was performed by singular value decomposition of the scaled RPKM values of all genes in the data set. Figure 6.5 illustrates such analysis. The plot in Figure 6.5a displays the singular values for each principal component (eigengene), Figure 6.5b shows the percentage of variance explained per principal component and Figure 6.5c shows the cumulative variance explained by the consecutive combination of the principal components. The first three principal components are able to capture most of the variance in the data set, explaining around 60, 20 and 5 percent of the data set variance respectively with a cumulative variance explained of around 85 percent.

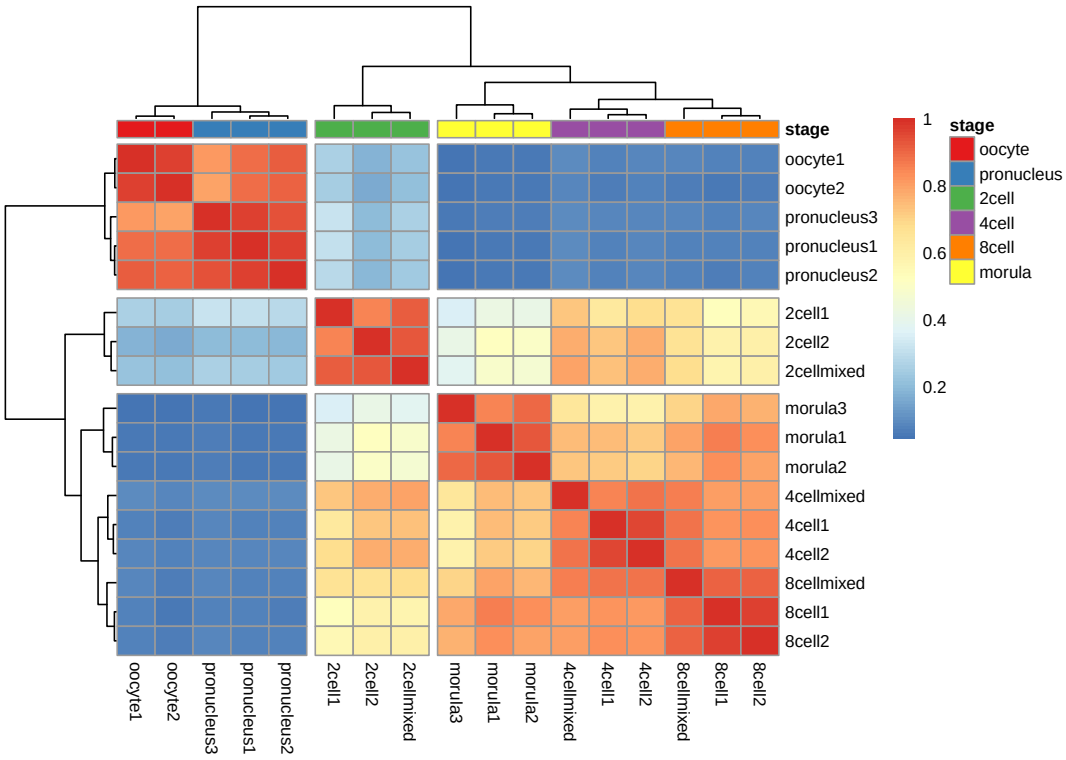
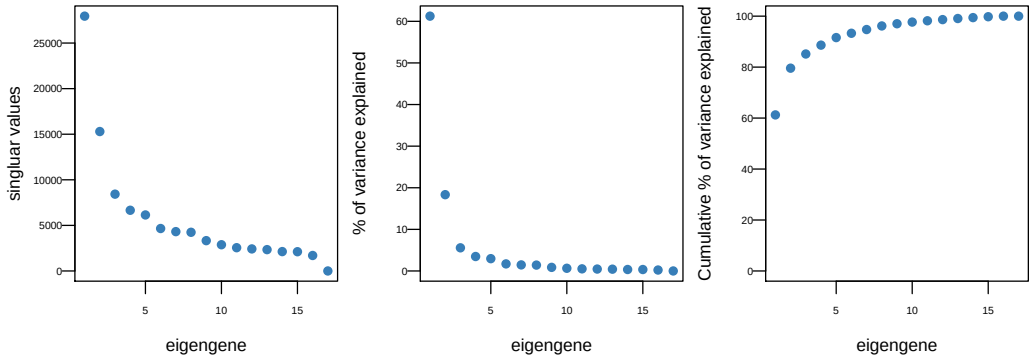


Fig. 6.4 **Xue data set sample clustering:** The heatmap represents the clustered pairwise Pearson correlations between all samples based on their log transformed RPKM expression of all 13879 genes analysed. The Pearson correlations range from 0 for no correlation (dark blue) to 1 for maximum correlation (red) as depicted in the legends next to the plot. The sample types were also color coded to assist in the sample clustering visualisation.

Figure 6.6 shows that the first 3 principal components are well suited for separating all the sample groups. The principal component analysis using the first 2 principal components clearly illustrated the appropriate ontological sequence of embryonal stages forming a well defined path from oocytes to morula. The ability to reproduce well what has been demonstrated in the original publication by Xue et al. (2013) gives support to the quality and robustness of the data set.



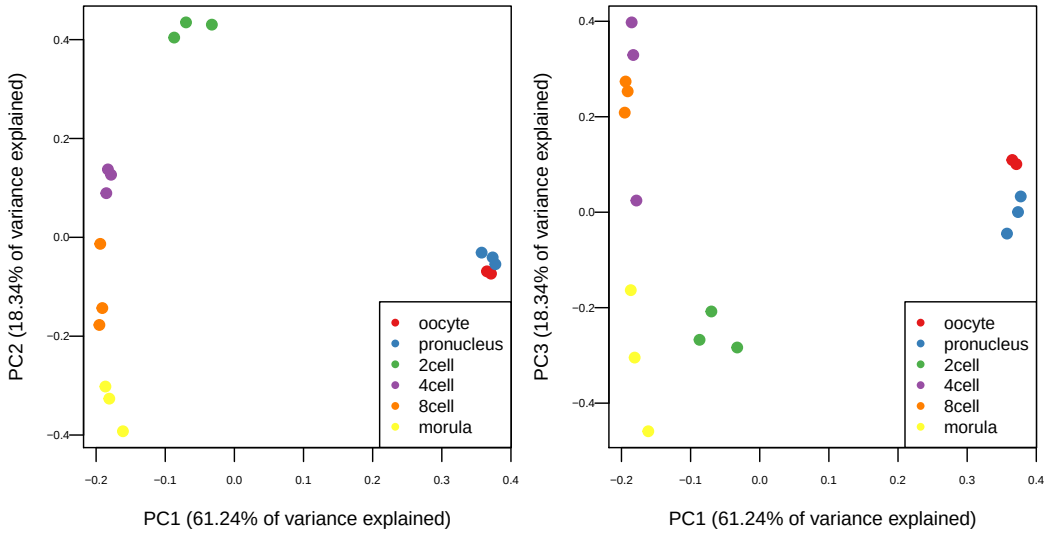
(a) Singular values (b) Variance explained (%) (c) Cummulative variance

Fig. 6.5 **Xue data set singular value decomposition:** Principal components (eigengenes) were calculated via singular value decomposition.

6.2.3 Weighted Gene Co-Expression Network Analysis

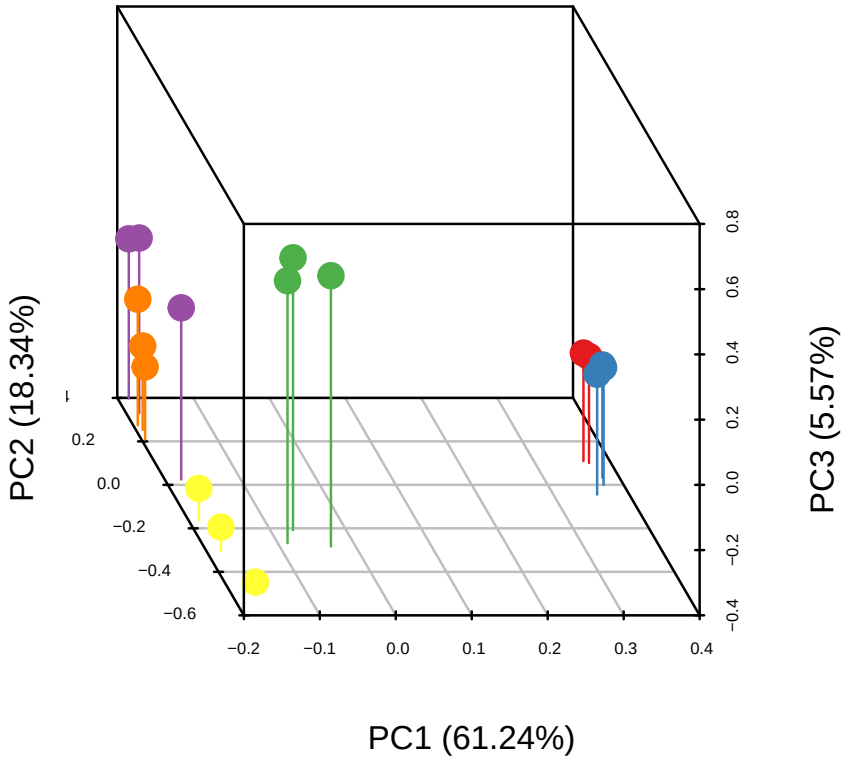
Scale Free Topology Fitting

As previously described in the two previous chapters an optimum power β was selected as a soft-threshold in order to build a weighted gene co-expression network which is able to approximately satisfy the scale-free topology criterion. Figure 6.7 illustrates how the network properties such as connectivity and scale-free fit index vary as a function of power β values ranging from 1 to 35.



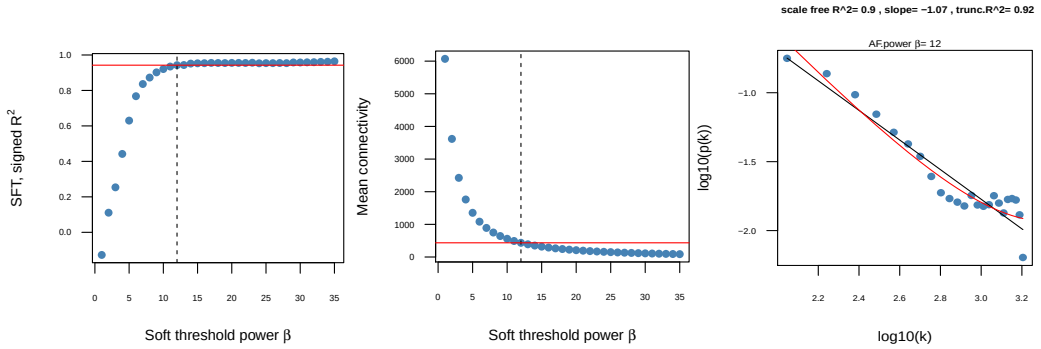
(a) PCA cell type (PC1/PC2)

(b) PCA cell type (PC1/PC3)



(c) 3D PCA cell type

Fig. 6.6 **Xue data set principal component analysis:** Most of the variance between samples from each embryonic stage is captured by the first 3 principal components creating a clear path reflecting the actual ontological order.



(a) Soft threshold selection (b) Mean connectivity (c) Scale free fitting

Fig. 6.7 **Selection of power β for the Xue data set:** The power β 12 produces a weighted gene co-expression network that satisfy the most the scale free topology criteria.

A spreadsheet (Xue/xue_power_threshold.csv) with the resulting network properties as a function of the tested power threshold range can be found in the attached media. The network properties tested were mean, median and maximum connectivities, network density, centralization and heterogeneity as well as the scale free topology indexes and slope. The WGCNA function *pickSoftThreshold* was used for such calculations by transforming the data set into a similarity matrix and raising its values by the soft-threshold values using the network type setting to signed network. We can see in Figure 6.7a that the power β value of 12, which is usually the recommended method default, produces a scale free fitting index of around 0.95 and this represent the point where scale free fitting index begins to saturate. The plot in Figure 6.7b shows the mean connectivity is a decreasing function of the power β and reaches a mean connectivity value of 435 when the soft-threshold is set to 12. Figure 6.7c illustrates that the network produced the power β set to 12 fits very well a scale-free topology producing a near perfect fit

when the log10 of the connectivities (log10(κ)) was plotted as a function of the log10 of the frequency of the connectivities (log10(p(κ))).

Construction of a Weighted Gene Co-Expression Network

Using the selected power β set to 12 I created a signed weighted gene co-expression network following the guidelines from the authors of the WGCNA method (Zhang and Horvath, 2005). I was able to produce a network topology very similar to that in the original publication (Xue et al., 2013). Figure 6.8 illustrates the resulting network module definition for the 13879 genes analysed. The second color band below the clustering tree represents the final module assignments after merging modules which eigengenes had similar expression profiles. The network modules produced remarkably well defined gene clusters with clear differential associations with the different embryonic stages analysed in this data set. We can verify that most modules have an almost exclusive positive association to one of the embryonic stages as illustrated by the red colours in the gene heatmap below the clustering tree in Figure 6.8. Some of the final modules still had a relatively high correlation to each other as illustrated in the correlation heatmap in Figure 6.9. A spreadsheet containing the module assignment (mergedColors) per gene can be found in the attached media (Xue/xue_module_assignment.csv).

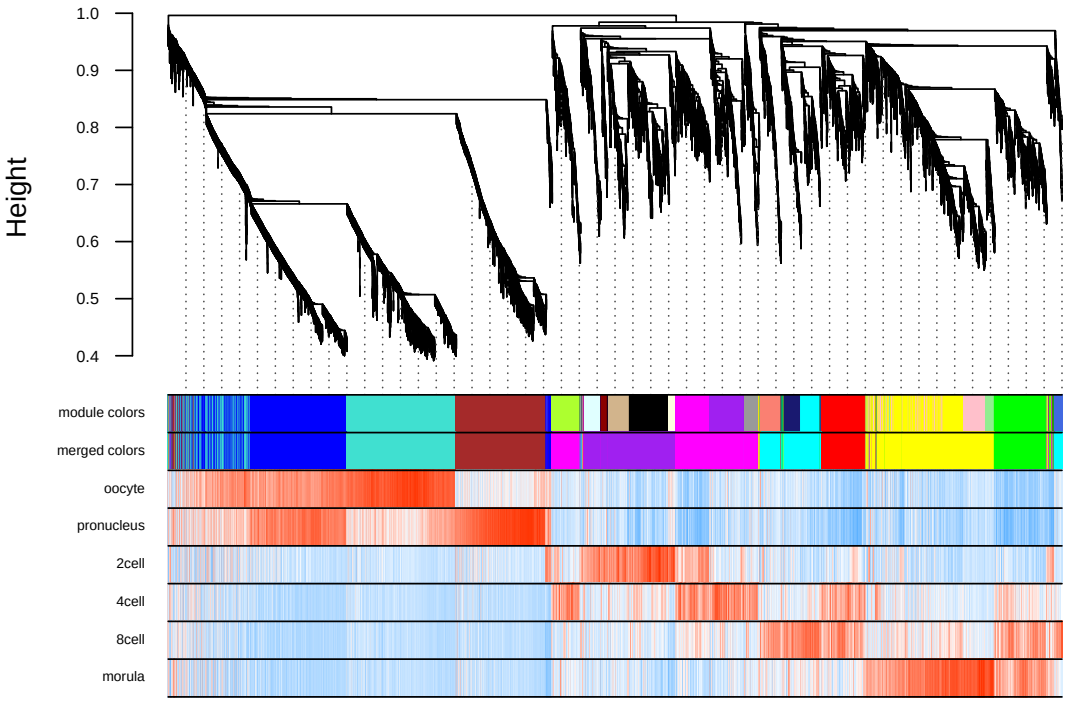


Fig. 6.8 **Network module definition for the Xue data set:** Gene dendrogram obtained from the single cell RNA-Seq dataset of Xue et al. (2013). This tree was defined by hierarchical clustering the dissimilarity based on consensus Topological Overlap measures using average linkage of the 13879 genes analysed. The color coded modules were assigned based on the dynamic hybrid branch cutting method represented as the first color band below the the dendrogram. The second color coded band represent the final module assignment after merging similar modules. The several other color bands represent the gene significance measures to all the cell samples analysed: "red" indicates a high positive correlation with the a cell type and "blue", high negative correlation.

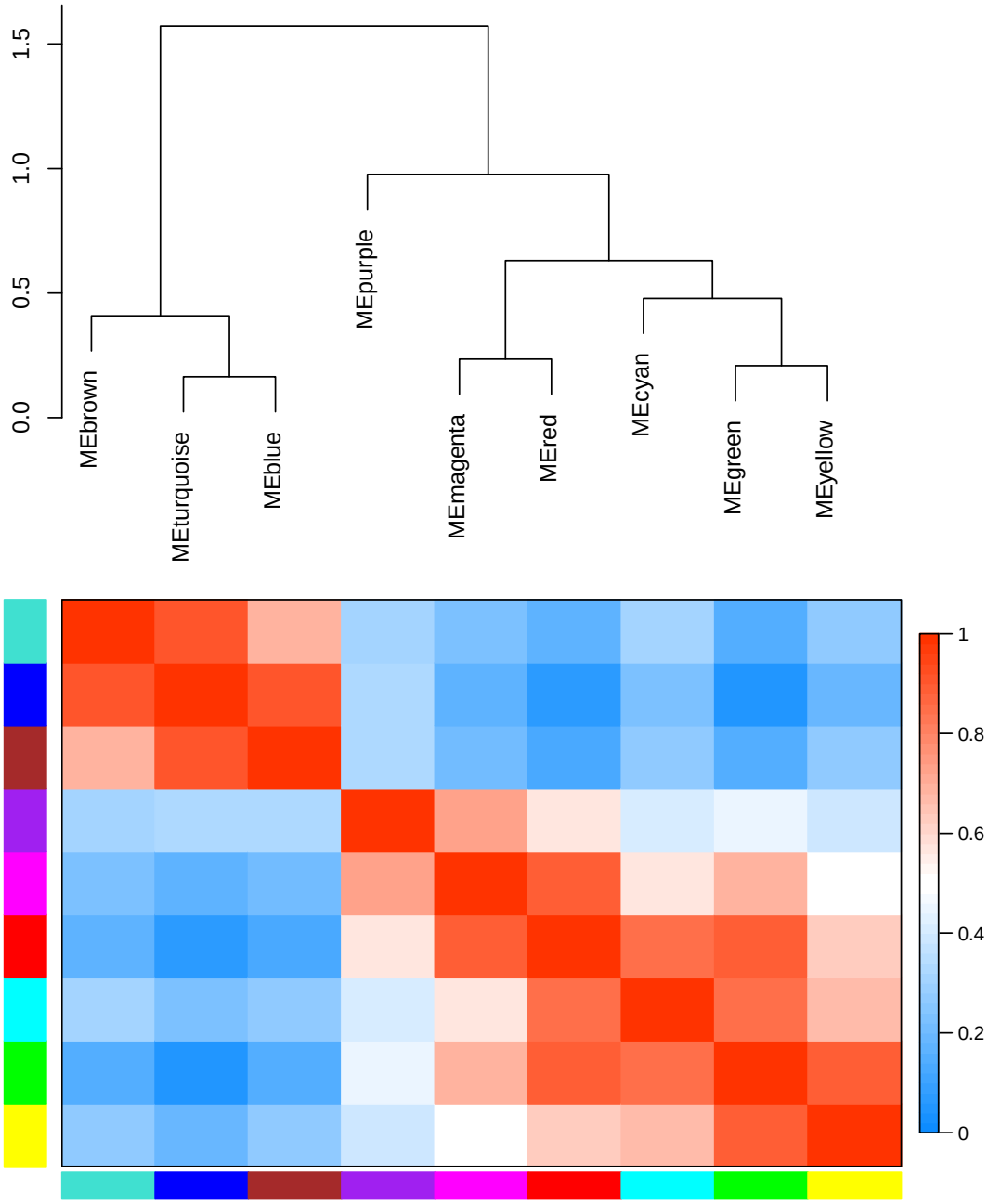


Fig. 6.9 **Module eigengene similarities and correlations for the Xue data set:** The dendrogram at the top show the similarity relationships between the module eigengene based on their Euclidean distances. The heatmap at the bottom shows the pairwise correlations of the module eigengenes measured by their Pearson correlation coefficient values as per legend on the right.

Relating Module Eigengenes to Embryonic Stages

The profile of the defined module eigengenes has differential correlation with the different cellular traits analysed as illustrated in Figure 6.10. The values displayed in each cell in the heatmap refer to the pairwise correlation coefficients for each module eigengene-cellular trait which is color coded as pallet from red (positive correlation) to blue (negative correlation) as per figure legend. The value between parentheses correspond significance of such correlation (student asymptotic p-values). The fact that each module eigengene has a very strong correlation with a distinct pre-implantation embryonic stage indicates that their expression profiles represent well the gene expression signatures from each stage and is able to capture their most salient differences. These results are very similar to what was published by Xue et al. (2013) and the small differences may be due to different choices in the tree cutting setting and selection of similar modules to merge. However, these small differences should not affect the overall final network structure, especially for the hub genes whose connectivities are very robust.

6.2.4 Network Based Gene Set Enrichment Analysis

As in the previous two chapters, the Figures 6.12 to 6.20 illustrate how the characterisation of each of the defined modules was performed as well as how the hub genes were selected and used as input for gene set enrichment analysis. The gene expression heatmaps in sub-figures a (Figures 6.12a to 6.20a) at the top of each panel serve to illustrate the gene expression profile of each module along the

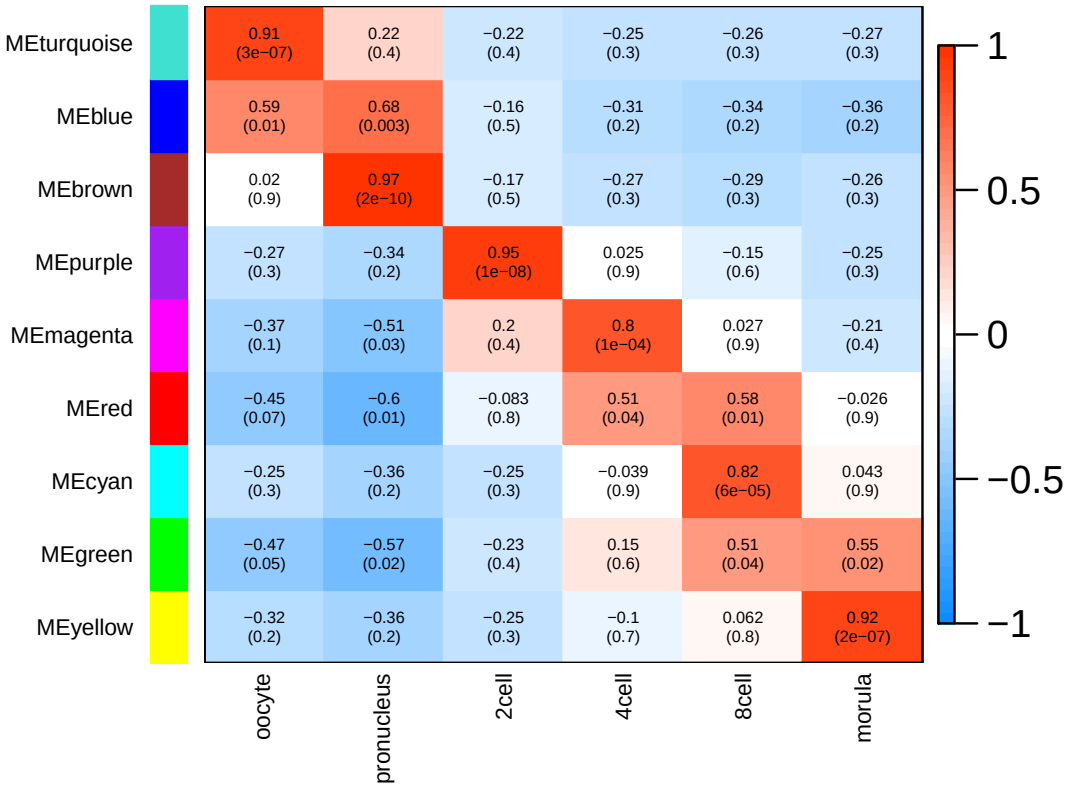


Fig. 6.10 **Module eigengene association with cell type:** This heatmap represent a high level representation of the identified network modules and the highly correlated modules to embryonic stages analysed providing an efficient mean of prioritisation of modules for further analysis. The module eigengene values were correlated with each embryonic stage analysed. The color coded heatmaps represent the level of correlation between the module eigengenes (rows) and embryonic stage (columns). As shown by the color legend guides on the right, red represent strong positive correlation, white stands for no correlation and blue represent strong negative correlation. Each correlation coefficient value is printed in the corresponding table cell along with the association significance level between brackets. We can see that for each embryonic stage there is at least one module highly correlated with the correlation coefficient value of at least 0.8.

different samples from the various embryonic stages analysed. The number of genes is shown between parentheses next to the module name. Likewise the bar plots below the heatmaps serve to summarise the expression of the genes in each module as the first principal component also referred as the module eigengene expression.

The process of selection of the hub gene sets is illustrated in the scatter plots in Figures 6.12b to 6.20b. The gene significance to an embryonic stage is plotted as a function of their module membership values for the selected module. The embryonic stage chosen for performing gene significance analysis is the one with the highest association to that module as viewed in Figure 6.10. The thresholds marked as dashed lines are arbitrarily selected values for module membership and gene significance in order to shortlist the potential hub genes with the highest values for both measures. The selected hub gene sets are then used as input for gene set enrichment analysis using the online tool Enrichr (Chen et al., 2013a) and the results are summarised as bar plots in sub-figures 6.12c to 6.20c. The full list of short listed genes per module containing their gene symbols and module membership values can be found in the attached media (Xue/xue_turquoise_genes_4_Erichr.csv to Xue/xue_yellow_genes_4_Erichr.csv). These files are already formatted in a way that can be directly uploaded into the Enrichr tool (http://amp.pharm.mssm.edu/Enrichr/) to reproduce the gene set enrichment analysis displayed in this chapter. The sub-figures 6.12c to 6.20c shows the Enrichr results for Gene Ontology - Biological Processes 2017.

Figure 6.11 summarises the results of all Figures 6.12 to 6.20. The module expression profiles and their most associated Biological Processes produced in this re-analysis of Xue et al. (2013) data set are very similar to what the authors have published in the original article which gives confidence in the reproducibility and robustness of the method as well as in the quality of the data set.

The original publication summarised their results into 5 broad functional terms, (a) cell cycle regulation (From oocyte to zygote), (b) transcription regulation (From zygote to 4 cell stage), (c) RNA processing/ribosome biogenesis (From zygote to 8 cell stage), (d) Translation (From 8 cell stage to morula) and (e) Mitochondrion/Energy metabolism (Morula). In my re-analysis I characterised each of the 9 final modules obtained with WGCNA which increased the resolution of the enriched functions associated with each embryonic stage. The summary of the enriched functions obtained in this re-analysis can be verified in Figure 6.11. The oocyte specific turquoise module is associated with hormonal response which gets down-regulated after fecundation. The blue module is expressed up until before 2 cell stage and is associated with cell division regulation. The brown module is specific to the pronuclear stage and is associated with regulation of DNA methylation reflecting the preparation of of the paternal genome DNA epigenetic reprogramming as a preparation for ZGA. The purple module is restricted to the 2 cell stage and is associated with chromatin remodelling and DNA repair which are strongly associated biological functions activated during ZGA. The pink module expression from 2 cell stage to 8 cell stage and is associated with mRNA stabili-

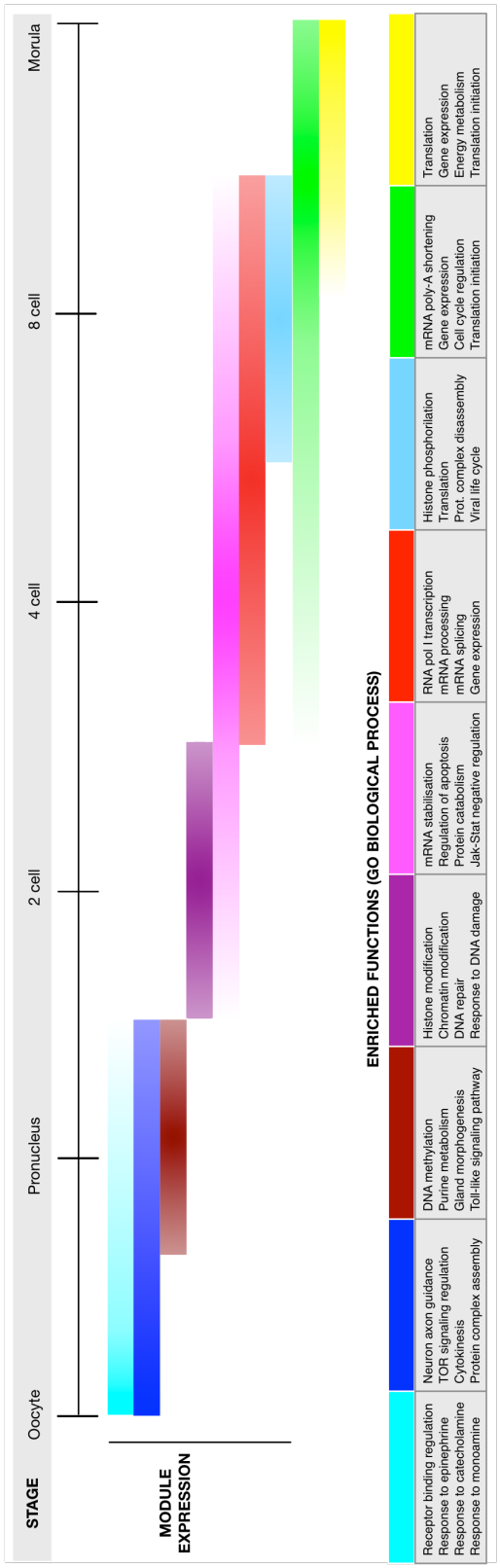
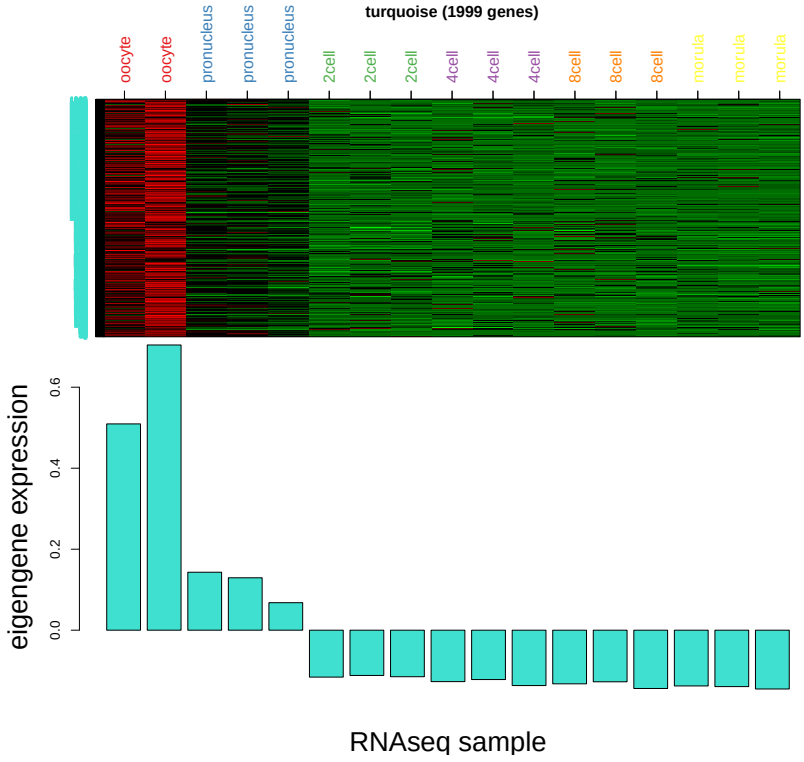


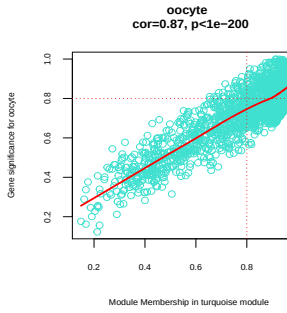
Fig. 6.11 **Dynamics of Biological Processes During Mammalian Pre-Implantation Development:** The embryonic stages are labelled in the grey bar on the top of the figure. The network eigengenes expressions are represented as horizontal color bars with the color intensity correlating with its expression levels. The most significantly enriched functions are summarised on the color coded boxes on the bottom of the figure.

sation. The red module has its expression restricted to 4 and 8 cell stage and is associated with gene expression regulation and post-transcriptional regulation during the first lineage segregation stage when pluripotency is being established. The cyan module is expressed only at 8 cell stage and is associated with translation regulation and chromatin remodelling. The green module expression extends from 4 cell stage to morula and is associated with gene expression, translation and cell cycle regulation. Finally the yellow module expression is restricted to morula stage and is associated with translation, gene expression and energy metabolism characteristic of the changes occurring prior to the formation of the blastocyst.

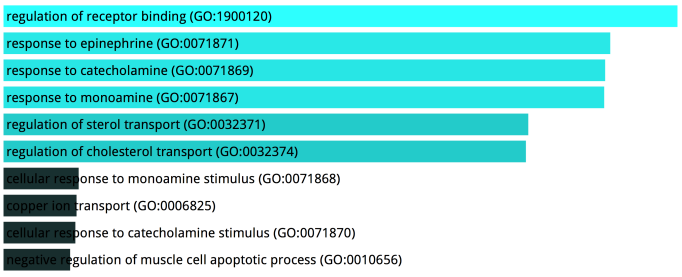
These results are in good agreement with the the biological events taking place in the pre-implantation development in the mouse and the fact that the defined network modules are able to represent well each of the embryonic stages gave me the opportunity to further inspect each intra-module network structure to try to identify commonalities between the reprogramming network modules defined in the previous two chapters and gain some insights into possible novel ways to try to improve the efficiency of somatic cell reprogramming.



(a) Turquoise module eigengene expression

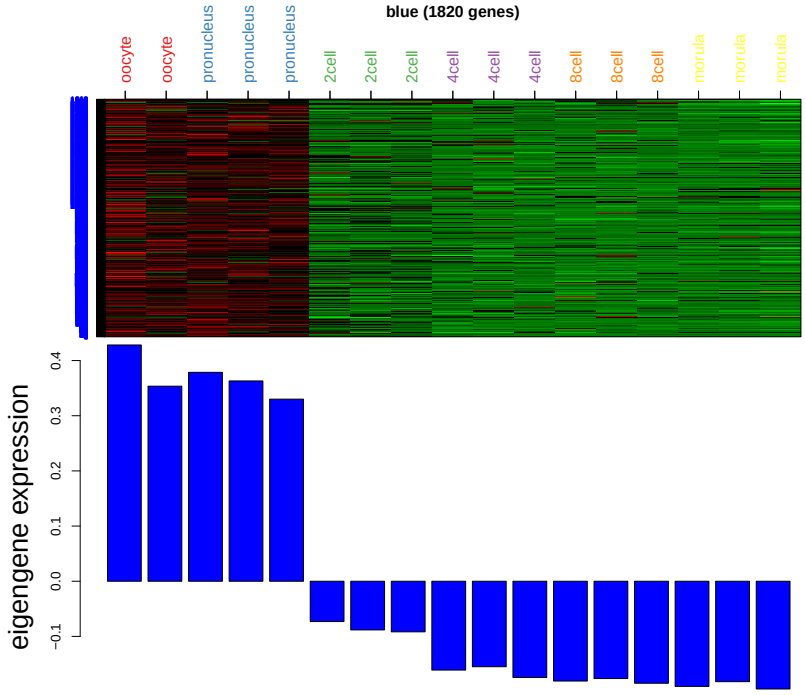


(b) Cutoff thresholds

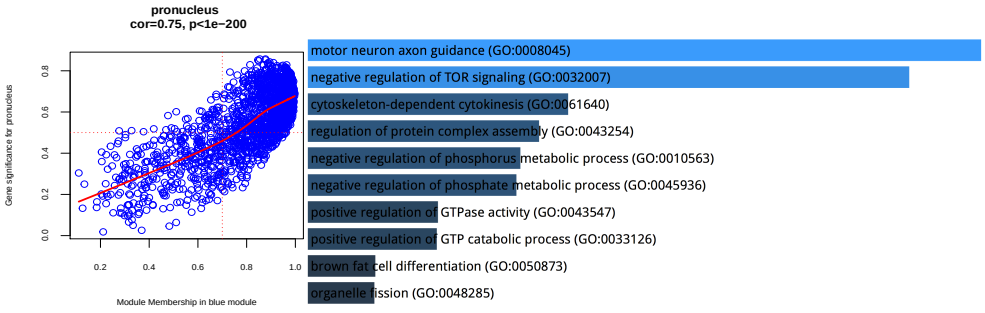


(c) Turquoise module Gene Ontology

Fig. 6.12 Turquoise module characterisation

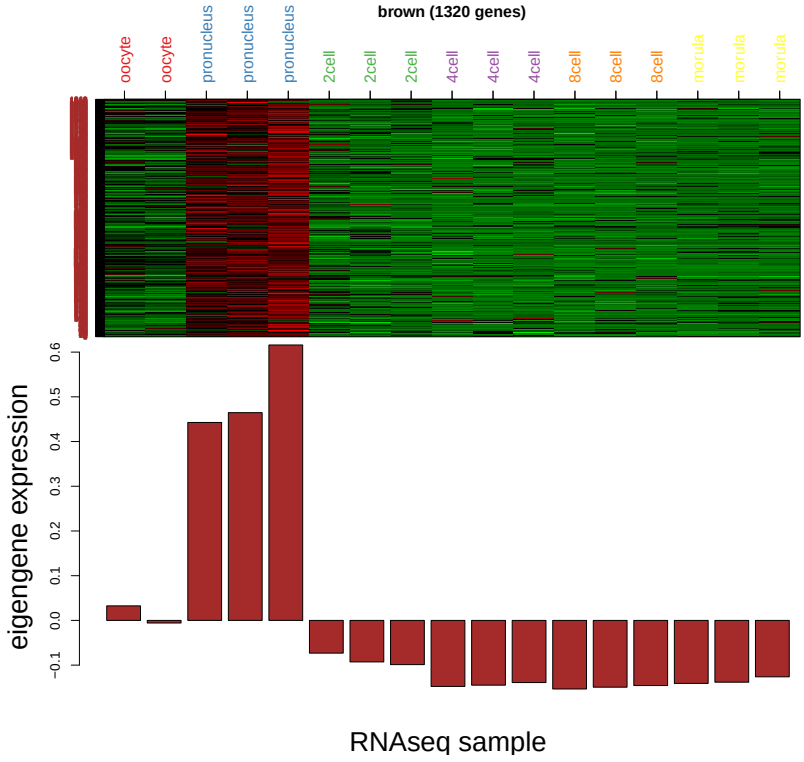


(a) Blue module eigengene expression

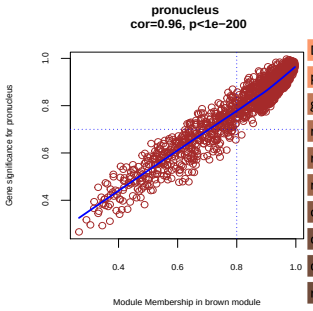


(b) Cutoff thresholds (c) Blue module Gene Ontology

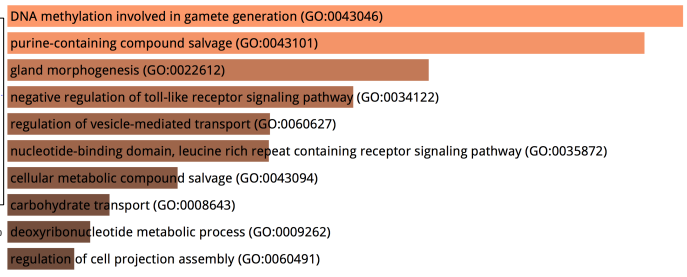
Fig. 6.13 Blue module characterisation



(a) Brown module eigengene expression

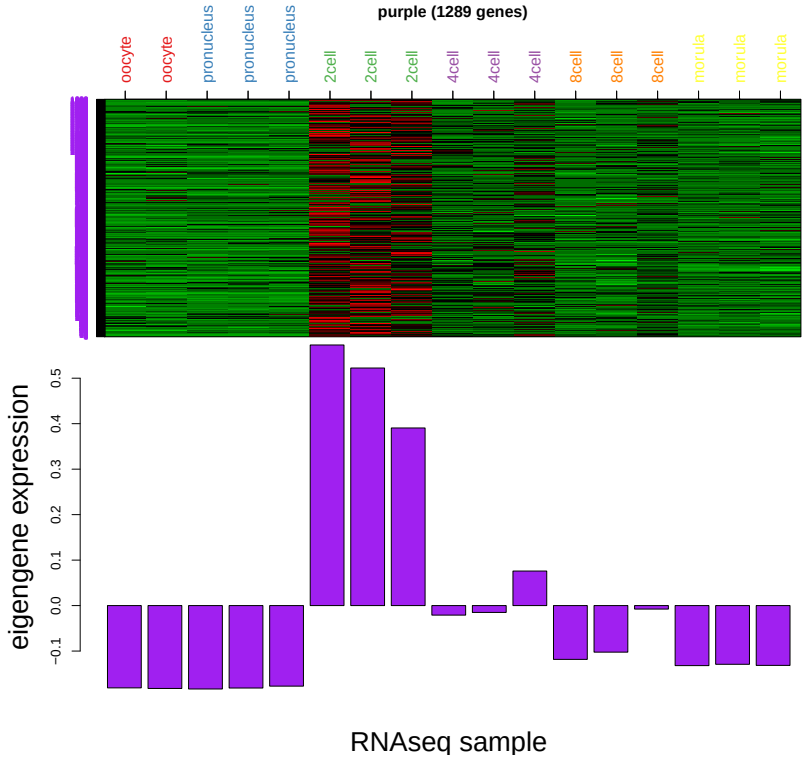


(b) Cutoff thresholds

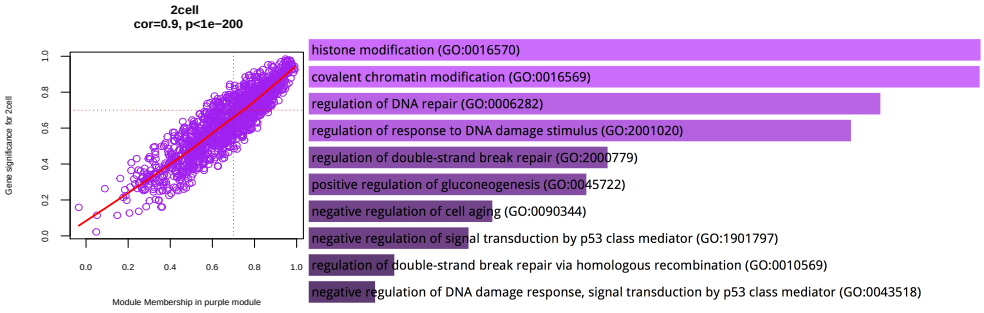


(c) Brown module Gene Ontology

Fig. 6.14 Brown module characterisation

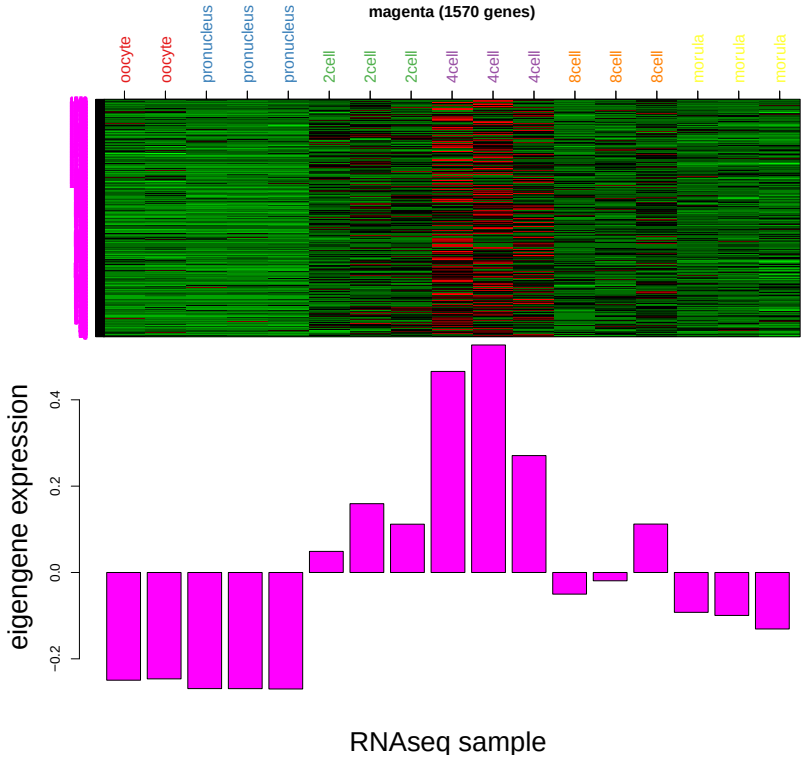


(a) Purple module eigengene expression

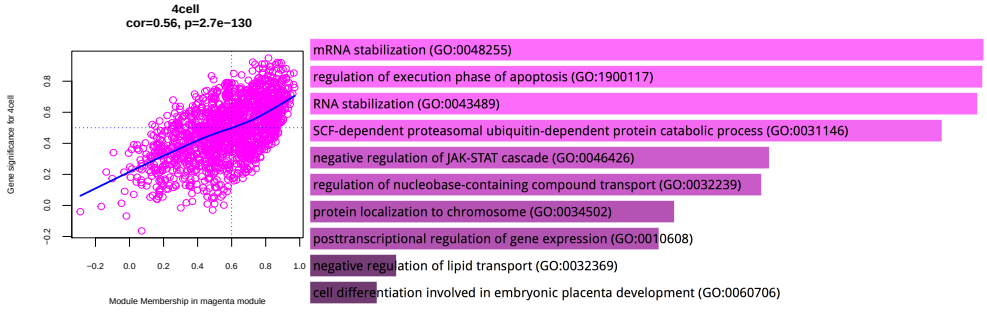


(b) Cutoff thresholds (c) Module Gene Ontology

Fig. 6.15 Purple module characterisation

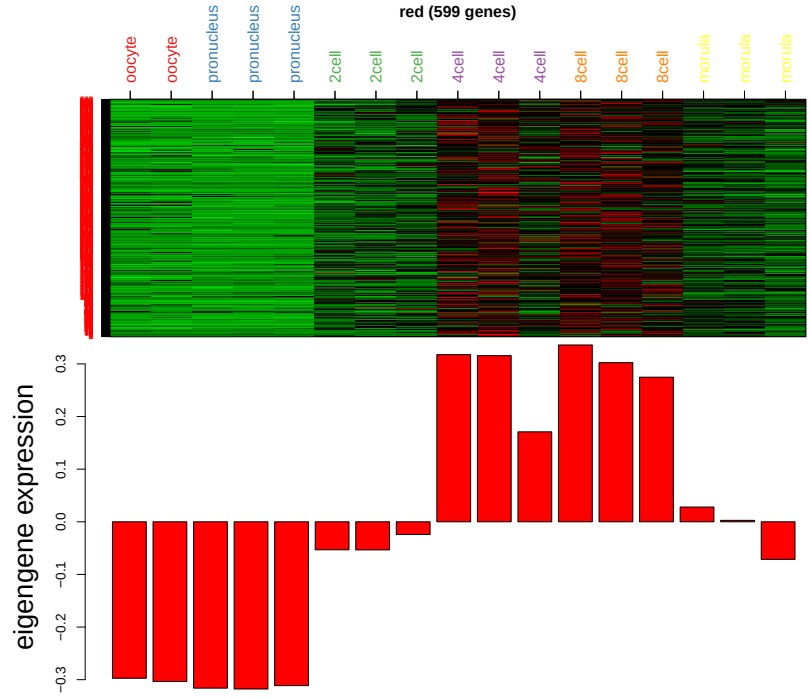


(a) Magenta module eigengene expression

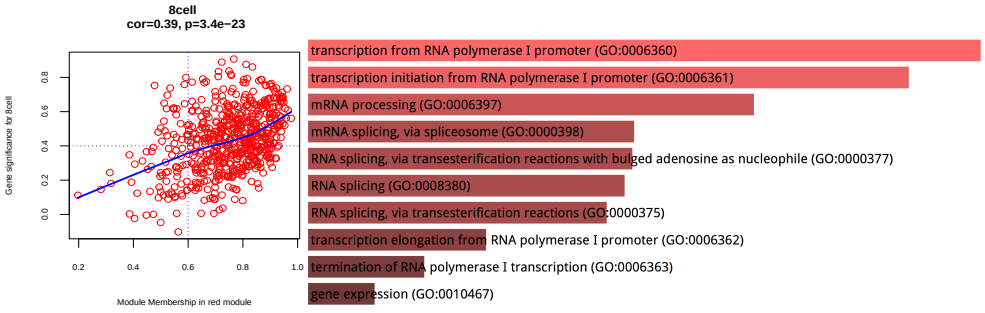


(b) Cutoff thresholds (c) Magenta module Gene Ontology

Fig. 6.16 Magenta module characterisation

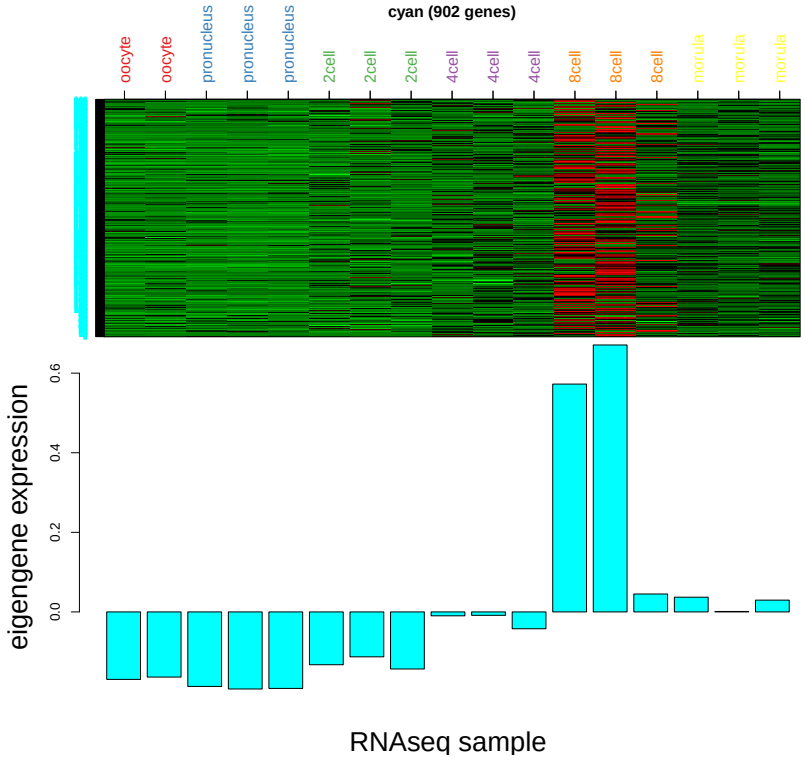


(a) Red module eigengene expression

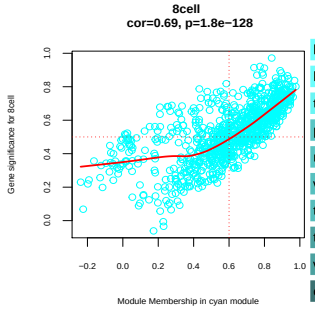


(b) Cutoff thresholds (c) Red module Gene Ontology

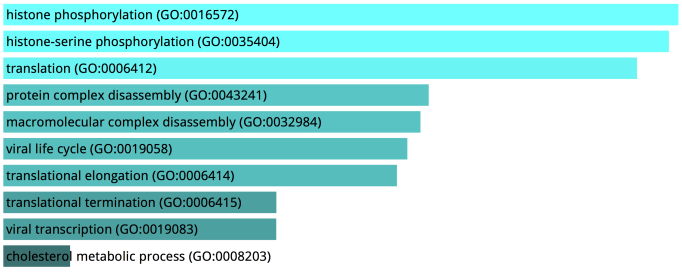
Fig. 6.17 Red module characterisation



(a) Cyan module eigengene expression

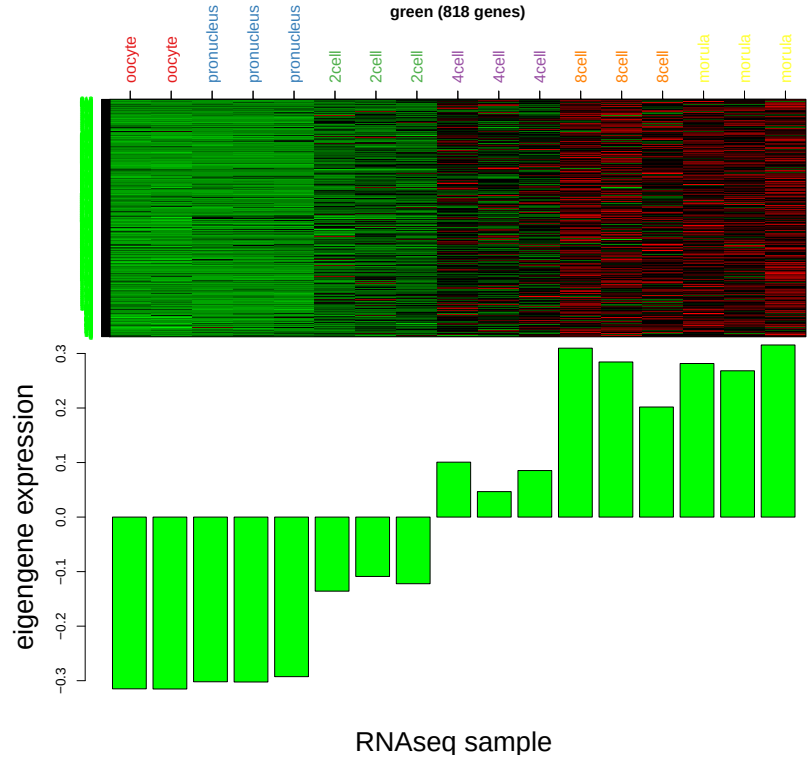


(b) Cutoff thresholds

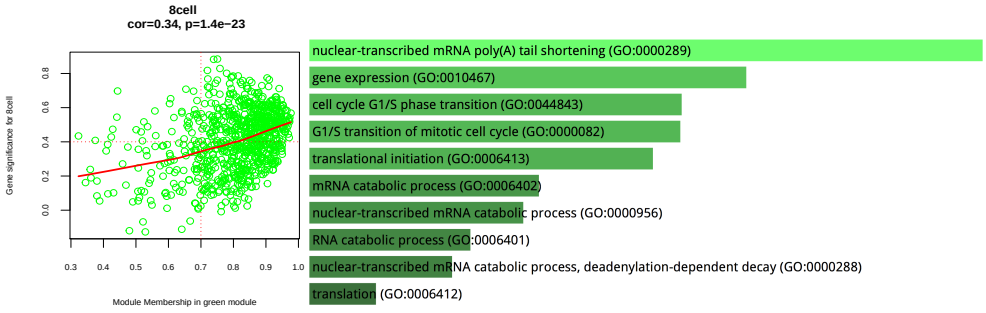


(c) Module Gene Ontology

Fig. 6.18 Cyan module characterisation

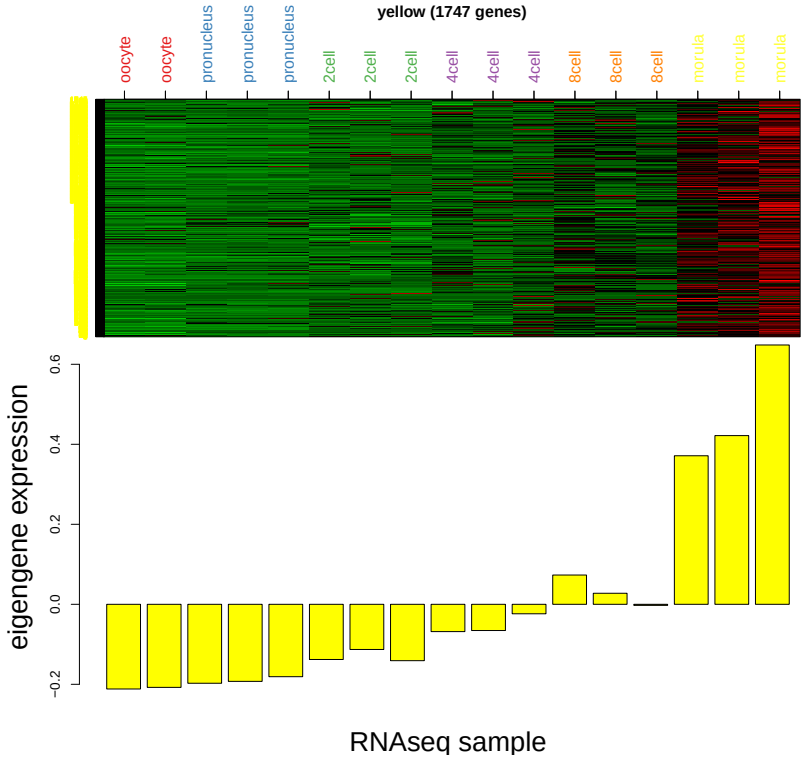


(a) Green module eigengene expression

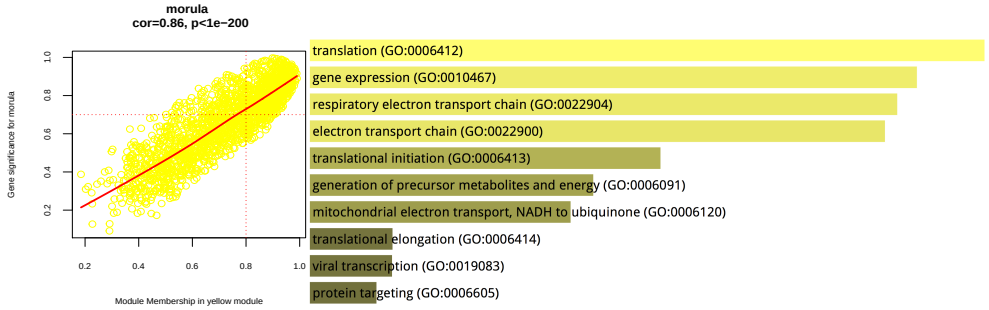


(b) Cutoff thresholds (c) Module Gene Ontology

Fig. 6.19 Green module characterisation



(a) Yellow module eigengene expression



(b) Cutoff thresholds (c) Yellow module Gene Ontology

Fig. 6.20 Yellow module characterisation

6.2.5 Network Module Preservation in Early Development and Reprogramming

In this section I compare the networks from pre-implantation development and somatic cell reprogramming to assess the possibility of finding preserved genetic circuits between the two processes and try to gain some insight from the embryo about how to improve the efficiency of iPS cell generation (Figure 6.21).

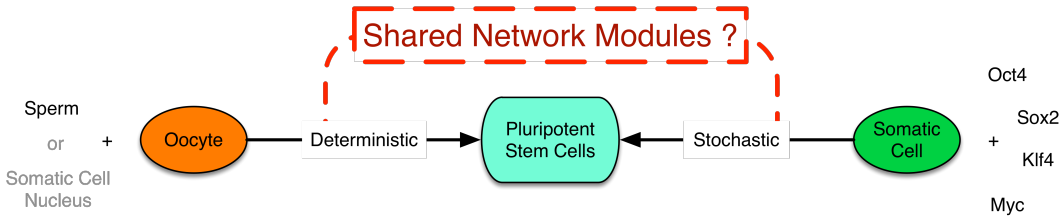


Fig. 6.21 Gene network analysis of shared modules between iPS cell generation and early development.

As we have seen in the previous chapter the use of network properties statistics is a powerful method for increasing the chances of identifying key master regulators, common between different networks because it takes into consideration not only the overlapping occurrences of genes individually, but also evaluates how the they are co-expressed with each other. The genes that have the highest intra-modular connectivities, that is, have the highest co-expression rates within a module are considered as the hub genes and are very often the key drivers of important biological functions that are associated with that module. By assessing the preservation of hub genes between networks it is possible to identify common

co-occurring processes between different conditions and this forms the rational for the network data mining performed below.

As a first step I cross-tabulated all the genes from the various network modules from pre-implantation development defined in this chapter and from somatic cell reprogramming defined in Chapter 4 and inferred their overlap significance by Fisher's exact test as illustrated in Figure 6.22.

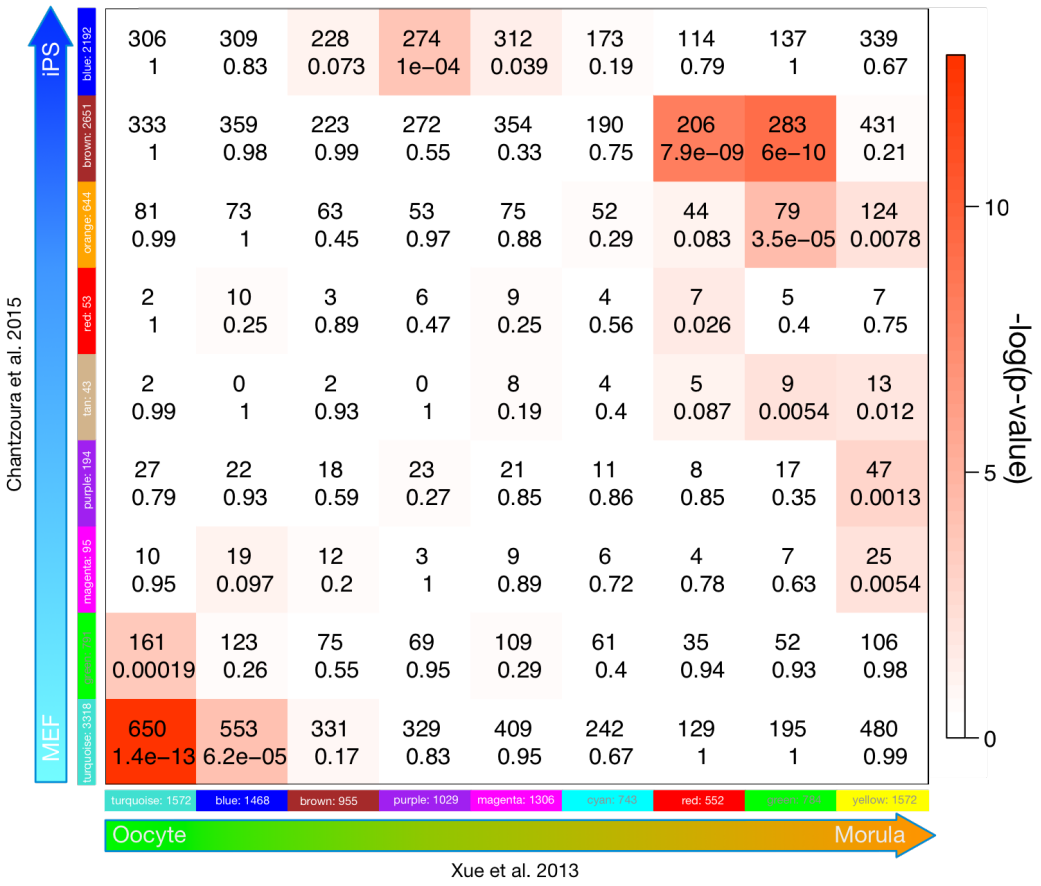


Fig. 6.22 **Chantzoura and Xue module genes cross tabulation:** Chantzoura modules are the rows and Xue modules the columns. Each row and column are labelled by the corresponding module color and the total number of genes in the module. In the table, numbers give counts of genes in the intersection of the corresponding row and column module. The table is color-coded by -log(p), the Fisher's exact test p value, according to the color legend on the right.

Next I performed the analysis of the WGCNA composite network preservation statistics Zsummary and MedianRank using the Xue data set as reference and Chantzoura as a test data set as well as the other way around switching the reference and test data sets (Figure 6.23).

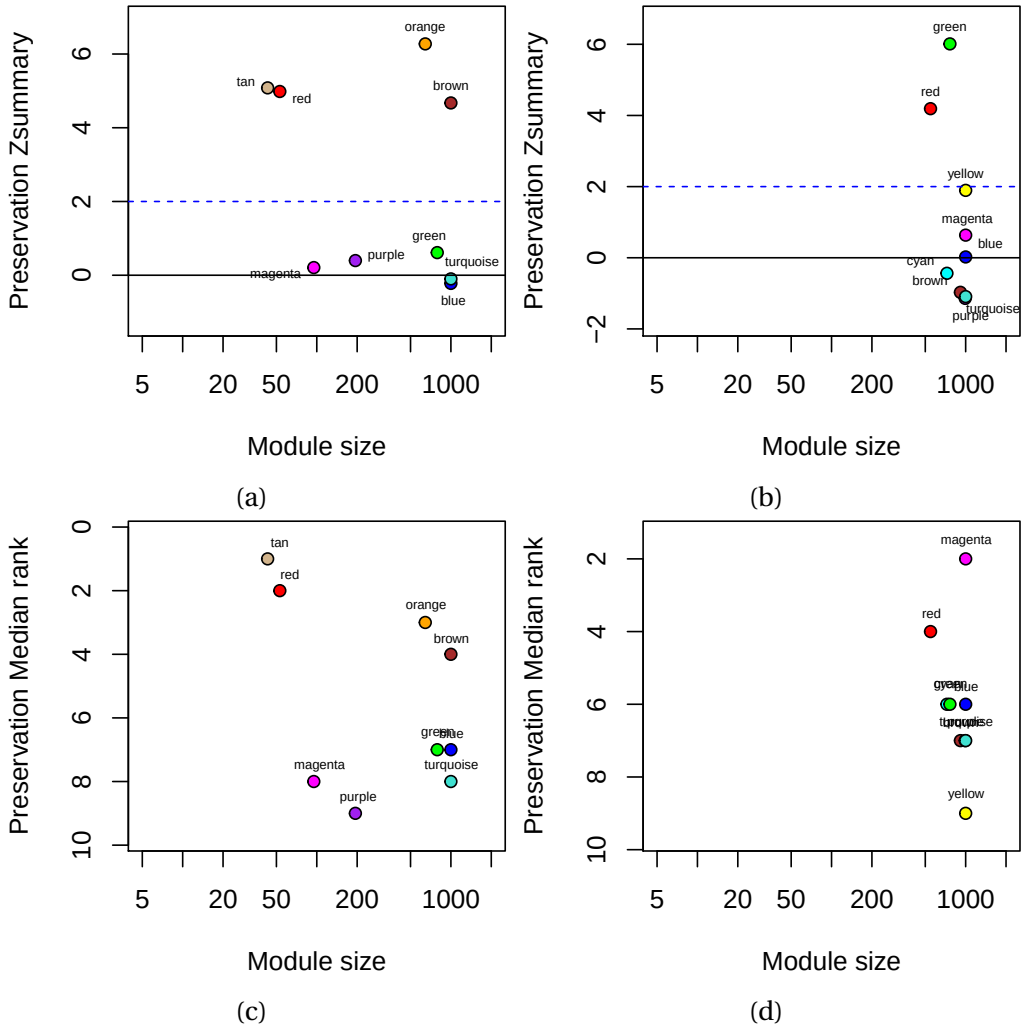
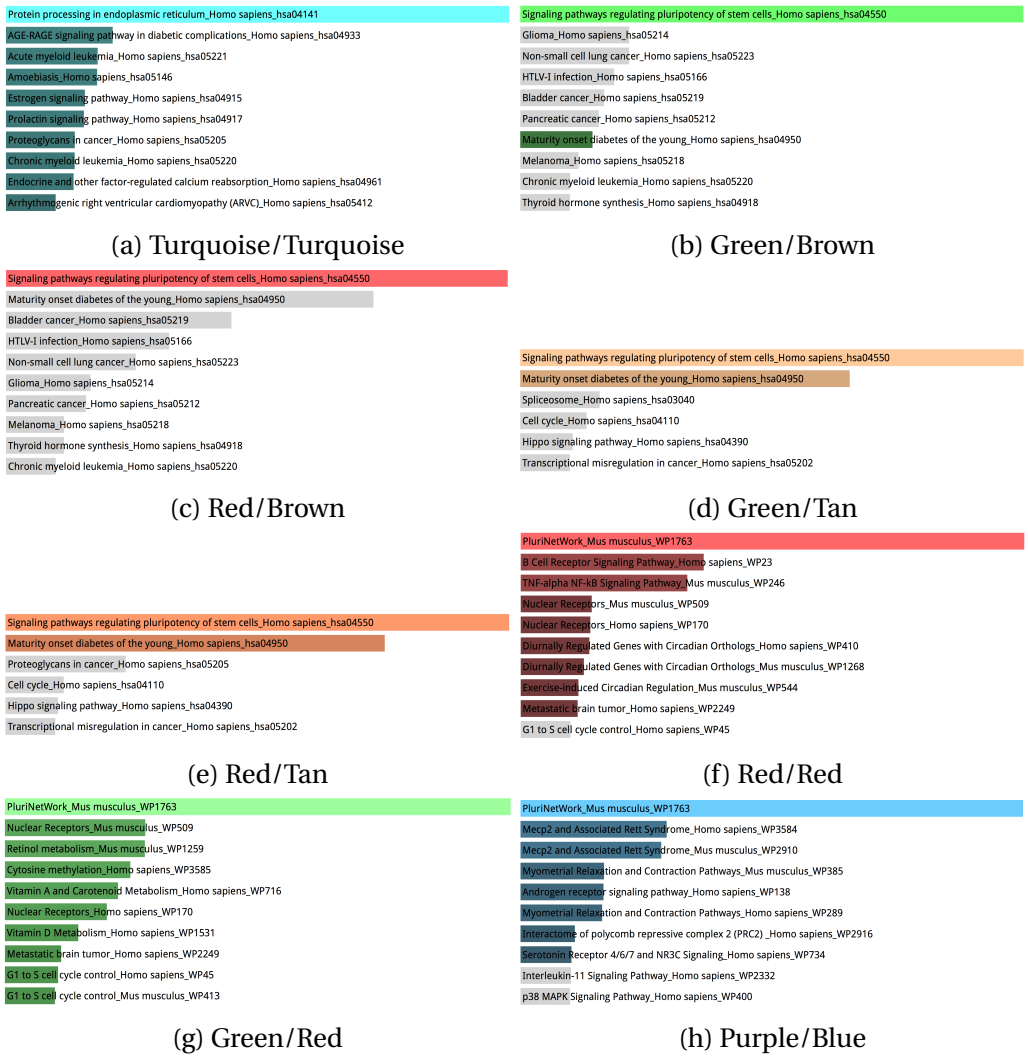


Fig. 6.23 **Module preservation statistics between and Xue and Chantzoura data sets:** In (a and c) Chantzoura was used as a reference data set and Xue as a test set and in (b and d) Xue was the reference and Chantzoura the test data set. (a and b) The summary statistic Zsummary (y-axis), as a function of the module size. Each point represents a module labelled by color. The dashed blue line indicate the threshold Zsummary = 2. (c an d) The composite statistic medianRank (y-axis), as a function of the module size. Each point represents a module, labelled by color as in panel a. As can be seen from the labels on the y-axis, the order on the y-axis is reversed. Low numbers on the y-axis indicate a high preservation.

The evaluation of the module preservation results illustrated in Figures 6.22 and 6.23 led me to some interesting observations. Curiously, the modules with the highest number of overlapping genes between the two networks were the turquoise modules in both data sets, which are associated with the oocyte in the Xue data set and MEFs in the Chantzoura data set. However the network preservation statistics between these two modules was not significant when preservation Zsummary was taken into account regardless of whether Xue or Chantzoura was used as a reference in the permutation tests (Figures 6.23 a and b). Of note, from the intersecting 650 genes between the turquoise modules of the two data sets 365 have module membership (aka intra-modular connectivity) values over 0.5 in both modules and are enriched for genes involved in protein processing in the endoplasmic reticulum and may not be relevant for the objectives of this work (Figure 6.24a).

The most preserved modules in the Xue network are the red, green and yellow modules, when considering preservation Zsummary with Xue as a reference and Chantzoura as a test set (Figures 6.23b and 6.22). More over, many preserved hub genes from these two modules had significant overlap between the four most preserved reprogramming modules tan, red, orange and brown (Figures 6.23a, 6.23c and 6.22). Interestingly the red and green pre-implantation modules are associated with the 4C-8C and 8C-morula stages respectively, which correspond to the final stages when the pluripotency network is being established (Figure 6.11). Similarly the reprogramming modules tan, red, orange and brown also correspond to the final stages of pluripotency acquisition (Figure 4.15). Indeed

the overlapping hub genes with module memberships values above 0.5 in both data sets are significantly associated with pluripotency related genes. Figures 6.24b, 6.24c, 6.24d, 6.24e, 6.24f, 6.24g illustrate the gene set enrichment analysis of the significant overlapping transcription factors in pre-implantation and reprogramming networks.



(a) Turquoise/Turquoise (b) Green/Brown

(c) Red/Brown (d) Green/Tan

(e) Red/Tan (f) Red/Red

(g) Green/Red (h) Purple/Blue

Fig. 6.24 **Gene set enrichment analysis of preserved hub genes:** Overlapping hub genes with module memberships (MM) over 0.5 were used for gene set enrichment analysis in Enrichr. The bar plot labels correspond to the module color labels as Xue / Chantzoura respectively. The Turquoise/Turquoise (a) gene list was composed of all genes meeting the MM thresholds. The gene lists from b to g were further filtered to contain only transcription factors. The Purple/Blue gene list was filtered to contain only transcription factors and chromatin remodellers.

Table 6.1 Preserved hub genes between pre-implantation development and somatic cell reprogramming. Genes are listed in order of their module membership values per module overlaps. TFs stands for Transcription factors and CRs stands for Chromatin Remodellers

Overlapping Modules	Module Associations	Genes
Green/Brown	8C-Morula/iPSCs	Zfp42, Dnajc2, Esrrb, Pou5f1, Zfp568, Gm14139, Cdc5l, Sall1, Mta3, Zfp296, Jarid2, Tfam, Sox15, Nr5a2, Terf1, Zfp599, Zfp110, Nfx1, Zfp143, Ttf1, Elf3, Gm3604, Zbtb3, Nfyb, E2f3, Zbtb48, Klf8, Tgif1, Snapc4, Zfp217, Klf2, Gabpa,Zfp157
Green/Red	8C-Morula/pre-iPSCs	Mbd3, Dnajc2, Rxrb, Snapc4, Sall1, E2f3, Gtf2i, Zfp747
Green/Tan	8C-Morula/pre-iPSCs	Zfp42, Cdc5l, Esrrb, Tead4, Sox15, Pou5f1, Tbx3, Gm14139, Nr5a2, Zfp296, Zfp568, Mta3, Klf8, Zbtb17, Terf1, Zfp143, Zfp599, Gm3604, Elf3
Red/Brown	4C-8C/iPSCs	Nr5a2, Baz2a, Dnajc2, Sox15, Zfp459, Esrrb, Sall1, Zfp568, Gm14139, Terf1, Zfp143, Rcor2, Nfyb, Nanog, Ttf1, Zbed4, Zfp110, Zfp42, Gm3604, Klf8 Hmgb2, Gabpa, Zfp599, E2f3, Zfp59, Grhl2, Klf2, Zfp217, Mga, Zfp763, Zbtb44, Zfp808, Zbtb8a, Mycn, Zfp317, Tgif1, Zfp566, Dmrt1, Zfp105, Sox21, Zfp273
Red/Red	4C-8C/pre-iPSCs	Dnajc2, Mbd3, Rcor2, Nr1d2, Gtf2i, Zfp740, Sox21, Yy1, Zfp105, Zfp292, Sall1, E2f3, Smarcc1, Pax6, Zbed4, Grhl2, Hmgb2, Mycn
Red/Tan	4C-8C/pre-iPSCs	Sox15, Obox6, Nr5a2, Zfp820, Tead4, Klf8, Esrrb, Baz2a, Zfp143, Gm14139, Zfp459, Terf1, Zfp568, Zbtb17, Tbx3, Zfp42, Zfp59, Nanog, Gm3604, Zbtb44, Zfp273, Zfp599, Zfp808
Purple/Blue (TFs)	2C/2C-like iPSCs	Kdm5b, Mga, Zfp367, Zscan4c, Zfp281, Zscan5b, Hic2, Rlf, Tfdp2, Etv5, Zscan4d, Zfp639, Zfp446, Zfp775, Zfp7, Yy2, Zfp352, Smarcc1, Gabpa, Zfp689, Nrf1, Ctcf, Zbtb8a, Atf1, Zbtb44, Atf2, Pou2f3, Zfp553, Usf1, Zfp142, Zfp263, Zfp574
Purple/Blue (CRs)	2C/2C-like iPSs	Kat2b, Tet1, Taf7, Suz12, Kdm4c, Cbx7, Trim33, Chd1

I also observed a significant overlap of hub genes from the Xue purple module and the Chantzoura blue module. This is very relevant because the purple module is exclusively associated with the 2C stage of pre-implantation marked by the transcriptional burst of major ZGA. Moreover the in depth evaluation of the genes associated with the reprogramming blue module indicates that this module is capturing the genetic network of the elusive 2C-like pluripotent stem cells that spontaneously cycle at a very low rate between this totipotent state and the most prominent naïve pluripotent state (Macfarlan et al., 2012). It is currently believed that somatic cell reprogramming into induced pluripotency occurs via an initial rate limiting stochastic stage followed by a more deterministic hierarchical late stage (Buganim et al., 2013). Therefore genetic dissection of networks operating in this early 2C embryonic stage may provide clues about achieving more efficient reprogramming of somatic cells.

A number of 2C/2C-like factors identified in this analysis (Table 6.1) have been linked to important functions in pluripotent stem cells and have even been shown to improve the efficiency and quality of iPS cell generation. Another handful of factors identified here as 2C/2C-like hub genes have not been functionally characterised and may prove to be key factors in the induction of totipotency and pluripotency. The list of 2C/2C-like hub genes identified are composed of 33 transcription factors and 8 chromatin remodellers and some of their known functions are described below (Table 6.1).

Zscan4 is a well known family of markers of ZGA with embryonic expression being exclusive to the 2C stage (Falco et al., 2007).Three members of the *Zscan* gene family (*Zscan4c*, *Zscan4d* and *Zscan5b*) have been identified as highly ranked 2C/2C-like hub genes. *Zscan4* has been shown to be a potent natural factor that enhances the efficiency of iPSC formation and reactivates early embryonic genes (Hirata et al., 2012). Moreover *Zscan4* promotes genomic stability during reprogramming and dramatically improves the quality of iPS cells as demonstrated by tetraploid complementation. *Zscan4* also enhances telomere lengthening and dramatically decreases the DNA damage response (DDR) during reprogramming (Jiang et al., 2013). *Zscan4* has also been studied in ES cells where it has been shown to regulate telomere elongation and promote genomic stability (Zalzman et al., 2010). The *Atf1* and *Atf2* factors have also been found to be essential for ZGA in mouse embryos (Jin and O'Neill, 2010, 2011, 2014) and *Trim33* involved in silencing of transposable elements and contribution to the maintenance of genomic stability (Isbel et al., 2015). Another well know maker of 2C cells also identified here as a hub gene is *Zfp553* which has been recently shown to be essential for the maintenance and acquisition of pluripotency in mice by activating the expression of *Oct4* and *Nanog* (Yu et al., 2016). *Etv5* was identified as being an early reprogramming regulator preceding activation of key pluripotency regulators such as *Rex1* (also known as *Zfp42*), *Dppa2*, *Nanog* and *Sox2*. *Etv5* was found to be preferentially expressed in reprogramming-prone cell populations and highly expressed in ESCs suggesting a functional role in poising early reprogramming (Lujan et al., 2015).

A number of identified factors are known to be involved in DNA demethylation such as *Tet1*. There are 3 known *Tet* genes and *Tet3* is the main transcript variant in oocytes which is key in the ability to reprogram the injected nuclei from somatic cells (Gu et al., 2011). However I was able to verify with the data set from Xue et al. (2013) a massive switch to *Tet1* between the pronuclear stage and the 2C stage where the major ZGA takes place, with the *Tet3* transcript being barely detectable at this stage. Mouse ES cells, on the other hand, expresses *Tet1* and *Tet2* which have distinct roles in regulating hidroxylation of 5-methyl cytosines. Tet1 is associated with the methyl-cytosine hydroxylation at transcription start sites (TSS), whereas Tet2 is mostly associated with this conversion in gene bodies (Huang et al., 2014b). Importantly *Tet1* facilitates iPSC induction by promoting *Oct4* demethylation and reactivation and can replace *Oct4* at the standard Yamanaka 4 factors to generate fully pluripotent iPSCs (Gao et al., 2013). Another identified 2C/2C-like factor involved in the maintenance of low levels of DNA methylation is Rearranged L-Myc Fusion, RLF which acts as an epigenetic modifier at CpG island shores and enhancers across the genome.(Bourke et al., 2017).

While many other factors such as *Cbx7* (Morey et al., 2012), *Chd1* (Gaspar-Maia et al., 2009), *Zfp553* (Yu et al., 2016), *Smarcc1* (Huang et al., 2012), *Hic2* (Hansson et al., 2012) , *Gabpa* (Ueda et al., 2017), *Ctcf* (Wang et al., 2017) have all been associated with the maintenance and/or acquisition of pluripotency a few other factors may have negative effects in reprogramming. One of them is Lysine Demethylase 5B (KDM5B) which was the most highly ranked 2C/2C-like factor based on its module membership values. This transcription regulator is essen-

tial for early embryo development and is required for ZGA by actively removing broad H3K4me3 chromatin domains (Dahl et al., 2016). However it has also been linked to the modulation of mouse ES cell self-renewal and suggested to be a a barrier to the reprogramming process (Kidder et al., 2013). *Znf281* is another factor which has been linked to reprogramming as a roadblock which suggests a need for careful caution on the selection of some selected 2C/2C-like genes as candidate reprogramming factors. *Znf281* has been characterised as an EMT-inducing transcription factor down regulating E-cadherin (Hahn and Hermeking, 2014) and also been shown to mediate *Nanog* auto-repression through recruitment of the NuRD complex inhibiting somatic cell reprogramming(Fidalgo et al., 2012). Also *Yy2* which is from the same gene sub family as *Rex1* (Kim et al., 2007) has been recently identified to have a critical role in regulating mESC functions through control of key pluripotency factors, including *Oct4* and *Esrrb*. Importantly, over-expression of *Yy2* directs the differentiation of mESCs into cardiovascular lineages (Tahmasebi et al., 2016).

Therefore, whereas there is good evidence that a number of genes could directly contribute to the improvement of reprogramming efficiency to pluripotency, some factors might negatively interfere with pluripotency network but still be important for the ZGA process in the context of totipotency. This possible drawback may be circumvented if such factors are used only as a short transient pulse at the beginning of reprogramming test experiments. The identification of network hub genes was instrumental in the arduous task of short listing genes with important biological functions. The list of genes identified in this analysis should provide a

platform for the development of new improved reprogramming cocktails as well as the reference for further functional studies in pre-implantation embryology.

6.2.6 Conclusion

In this chapter I performed a weighted gene co-expression network analysis of a single-cell RNA-Seq data set from mouse pre-implantation embryos produced by Xue et al. (2013) with the purpose of trying to identify gene networks which could also be present during the process of factor induced somatic cell reprogramming and harness this information to select hub genes with potential for improving the efficiency of reprogramming. The embryonic period from 8 cell to morula stages showed a moderate network overlap with that of late stage reprogramming marked by reactivation of endogenous *Sall1*, *Pou5f1*, *Zfp42*, *Nr5a2* and *Esrrb* among other well known pluripotency genes. Interestingly the embryonic 2 cell stage specific network had a significant structure overlap of hub gene connections when compared to a pluripotency network module associated with a 2C-like properties which spontaneously appear a low frequency in prolonged culture of pluripotent stem cell cells (Macfarlan et al., 2012). Given that the totipotent 2C embryonic stage is the period when the zygotic genome becomes fully reprogrammed and reactivated I speculate that the hub genes in common between the 2C embryonic stage and the 2C-like stem cell derived cell populations could be used as reprogramming factor candidates for improving somatic cell reprogramming efficiency. A number of the highly ranked hub genes have already been reported to indeed

enhance iPS cell generation efficiency and the selection of such factors together with other closely associated network hub genes could potentially help kick start a totipotent-like stage prior to the acquisition of pluripotency *in vitro* mirroring the deterministic process that occurs *in vivo* in pre-implantation embryos.

Chapter 7

Conclusion

The seminal discovery by Takahashi and Yamanaka that somatic cells could be reprogrammed to pluripotency by the four OSKM factors (Oct4, Sox2, Klf4, c-Myc) has revolutionised understanding about cell identity (Takahashi et al., 2007; Takahashi and Yamanaka, 2006b). Despite the numerous advances in the study of the mechanisms of iPS cell generation it remains a poorly understood phenomenon. It is believed that OSKM factors induce somatic cells to go through a long and inefficient stochastic initial phase followed by a more deterministic stabilisation stage before full pluripotency acquisition (Buganim et al., 2013). A number of studies have unravelled critical roadblocks to reprogramming (Ebrahimi, 2015). The study of maternal factors and the mechanisms of zygotic genome activation should hold important clues for the improvement of reprogramming efficiency since the embryonic development must occur in a deterministic step wise fashion.

The work presented here began by identifying *H2afz* as one of the most expressed genes in the oocytes of the *Axolotl* salamander. I set out to investigate a role of *H2afz* in mamalian iPS cell generation motivated by the fact that this histone variant is also the most abundant histone in ES cells (Kafer et al., 2010) and has proven to be important for pluripotency regulation (Creyghton et al., 2008; Hu et al., 2013). The over-expression of this histone variant was able to slightly improve the efficiency of reprogramming of two different pre-iPS cells which over-expresses classical Yamanka factors and acquire pluripotency at a very low frequency. Interestingly *Hist1h2aa*, a maternal canonical histone, was also able to improve reprogramming efficiency to a lesser extent and this has been independently reproduced (Shinagawa et al., 2014). Moreover the role of *H2afz* in reprogramming enhancement has also now been published (Wang et al., 2015). This study also reported that the pluripotency activity of Nanog is dependent of H2AZ. Surprisingly my results showed that somatic cells deficient for Nanog can still be reprogrammed by Yamanka factors if *H2afz* is over-expressed.

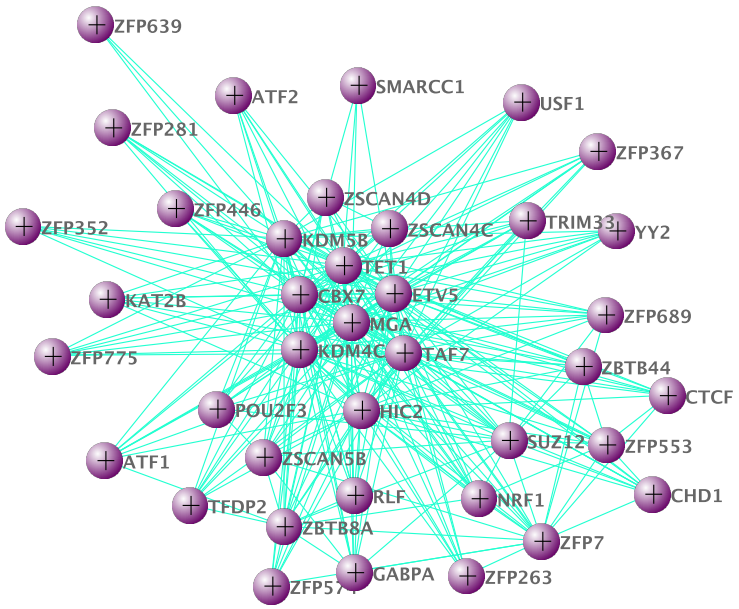
With the assumption that natural mechanisms occurring in early pre-implantation embryos could reveal more efficient ways of generating iPS cells I turned to re-analysing a number of published gene expression data sets. I performed gene co-expression network analyses with a single cell RNA-seq data set from mouse pre-implantation development (Xue et al., 2013) as well as with two reprogramming studies which utilised cell sorting to identify intermediate stages of reprogramming (Chantzoura et al., 2015; O'Malley et al., 2013b). The network analysis of iPS cell reprogramming generated a higher resolution analysis than first published

and resulted in highly reproducible network modules with significant associations to the various intermediate reprogramming stages. One of these preserved network modules had *H2afz* as one of the hub genes along with a number of chromatin remodelling associated genes which could be tested for their ability to complement the function of *H2afz* in enhancing reprogramming efficiency. The biological function enrichment analysis of intra-modular hub genes were instrumental in identifying the most salient events on the route to induced pluripotency. Interestingly this network analysis was able to distinguish two pluripotency gene networks, one associated with the well know ground state pluripotency and another one associated with a recently reported transient cell state similar to the totipotent 2 cell stage in pre-implantation embryos (Macfarlan et al., 2012).

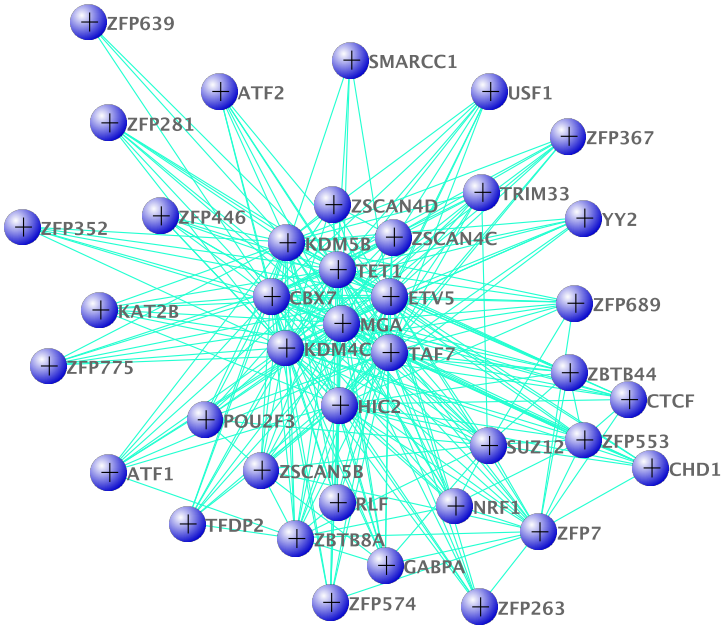
The analysis of the pre-implantation data produced well defined network modules which represented the gradual transitions from oocyte to morula and the enrichment analysis of intra-modular hub genes revealed the biological processes associated to each embryonic stage. Lastly the network statistical analysis between early development and reprogramming revealed a moderate preservation of modules associated with the later stages of pluripotency acquisition in both conditions.

Moreover, there was also a significant overlap between a module associated with the 2 cell stage in embryos (purple) and a module associated with 2C-like cell state of pluripotent stem cells (blue). The embryonic 2 cell stage is associated with the zygotic genome activation, therefore the analysis of network properties of

the top hub genes common between the two 2C conditions is highly relevant for reprogramming. Indeed their network structures proved to be extremely similar (Figure 7.1).



(a) Xue purple module (2C Stage)



(b) Chantzoura blue module (2C-like cells)

Fig. 7.1 **Gene Networks in 2C stage and 2C-like cells:** The most highly connected intramodular hub genes are indicated by purple and blue dots (nodes) for the 2C pre-implantation stage (Figure 7.1a) and 2C-like cells (Figure 7.1b) respectively. The thickness of the green lines (edges) represent the topological overlap measures between two nodes. The most connected nodes are clustered in the centre of the networks.

The intra-modular analysis of the most preserved hub genes between the two conditions uncovers relevant genes with could be used for new reprogramming cocktail compositions. The topological overlap measure (TOM) takes into account not only the pairwise correlations (co-expressions) but also measures how many common genes each one correlates with. Taking TOM into consideration, the most highly connected genes both in 2C stage and 2C-like cells are *Tet1*, *Cbx7*, *Etv5*, *Mga*, *Kdm5b*, *Kdm4c*, *Taf7*, *Zscan4c* and *Zscan4d*.

To my knowledge still there is no report describing an attempt to reprogram somatic cells to pluripotency via a 2C-like stage. The forced activation of such a network may help overcome important reprogramming barriers. *Tet1*,*Etv5* and *Zscan4* have all been shown to have potent reprogramming enhancing activity when combined with the 4 Yamanaka factors (Costa et al., 2013; Gao et al., 2013; Hirata et al., 2012; Lujan et al., 2015). Moreover, somatic cell nuclear transfer (SCNT) studies identified reprogramming resistant regions (RRRs) enriched for H3K9me3, that are expressed normally at 2 cell mouse embryos generated by *in vitro* fertilisation but normally not by SCNT. These RRRs could be successfully reprogrammed by forced expression of *Kdm4d* greatly improving SCNT efficiency (Matoba et al., 2014). Likewise the over-expression of *Kdm5b* and *Kdm4c* may also help improve reprogramming efficiency *in vitro* prior to pluripotency acquisition. Taken together the evidence of these being hub genes in a network preserved between 2 cell embryos and 2C-like stem cells and reports linking some of these genes to reprogramming enhancement it is likely that a combination of them might more efficiently kick start the early phase of reprogramming reducing

its stochasticity. It would be worth evaluating this 2C factors cocktail with and without combining them with Yamanaka factors. Since the 2C network precedes the pluripotency network it might be possible to completely omit the Yamanaka factors altogether with the right combination of 2C factors.

References

Abbott, D. W., Ivanova, V. S., Wang, X., Bonner, W. M., and Ausió, J. (2001). Characterization of the Stability and Folding of H2A.Z Chromatin Particles: Implications for Transcriptional Activation. *Journal of Biological Chemistry*, 276(45):41945–41949.

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M. S., Tomasini, L., Ferrandino, A. F., Rosenberg Belmaker, L. A., Szekely, A., Wilson, M., Kocabas, A., Calixto, N. E., Grigorenko, E. L., Huttner, A., Chawarska, K., Weissman, S., Urban, A. E., Gerstein, M., and Vaccarino, F. M. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature*, 492(7429):438–442.

Adam, M., Robert, F., Larochelle, M., and Gaudreau, L. (2001). H2A.Z Is Required for Global Chromatin Integrity and for Recruitment of RNA Polymerase II under Specific Conditions. *Molecular and cellular biology*, 21(18):6270–6279.

Ai, W., Zheng, H., Yang, X., Liu, Y., and Wang, T. C. (2007). Tip60 functions as a potential corepressor of KLF4 in regulation of HDC promoter activity. *Nucleic acids research*, 35(18):6137–6149.

Allard, S. (1999). NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p. *The EMBO journal*, 18(18):5108–5119.

Allis, C. D., Richman, R., Gorovsky, M. A., Ziegler, Y. S., Touchstone, B., Bradley, W. A., and Cook, R. G. (1986). hv1 is an evolutionarily conserved H2A variant that is preferentially associated with active genes. *Journal of Biological Chemistry*, 261(4):1941–1948.

Altaf, M., Auger, A., Monnet-Saksouk, J., Brodeur, J., Piquet, S., Cramet, M., Bouchard, N., Lacoste, N., Utley, R. T., Gaudreau, L., and Cote, J. (2010). NuA4-dependent Acetylation of Nucleosomal Histones H4 and H2A Directly Stimulates Incorporation of H2A.Z by the SWR1 Complex. *Journal of Biological Chemistry*, 285(21):15966–15977.

Amat, R. and Gudas, L. J. (2011). RARγ is required for correct deposition and removal of Suz12 and H2A.Z in embryonic stem cells. *Journal of cellular physiology*, 226(2):293–298.

Andrews, P. W. (1984). Retinoic acid induces neuronal differentiation of a cloned human embryonal carcinoma cell line in vitro. *Developmental biology*.

Andrews, P. W. and Andrews, P. W. (2002). From teratocarcinomas to embryonic stem cells. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1420):405–417.

Auger, A., Galarneau, L., Altaf, M., Nourani, A., Doyon, Y., Utley, R. T., Cronier, D., Allard, S., and Côté, J. (2008). Eaf1 is the platform for NuA4 molecular assembly that evolutionarily links chromatin acetylation to ATP-dependent exchange of histone H2A variants. *Molecular and cellular biology*, 28(7):2257–2270.

Babiarz, J. E. (2006). Telomeric heterochromatin boundaries require NuA4-dependent acetylation of histone variant H2A.Z in Saccharomy cescerevisiae. *Genes & development*, 20(6):700–710.

Bachvarova, R. F., Masi, T., Drum, M., Parker, N., Mason, K., Patient, R., and Johnson, A. D. (2004). Gene expression in the axolotl germ line: Axdazl, Axvh, Axoct-4, and Axkit. *Developmental dynamics : an official publication of the American Association of Anatomists*, 231(4):871–880.

Banaszynski, L. A., Allis, C. D., and Lewis, P. W. (2010). Histone Variants in Metazoan Development. *Developmental cell*, 19(5):662–674.

Bao, Y. and Shen, X. (2007). INO80 subfamily of chromatin remodeling complexes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 618(1-2):18–29.

Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R. M., Ramaswamy, S., and Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous iPSC generation. *Nature methods*, 11(11):1170–1176.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, NY)*, 286(5439):509–512.

Barrero, M. J., Sese, B., Kuebler, B., Bilic, J., Boué, S., Martí, M., and Izpisúa Belmonte, J. C. (2013). Macrohistone Variants Preserve Cell Identity by Preventing the Gain of H3K4me2 during Reprogramming to Pluripotency. *Cell Reports*, 3(4):1005–1011.

Batlle-Morera, L., Smith, A., and Nichols, J. (2008). Parameters influencing derivation of embryonic stem cells from murine embryos. *Genesis (New York, NY : 2000)*, 46(12):758–767.

Bellucci, L., Dalvai, M., Kocanova, S., Moutahir, F., and Bystricky, K. (2013). Activation of p21 by HDAC Inhibitors Requires Acetylation of H2A.Z. *PloS one*, 8(1):e54102.

Benham, F. J., Andrews, P. W., Knowles, B. B., Bronson, D. L., and Harris, H. (1981). Alkaline phosphatase isozymes as possible markers of differentiation in human testicular teratocarcinoma cell lines. *Developmental biology*, 88(2):279–287.

Berstine, E. G., Hooper, M. L., Grandchamp, S., and Ephrussi, B. (1973). Alkaline phosphatase activity in mouse teratoma. *Proceedings of the National Academy of Sciences*, 70(12):3899–3903.

Bertone, P., Hendrich, B., and Silva, J. C. R. (2015). Mbd3 and deterministic reprogramming. *BioRXiv*, page 013904.

Bian, Y., Alberio, R., Allegrucci, C., Campbell, K. H., and Johnson, A. D. (2009). Epigenetic marks in somatic chromatin are remodelled to resemble pluripotent nuclei by amphibian oocyte extracts. *Epigenetics : official journal of the DNA Methylation Society*, 4(3):194–202.

Binda, O., Sevilla, A., Leroy, G., Lemischka, I. R., Garcia, B. A., and Richard, S. (2013). SETD6 monomethylates H2AZ on lysine 7 and is required for the maintenance of embryonic stem cell self-renewal. *Epigenetics : official journal of the DNA Methylation Society*, 8(2):177–183.

Blelloch, R. H., Hochedlinger, K., Yamada, Y., Brennan, C., Kim, M., Mintz, B., Chin, L., and Jaenisch, R. (2004). Nuclear cloning of embryonal carcinoma cells. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13985–13990.

Bonenfant, D., Towbin, H., Coulot, M., Schindler, P., Mueller, D. R., and van Oostrum, J. (2007). Analysis of Dynamic Changes in Post-translational Modifications of Human Histones during Cell Cycle by Mass Spectrometry. *Molecular & Cellular Proteomics*, 6(11):1917–1932.

Boroviak, T., Loos, R., Lombard, P., Okahara, J., Behr, R., Sasaki, E., Nichols, J., Smith, A., and Bertone, P. (2015). Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Developmental cell*, 35(3):366–382.

Boskovic, A., Bender, A., Gall, L., Ziegler-Birling, C., Beaujean, N., and Torres-Padilla, M. E. (2012). Analysis of active chromatin modifications in early mammalian embryos reveals uncoupling of H2A.Z acetylation and H3K36 trimethylation from embryonic genome activation. *Epigenetics : official journal of the DNA Methylation Society*, 7(7):747–757.

Bourke, L. M., del Monte-Nieto, G., Outhwaite, J. E., Bharti, V., Pollock, P. M., Simmons, D. G., Adam, A., Hur, S. S. J., Maghzal, G. J., Whitelaw, E., Stocker, R., Suter, C. M., Harvey, R. P., and Harten, S. K. (2017). Loss of Rearranged L-Myc Fusion (RLF) results in defects in heart development in the mouse. *Differentiation; research in biological diversity*, 94:8–20.

Boyne, M. T., Pesavento, J. J., Mizzen, C. A., and Kelleher, N. L. (2006). Precise characterization of human histones in the H2A gene family by top down mass spectrometry. *Journal of proteome research*, 5(2):248–253.

Bradley, A., Evans, M., Kaufman, M. H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, 309(5965):255–256.

Briggs, R. and King, T. J. (1952). Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proceedings of the National Academy of Sciences*, 38(5):455–463.

Brinster, R. L. (1974). The effect of cells transferred into the mouse blastocyst on subsequent development. *The Journal of experimental medicine*, 140(4):1049–1056.

Brons, I. G. M., Smithers, L. E., Trotter, M. W. B., Rugg-Gunn, P. J., Sun, B., de Sousa Lopes, S. M. C., Howlett, S. K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. A., and Vallier, L. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448(7150):191–195.

Brook, F. A. and Gardner, R. L. (1997). The origin and efficient derivation of embryonic stem cells in the mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 94(11):5709–5712.

Bruce, K., Myers, F. A., Mantouvalou, E., Lefevre, P., Greaves, I., Bonifer, C., Tremethick, D. J., Thorne, A. W., and Crane-Robinson, C. (2005). The replacement histone H2A.Z in a hyperacetylated form is a feature of active genes in the chicken. *Nucleic acids research*, 33(17):5633–5639.

Buecker, C., Chen, H.-H., Polo, J. M., Daheron, L., Bu, L., Barakat, T. S., Okwieka, P., Porter, A., Gribnau, J., Hochedlinger, K., and Geijsen, N. (2010). A Murine ESC-like State Facilitates Transgenesis and Homologous Recombination in Human Pluripotent Stem Cells. *Cell stem cell*, 6(6):535–546.

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(7):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(6):1209–1222.

Buganim, Y., Faddah, D. A., and Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. *Nature reviews Genetics*, 14(6):427–439.

Cai, Y. (2005). The Mammalian YL1 Protein Is a Shared Subunit of the TRRAP/TIP60 Histone Acetyltransferase and SRCAP Complexes. *Journal of Biological Chemistry*, 280(14):13665–13670.

Cai, Y., Jin, J., Tomomori-Sato, C., Sato, S., Sorokina, I., Parmely, T. J., Conaway, R. C., and Conaway, J. W. (2003). Identification of new subunits of the multiprotein mammalian TRRAP/TIP60-containing histone acetyltransferase complex. *Journal of Biological Chemistry*, 278(44):42733–42736.

Campbell, P. A. and Rudnicki, M. A. (2013). Oct4 Interaction with Hmgb2 Regulates Akt Signaling and Pluripotency. *Stem cells (Dayton, Ohio)*, 31(6):1107–1120.

Cao, R., Tsukada, Y.-I., and Zhang, Y. (2005). Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Molecular cell*, 20(6):845–854.

Capecchi, M. R. (2008). *The making of a scientist II (Nobel Lecture).*, volume 9.

Carey, B. W., Markoulaki, S., Hanna, J., Saha, K., Gao, Q., Mitalipova, M., and Jaenisch, R. (2009). Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proceedings of the National Academy of Sciences*, 106(1):157–162.

Carey, B. W., Markoulaki, S., Hanna, J. H., Faddah, D. A., Buganim, Y., Kim, J., Ganz, K., Steine, E. J., Cassady, J. P., Creyghton, M. P., Welstead, G. G., Gao, Q., and Jaenisch, R. (2011). Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell stem cell*, 9(6):588–598.

Carr, A. M., Dorrington, S. M., Hindley, J., Phear, G. A., Aves, S. J., and Nurse, P. (1994). Analysis of a histone H2A variant from fission yeast: evidence for a role in chromosome stability. *Molecular & general genetics : MGG*, 245(5):628–635.

Carter, A. C., Davis-Dusenbery, B. N., Koszka, K., Ichida, J. K., and Eggan, K. (2014). Nanog-independent reprogramming to iPSCs with canonical factors. *Stem Cell Reports*, 2(2):119–126.

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. 113(5):643–655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007a). Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007b). Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234.

Chan, H.-M., Narita, M., Lowe, S. W., and Livingston, D. M. (2005). The p400 E1A-associated protein is a novel component of the p53 –> p21 senescence pathway. *Genes & development*, 19(2):196–201.

Chan, Y.-S., Göke, J., Ng, J.-H., Lu, X., Gonzales, K. A. U., Tan, C.-P., Tng, W.-Q., Hong, Z.-Z., Lim, Y.-S., and Ng, H.-H. (2013). Induction of a Human Pluripotent State with Distinct Regulatory Circuitry that Resembles Preimplantation Epiblast. *Cell stem cell*, 13(6):663–675.

Chantzoura, E., Skylaki, S., Menendez, S., Kim, S.-I., Johnsson, A., Linnarsson, S., Woltjen, K., Chambers, I., and Kaji, K. (2015). Reprogramming Roadblocks Are System Dependent. *Stem Cell Reports*, 5(3):350–364.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013a). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128.

Chen, P. B., Hung, J.-H., Hickman, T. L., Coles, A. H., Carey, J. F., Weng, Z., Chu, F., and Fazzio, T. G. (2013b). Hdac6 regulates Tip60-p400 function in stem cells. *eLife*, 2(0):e01557.

Cheng, L., Hansen, N. F., Zhao, L., Du, Y., Zou, C., Donovan, F. X., Chou, B.-K., Zhou, G., Li, S., Dowey, S. N., Ye, Z., Chandrasekharappa, S. C., Yang, H., Mullikin, J. C., and Liu, P. P. (2012). Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. *Cell stem cell*, 10(3):337–344.

Chevillard-Briet, M., Quaranta, M., Grezy, A., Mattera, L., Courilleau, C., Philippe, M., Mercier, P., Corpet, D., Lough, J., Ueda, T., Fukunaga, R., Trouche, D., and Escaffit, F. (2014). Interplay between chromatin-modifying enzymes controls colon cancer progression through Wnt signaling. *Human molecular genetics*, 23(8):2120–2131.

Chin, M. H., Mason, M. J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., Khvorostov, I., Ott, V., Grunstein, M., Lavon, N., Benvenisty, N., Croce, C. M., Clark, A. T., Baxter, T., Pyle, A. D., Teitell, M. A., Pelegrini, M., Plath, K., and Lowry, W. E. (2009). Induced Pluripotent Stem Cells and Embryonic Stem Cells Are Distinguished by Gene Expression Signatures. *Cell stem cell*, 5(1):111–123.

Choi, J., Heo, K., and An, W. (2009). Cooperative action of TIP48 and TIP49 in H2A.Z exchange catalyzed by acetylation of nucleosomal H2A. *Nucleic acids research*, 37(18):5993–6007.

Choi, J., Lee, S., Mallard, W., Clement, K., Tagliazucchi, G. M., Lim, H., Choi, I. Y., Ferrari, F., Tsankov, A. M., Pop, R., Lee, G., Rinn, J. L., Meissner, A., Park, P. J., and Hochedlinger, K. (2015). A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. *Nature biotechnology*, 33(11):1173–1181.

Clarkson, M. J., Saint, R., Tremethick, D. J., Wells, J. R., Gibson, F., Clarkson, M. J., Wells, J. R. E., Saint, R., Gibson, F., and Tremethick, D. J. (1999). Regions of variant histone His2AvD required for Drosophila development. *Nature*, 399(6737):694–697.

Collins, F. S., Rossant, J., and Wurst, W. (2007). A mouse for all reasons. *Cell*, 128(1):9–13.

Conerly, M. L., Teves, S. S., Diolaiti, D., Ulrich, M., Eisenman, R. N., and Henikoff, S. (2010). Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. *Genome research*, 20(10):1383–1390.

Costa, Y., Ding, J., Theunissen, T. W., Faiola, F., Hore, T. A., Shliaha, P. V., Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S., Das, S., Levasseur, D. N., Li, Z., Xu, M., Reik, W., Silva, J. C. R., and Wang, J. (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature*, 495(7441):370–374.

Creyghton, M. P., Markoulaki, S., Levine, S. S., Hanna, J., Lodato, M. A., Sha, K., Young, R. A., Jaenisch, R., and Boyer, L. A. (2008). H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is Necessary for Lineage Commitment. *Cell*, 135(4):649–661.

Csányi, G. and Szendrői, B. (2004). Structure of a large social network. *Physical Review E*, 69(3):036131.

Cui, K., Zang, C., Roh, T.-Y., Schones, D. E., Childs, R. W., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell stem cell*, 4(1):80–93.

Dahl, J. A., Jung, I., Aanes, H., Greggains, G. D., Manaf, A., Lerdrup, M., Li, G., Kuan, S., Li, B., Lee, A. Y., Preissl, S., Jermstad, I., Haugen, M. H., Suganthan, R., Bjørås, M., Hansen, K., Dalen, K. T., Fedorcsak, P., Ren, B., and Klungland, A. (2016). Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature*, 537(7621):548–552.

Dalvai, M., Bellucci, L., Fleury, L., Lavigne, A.-C., Moutahir, F., and Bystricky, K. (2012). H2A.Z-dependent crosstalk between enhancer and promoter regulates Cyclin D1 expression. *Oncogene*.

Dalvai, M., Fleury, L., Bellucci, L., Kocanova, S., and Bystricky, K. (2013). TIP48/Reptin and H2A.Z requirement for initiating chromatin remodeling in estrogen-activated transcription. *PLoS genetics*, 9(4):e1003387.

Davidson, K. C., Mason, E. A., and Pera, M. F. (2015). The pluripotent state in mouse and human. *Development*, 142(18):3090–3099.

Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6):987–1000.

De Los Angeles, A., Ferrari, F., Xi, R., Fujiwara, Y., Benvenisty, N., Deng, H., Hochedlinger, K., Jaenisch, R., Lee, S., Leitch, H. G., Lensch, M. W., Lujan, E., Pei, D., Rossant, J., Wernig, M., Park, P. J., and Daley, G. Q. (2015). Hallmarks of pluripotency. *Nature*, 525(7570):469–478.

De Napoles, M., Mermoud, J. E., Wakao, R., Tang, Y. A., Endoh, M., Appanah, R., Nesterova, T. B., Silva, J. C. R., Otte, A. P., and Vidal, M. (2004). Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation. *Developmental cell*, 7(5):663–676.

Deneault, E., Cellot, S., Faubert, A., Laverdure, J. P., and al, e. (2009). A functional screen to identify novel effectors of hematopoietic stem cell activity. *Cell*.

Dhillon, N., Oki, M., Szyjka, S. J., Aparicio, O. M., and Kamakaka, R. T. (2006). H2A.Z functions to regulate progression through the cell cycle. *Molecular and cellular biology*, 26(2):489–501.

Diwan, S. B. and Stevens, L. C. (1976). Development of teratomas from the ectoderm of mouse egg cylinders. *Journal of the National Cancer Institute*, 57(4):937–942.

Dixon, J. E., Allegrucci, C., Redwood, C., Kump, K., Bian, Y., Chatfield, J., Chen, Y.-H., Sottile, V., Voss, S. R., Alberio, R., and Johnson, A. D. (2010). Axolotl Nanog activity in mouse embryonic stem cells demonstrates that ground state pluripotency is conserved from urodele amphibians to mammals. *Development*, 137(18):2973–2980.

Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G. Q., and Feinberg, A. P. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics*, 41(12):1350–1353.

dos Santos, R. L., Tosti, L., Radzisheuskaya, A., Caballero, I. M., Kaji, K., Hendrich, B., and Silva, J. C. R. (2014). MBD3/NuRD Facilitates Induction of Pluripotency in a Context-Dependent Manner. *Cell stem cell*, 15(1):102–110.

Doyon, Y., Selleck, W., Lane, W. S., Tan, S., and Côté, J. (2004). Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans. *Molecular and cellular biology*, 24(5):1884–1896.

Draker, R. and Cheung, P. (2009). Transcriptional and epigenetic functions of histone variant H2A. Z. *Biochemistry and Cell Biology*, 87(1):19–25.

Dryhurst, D., Ishibashi, T., Rose, K. L., Eirín-López, J. M., McDonald, D., Silva-Moreno, B., Veldhoen, N., Helbing, C. C., Hendzel, M. J., Shabanowitz, J., Hunt, D. F., and Ausió, J. (2009). Characterization of the histone H2A.Z-1 and H2A.Z-2 isoforms in vertebrates. *BMC biology*, 7(1):86–16.

Dryhurst, D., Thambirajah, A. A., and Ausió, J. (2004). New twists on H2A.Z: a histone variant with a controversial structural and functional past. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 82(4):490–497.

Dunican, D. S., McWilliam, P., Tighe, O., Parle-McDermott, A., and Croke, D. T. (2002). Gene expression differences between the microsatellite instability (MIN) and chromosomal instability (CIN) phenotypes in colorectal cancer revealed by high-density cDNA array hybridization. *Oncogene*, 21(20):3253–3257.

Durant, M. and Pugh, B. F. (2007). NuA4-directed chromatin transactions throughout the Saccharomyces cerevisiae genome. *Molecular and cellular biology*, 27(15):5327–5335.

Durcova-Hills, G., Adams, I. R., Barton, S. C., Surani, M. A., and McLaren, A. (2006). The role of exogenous fibroblast growth factor-2 on the reprogramming of primordial germ cells into pluripotent stem cells. *Stem cells (Dayton, Ohio)*, 24(6):1441–1449.

Ebrahimi, B. (2015). Reprogramming barriers and enhancers: strategies to enhance the efficiency and kinetics of induced pluripotency. *Cell regeneration (London, England)*, 4(1):10.

Eirín-López, J. A. J. (2007). H2A.Z-Mediated Genome-Wide Chromatin Specialization. *Current Genomics*, 8(1):59.

Eissenberg, J. C., Wong, M., and Chrivia, J. C. (2005). Human SRCAP and Drosophila melanogaster DOM are homologs that function in the notch signaling pathway. *Molecular and cellular biology*, 25(15):6559–6569.

Evans, M. (2011). Discovering pluripotency: 30 years of mouse embryonic stem cells. *Nature reviews Molecular cell biology*, 12(10):680–686.

Evans, M. J. (1972). The isolation and properties of a clonal tissue culture strain of pluripotent mouse teratoma cells. *Journal of embryology and experimental morphology*, 28(1):163–176.

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

Faast, R., Thonglairoam, V., Schulz, T. C., Beall, J., Wells, J. R., Taylor, H., Matthaei, K., Rathjen, P. D., Tremethick, D. J., and Lyons, I. (2001). Histone variant H2A.Z is required for early mammalian development. *Current biology : CB*, 11(15):1183–1187.

Falco, G., Lee, S.-L., Stanghellini, I., Bassey, U. C., Hamatani, T., and Ko, M. S. H. (2007). Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental biology*, 307(2):539–550.

Fan, J. Y., Gordon, F., Luger, K., Hansen, J. C., and Tremethick, D. J. (2002). The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. *Nature structural & molecular biology*, 9(3):172–176.

Fan, J. Y., Rangasamy, D., Luger, K., and Tremethick, D. J. (2004). H2A.Z Alters the Nucleosome Surface to Promote HP1α-Mediated Chromatin Fiber Folding. *Molecular cell*, 16(4):655–661.

Fang, R., Liu, K., Zhao, Y., Li, H., Zhu, D., Du, Y., Xiang, C., Li, X., Liu, H., Miao, Z., Zhang, X., Shi, Y., Yang, W., Xu, J., and Deng, H. (2014). Generation of naive induced pluripotent stem cells from rhesus monkey fibroblasts. *Cell stem cell*, 15(4):488–496.

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(1):162–174.

Fenderson, B. A., Andrews, P. W., Nudelman, E., Clausen, H., and Hakomori, S. (1987). Glycolipid core structure switching from globo- to lacto- and ganglioseries during retinoic acid-induced differentiation of TERA-2-derived human embryonal carcinoma cells. *Developmental biology*, 122(1):21–34.

Festuccia, N., Osorno, R., Halbritter, F., Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F., Yates, A., Tomlinson, S. R., and Chambers, I. (2012). Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell stem cell*, 11(4):477–490.

Fidalgo, M., Fidalgo, M., Faiola, F., Faiola, F., Pereira, C. F., Pereira, C. F., Ding, J., Ding, J., Saunders, A., Saunders, A., Gingold, J., Gingold, J., Schaniel, C., Schaniel, C., Lemischka, I. R., Lemischka, I. R., Silva, J. C. R., Silva, J. C. R., Wang, J., and Wang, J. (2012). Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming. *Proceedings of the National Academy of Sciences*, 109(40):16202–16207.

Frank, S. R., Parisi, T., Taubert, S., Fernandez, P., Fuchs, M., Chan, H.-M., Livingston, D. M., and Amati, B. (2003). MYC recruits the TIP60 histone acetyltransferase complex to chromatin. *EMBO reports*, 4(6):575–580.

Frank, S. R., Schroeder, M., Fernandez, P., Taubert, S., and Amati, B. (2001). Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. *Genes & development*, 15(16):2069–2082.

Fritsch, O., Benvenuto, G., Bowler, C., Molinier, J., and Hohn, B. (2004). The INO80 protein controls homologous recombination in Arabidopsis thaliana. *Molecular cell*, 16(3):479–485.

Fujii, T., Ueda, T., Nagata, S., and Fukunaga, R. (2010). Essential role of p400/mDomino chromatin-remodeling ATPase in bone marrow hematopoiesis and cell-cycle progression. *Journal of Biological Chemistry*, 285(39):30214–30223.

Fusaki, N., Ban, H., Nishiyama, A., Saeki, K., and Hasegawa, M. (2009). Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proceedings of the Japan Academy, Series B*, 85(8):348–362.

Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., Rais, Y., Shipony, Z., Mukamel, Z., Krupalnik, V., Zerbib, M., Geula, S., Caspi, I., Schneir, D., Shwartz, T., Gilad, S., Amann-Zalcenstein, D., Benjamin, S., Amit, I., Tanay, A., Massarwa, R., Novershtern, N., and Hanna, J. H. (2013). Derivation of novel human ground state naive pluripotent stem cells. *Nature*, 504(7479):282–286.

Gagliardi, A., Mullin, N. P., Ying Tan, Z., Colby, D., Kousa, A. I., Halbritter, F., Weiss, J. T., Felker, A., Bezstarosti, K., Favaro, R., Demmers, J., Nicolis, S. K., Tomlinson, S. R., Poot, R. A., and Chambers, I. (2013). A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *The EMBO journal*, 32(16):2231–2247.

Galarneau, L., Nourani, A., Boudreault, A. A., Zhang, Y., Heliot, L., Allard, S., Savard, J., Lane, W. S., Stillman, D. J., and Cote, J. (2000). Multiple links between the NuA4 histone acetyltransferase complex and epigenetic control of transcription. *Molecular cell*, 5(6):927–937.

Gallant-Behm, C. L., Ramsey, M. R., Bensard, C. L., Nojek, I., Tran, J., Liu, M., Ellisen, L. W., and Espinosa, J. M. (2012). ΔNp63α represses anti-proliferative genes via H2A.Z deposition. *Genes & development*, 26(20):2325–2336.

Gao, S., Chung, Y.-G., Parseghian, M. H., King, G. J., Adashi, E. Y., and Latham, K. E. (2004). Rapid H1 linker histone transitions following fertilization or somatic cell nuclear transfer: evidence for a uniform developmental program in mice. *Developmental biology*, 266(1):62–75.

Gao, S., Zheng, C., Chang, G., Liu, W., Kou, X., Tan, K., Tao, L., Xu, K., Wang, H., Cai, J., Tian, J., and Gao, S. (2015). Unique features of mutations revealed by sequentially reprogrammed induced pluripotent stem cells. *Nature communications*, 6:6318.

Gao, Y., Chen, J., Li, K., Wu, T., Huang, B., Liu, W., Kou, X., Zhang, Y., Huang, H., Jiang, Y., Yao, C., Liu, X., Lu, Z., Xu, Z., Kang, L., Chen, J., Wang, H., Cai, T., and Gao, S. (2013). Replacement of Oct4 by Tet1 during iPSC Induction

Reveals an Important Role of DNA Methylation and Hydroxymethylation in Reprogramming. *Stem Cell*, 12(4):453–469.

Gardner, R. L. and Brook, F. A. (1997). Reflections on the biology of embryonic stem (ES) cells. *The International journal of developmental biology*, 41(2):235–243.

Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nature reviews Molecular cell biology*, 12(1):36–47.

Gaspar-Maia, A., Alajem, A., Polesso, F., Sridharan, R., Mason, M. J., Heidersbach, A., Ramalho-Santos, J., McManus, M. T., Plath, K., Meshorer, E., and Ramalho-Santos, M. (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature*, 460(7257):863–868.

Gaspar-Maia, A., Qadeer, Z. A., Hasson, D., Ratnakumar, K., Leu, N. A., Leroy, G., Liu, S., Costanzi, C., Valle-Garcia, D., Schaniel, C., Lemischka, I., Garcia, B., Pehrson, J. R., and Bernstein, E. (2013). MacroH2A histone variants act as a barrier upon reprogramming towards pluripotency. *Nature communications*, 4:1565.

Geijsen, N., Horoschak, M., Kim, K., Gribnau, J., Eggan, K., and Daley, G. Q. (2003). Derivation of embryonic germ cells and male gametes from embryonic stem cells. *Nature cell biology*, 427(6970):148–154.

Gervais, A. L. and Gaudreau, L. (2009). Discriminating nucleosomes containing histone H2A.Z or H2A based on genetic and epigenetic information. *BMC molecular biology*, 10(1):18.

Gevry, N., Gévry, N., Chan, H. M., Chan, H.-M., Laflamme, L., Laflamme, L., Livingston, D. M., Livingston, D. M., Gaudreau, L., and Gaudreau, L. (2007). p21 transcription is regulated by differential localization of histone H2A.Z. *Genes & development*, 21(15):1869–1881.

Ghosh, Z., Wilson, K. D., Wu, Y., Hu, S., Quertermous, T., and Wu, J. C. (2010). Persistent Donor Cell Gene Expression among Human Induced Pluripotent Stem Cells Contributes to Differences with Human Embryonic Stem Cells. *PloS one*, 5(2):e8975.

Gillespie, P. J. (2000). Nucleoplasmin-mediated chromatin remodelling is required for Xenopus sperm nuclei to become licensed for DNA replication. *Nucleic acids research*, 28(2):472–480.

Gkountela, S., Li, Z., Vincent, J. J., Zhang, K. X., Chen, A., Pellegrini, M., and Clark, A. T. (2013). The ontogeny of cKIT+ human primordial germ cells proves to be a resource for human germ line reprogramming, imprint erasure and in vitro differentiation. *Nature cell biology*, 15(1):113–122.

Gorrini, C., Squatrito, M., Luise, C., Syed, N., Perna, D., Wark, L., Martinato, F., Sardella, D., Verrecchia, A., Bennett, S., Confalonieri, S., Cesaroni, M., Marchesi, F., Gasco, M., Scanziani, E., Capra, M., Mai, S., Nuciforo, P., Crook, T., Lough, J., and Amati, B. (2007). Tip60 is a haplo-insufficient tumour suppressor required for an oncogene-induced DNA damage response. *Nature*, 448(7157):1063–1067.

Grant, P. A., Schieltz, D., Pray-Grant, M. G., Yates, J. R., and Workman, J. L. (1998). The ATM-related cofactor Tra1 is a component of the purified SAGA complex. *Molecular cell*, 2(6):863–867.

Greaves, I. K., Rangasamy, D., Devoy, M., Graves, J. A. M., and Tremethick, D. J. (2006). The X and Y chromosomes assemble into H2A.Z-containing [corrected] facultative heterochromatin [corrected] following meiosis. *Molecular and cellular biology*, 26(14):5394–5405.

Greaves, I. K., Rangasamy, D., Ridgway, P., and Tremethick, D. J. (2007). H2A.Z contributes to the unique 3D structure of the centromere. *Proceedings of the National Academy of Sciences of the United States of America*, 104(2):525–530.

Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S.-g., Iqbal, K., Shi, Y. G., Deng, Z., Szabó, P. E., Pfeifer, G. P., Li, J., and Xu, G.-L. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366):606–610.

Guillemette, B., Bataille, A. R., Gévry, N., Adam, M., Blanchette, M., Robert, F., and Gaudreau, L. (2005). Variant Histone H2A.Z Is Globally Localized to the Promoters of Inactive Yeast Genes and Regulates Nucleosome Positioning. *PLoS biology*, 3(12):e384.

Guillemette, B. and Gaudreau, L. (2006). Reuniting the contrasting functions of H2A.Z. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 84(4):528–535.

Guo, G., von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A., and Nichols, J. (2016). Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports*, 6(4):437–446.

Guo, G., Yang, J., Nichols, J., Hall, J. S., Eyres, I., Mansfield, W., and Smith, A. (2009). Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development*, 136(7):1063–1069.

Gurdon, J. B. (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of embryology and experimental morphology*, 10:622–640.

Gurdon, J. B., Elsdale, T. R., and Fischberg, M. (1958). Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. *Nature*, 182(4627):64–65.

Hackett, J. A. and Surani, M. A. (2014). Regulatory Principles of Pluripotency: From the Ground State Up. *Cell stem cell*, 15(4):416–430.

Hager, G. L., McNally, J. G., and Misteli, T. (2009). Transcription dynamics. *Molecular cell*, 35(6):741–753.

Hahn, S. and Hermeking, H. (2014). ZNF281/ZBP-99: a new player in epithelial-mesenchymal transition, stemness, and cancer. *Journal of Molecular Medicine*, 92(6):571–581.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U. C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M. A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature*, 452(7189):877–881.

Hall, J., Guo, G., Wray, J., Eyres, I., Nichols, J., Grotewold, L., Morfopoulou, S., Humphreys, P., Mansfield, W., Walker, R., Tomlinson, S., and Smith, A. (2009). Oct4 and LIF/Stat3 Additively Induce Krüppel Factors to Sustain Embryonic Stem Cell Self-Renewal. *Cell stem cell*, 5(6):597–609.

Halley-Stott, R. P., Pasque, V., Astrand, C., Miyamoto, K., Simeoni, I., Jullien, J., and Gurdon, J. B. (2010). Mammalian nuclear transplantation to Germinal Vesicle stage Xenopus oocytes – A method for quantitative transcriptional reprogramming. *Methods (San Diego, Calif)*, 51(1):56–65.

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. *Proceedings of the National Academy of Sciences*, 107(20):9222–9227.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C. J., Creyghton, M. P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration . *Nature*, 462(7273):595–601.

Hansson, J., Rafiee, M. R., Reiland, S., Polo, J. M., Gehring, J., Okawa, S., Huber, W., Hochedlinger, K., and Krijgsveld, J. (2012). Highly Coordinated Proteome Dynamics during Reprogramming of Somatic Cells to Pluripotency. *Cell Reports*, 2(6):1579–1592.

Hardy, S., Jacques, P.-É., Gévry, N., Forest, A., Fortin, M.-È., Laflamme, L., Gaudreau, L., and Robert, F. (2009). The Euchromatic and Heterochromatic Landscapes Are Shaped by Antagonizing Effects of Transcription on H2A.Z Deposition. *PLoS genetics*, 5(10):e1000687.

Harvey, R. P., Whiting, J. A., Coles, L. S., Krieg, P. A., and Wells, J. R. (1983). H2A.F: an extremely variant histone H2A sequence expressed in the chicken embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 80(10):2819–2823.

Henderson, J. K., Draper, J. S., Baillie, H. S., Fishel, S., Thomson, J. A., Moore, H., and Andrews, P. W. (2002). Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens. *Stem cells (Dayton, Ohio)*, 20(4):329–337.

Herceg, Z., Hulla, W., Gell, D., Cuenin, C., Lleonart, M., Jackson, S., and Wang, Z.-Q. (2001). Disruption of Trrap causes early embryonic lethality and defects in cell cycle progression. *Nature genetics*, 29(2):206–211.

Hirata, T., Amano, T., Nakatake, Y., Amano, M., Piao, Y., Hoang, H. G., and Ko, M. S. H. (2012). Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. *Scientific reports*, 2(1):srep00208.

Hochedlinger, K. and Jaenisch, R. (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature*, 415(6875):1035–1038.

Horvath, S. (2011). Weighted Network Analysis: Applications in Genomics and Systems Biology. Springer-Verlag New York.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., and Zhao, K. (2013). H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell stem cell*, 12(2):180–192.

Hu, Y., Fisher, J. B., Koprowski, S., McAllister, D., Kim, M.-S., and Lough, J. (2009). Homozygous disruption of the tip60gene causes early embryonic lethality. *Developmental dynamics : an official publication of the American Association of Anatomists*, 238(11):2912–2921.

Hua, S., Kallen, C. B., Dhar, R., Baquero, M. T., Mason, C. E., Russell, B. A., Shah, P. K., Liu, J., Khramtsov, A., Tretiakova, M. S., Krausz, T. N., Olopade, O. I., Rimm, D. L., and White, K. P. (2008). Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression. *Molecular Systems Biology*, 4:1587–14.

Huang, K., Maruyama, T., and Fan, G. (2014a). The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. *Cell stem cell*, 15(4):410–415.

Huang, X., Tian, C., Liu, M., Wang, Y., Tolmachev, A. V., Sharma, S., Yu, F., Fu, K., Zheng, J., and Ding, S.-J. (2012). Quantitative Proteomic Analysis of Mouse Embryonic Fibroblasts and Induced Pluripotent Stem Cells Using 16O/ 18O Labeling. *Journal of proteome research*, 11(4):2091–2102.

Huang, Y., Chavez, L., Chang, X., Wang, X., Pastor, W. A., Kang, J., Zepeda-Martínez, J. A., Pape, U. J., Jacobsen, S. E., Peters, B., and Rao, A. (2014b). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):1361–1366.

Hussein, S. M., Batada, N. N., Vuoristo, S., Ching, R. W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., Lundin, K., Mikkola, M., Trokovic, R., Peitz, M., Brüstle, O., Bazett-Jones, D. P., Alitalo, K., Lahesmaa, R., Nagy, A., and Otonkoski, T. (2011). Copy number variation and selection during reprogramming to pluripotency. *Nature*, 471(7336):58–62.

Ikura, T., Ogryzko, V. V., Grigoriev, M., Groisman, R., Wang, J., Horikoshi, M., Scully, R., Qin, J., and Nakatani, Y. (2000). Involvement of the TIP60 histone acetylase complex in DNA repair and apoptosis. *Cell*, 102(4):463–473.

Illingworth, R. S., Botting, C. H., Grimes, G. R., Bickmore, W. A., and Eskeland, R. (2012). PRC1 and PRC2 are not required for targeting of H2A.Z to developmental genes in embryonic stem cells. *PloS one*, 7(4):e34848.

Illmensee, K. and Mintz, B. (1976). Totipotency and normal differentiation of single teratocarcinoma cells cloned by injection into blastocysts. *Proceedings of the National Academy of Sciences*, 73(2):549–553.

Iouzalen, N., Moreau, J., and Mechali, M. (1996). H2A.ZI, a new variant histone expressed during Xenopus early development exhibits several distinct features from the core histone H2A. *Nucleic acids research*, 24(20):3947–3952.

Isbel, L., Srivastava, R., Oey, H., Spurling, A., Daxinger, L., Puthalakath, H., and Whitelaw, E. (2015). Trim33 Binds and Silences a Class of Young Endogenous Retroviruses in the Mouse Testis; a Novel Component of the Arms Race between Retrotransposons and the Host Genome. *PLoS genetics*, 11(12):e1005693.

Ishibashi, T., Dryhurst, D., Rose, K. L., Shabanowitz, J., and al, e. (2009). Acetylation of Vertebrate H2A.Z and Its Effect on the Structure of the Nucleosome. *Biochemistry*, 48(22):5007–5017.

Jakob, H., Dubois, P., Eisen, H., and Jacob, F. (1978). *Effets de l'hexaméthylènebisacétamide sur la différenciation de cellules de carcinome embryonnaire.* CR Acad Sci Hebd Seances Acad Sci D.

Jensen, K., Santisteban, M. S., Urekar, C., and Smith, M. M. (2011). Histone H2A.Z acid patch residues required for deposition and function. *Molecular genetics and genomics : MGG*, 285(4):287–296.

Jiang, J., Lv, W., Ye, X., Wang, L., Zhang, M., Yang, H., Okuka, M., Zhou, C., Zhang, X., Liu, L., and Li, J. (2013). Zscan4 promotes genomic stability during reprogramming and dramatically improves the quality of iPS cells as demonstrated by tetraploid complementation. *Cell research*, 23(1):92–106.

Jin, X. L. and O'Neill, C. (2010). The Presence and Activation of Two Essential Transcription Factors (cAMP Response Element-Binding Protein and cAMP-Dependent Transcription Factor ATF1) in the Two-Cell Mouse Embryo. *Biology of reproduction*, 82(2):459–468.

Jin, X. L. and O'Neill, C. (2011). Regulation of the Expression of Proto-Oncogenes by Autocrine Embryotropins in the Early Mouse Embryo1. *Biology of reproduction*, 84(6):1216–1224.

Jin, X. L. and O'Neill, C. (2014). The regulation of the expression and activation of the essential ATF1 transcription factor in the mouse preimplantation embryo. *Reproduction (Cambridge, England)*, 148(2):147–157.

Johannesson, B., Sagi, I., Gore, A., Paull, D., Yamada, M., Golan-Lev, T., Li, Z., LeDuc, C., Shen, Y., Stern, S., Xu, N., Ma, H., Kang, E., Mitalipov, S., Sauer, M. V., Zhang, K., Benvenisty, N., and Egli, D. (2014). Comparable frequencies of coding mutations and loss of imprinting in human pluripotent cells derived by nuclear transfer and defined factors. *Cell stem cell*, 15(5):634–642.

Jullien, J., Astrand, C., Halley-Stott, R. P., Garrett, N., and Gurdon, J. B. (2010). Characterization of somatic cell nuclear reprogramming by oocytes in which a linker histone is required for pluripotency gene reactivation. *Proceedings of the National Academy of Sciences*, 107(12):5483–5488.

Kafer, G. R., Lehnert, S. A., Pantaleon, M., Kaye, P. L., and Moser, R. J. (2010). Expression of genes coding for histone variants and histone-associated proteins in pluripotent stem cells and mouse preimplantation embryos. *Gene expression patterns : GEP*, 10(6):299–305.

Kahan, B. W. and Ephrussi, B. (1970). Developmental potentialities of clonal in vitro cultures of mouse testicular teratoma. *Journal of the National Cancer Institute*, 44(5):1015–1036.

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(3):285–292.

Kaji, K., Norrby, K., Paca, A., Mileikovsky, M., Mohseni, P., and Woltjen, K. (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*, 458(7239):771–775.

Kalocsay, M., Hiller, N. J., and Jentsch, S. (2009). Chromosome-wide Rad51 spreading and SUMO-H2A.Z-dependent chromosome fixation in response to a persistent DNA double-strand break. *Molecular cell*, 33(3):335–343.

Kamine, J., Elangovan, B., Subramanian, T., Coleman, D., and Chinnadurai, G. (1996). Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator. *Virology*, 216(2):357–366.

Kanatsu-Shinohara, M., Inoue, K., Lee, J., Yoshimoto, M., Ogonuki, N., Miki, H., Baba, S., Kato, T., Kazuki, Y., Toyokuni, S., Toyoshima, M., Niwa, O., Oshimura, M., Heike, T., Nakahata, T., Ishino, F., Ogura, A., and Shinohara, T. (2004). Generation of pluripotent stem cells from neonatal mouse testis. *Cell*, 119(7):1001–1012.

Kannagi, R., Cochran, N. A., Ishigami, F., and Hakomori, S. (1983). Stage-specific embryonic antigens (SSEA-3 and-4) are epitopes of a unique globo-series ganglioside isolated from human teratocarcinoma cells.

Karagiannis, P. and Eto, K. (2016). Ten years of induced pluripotency: from basic mechanisms to therapeutic applications. *Development*, 143(12):2039–2043.

Kato, Y., Tani, T., Sotomaru, Y., Kurokawa, K., Kato, J.-y., Doguchi, H., Yasue, H., and Tsunoda, Y. (1998). Eight Calves Cloned from Somatic Cells of a Single Adult. *Science (New York, NY)*, 282(5396):2095–2098.

Keogh, M.-C., Mennella, T. A., Sawa, C., Berthelet, S., Krogan, N. J., Wolek, A., Podolny, V., Carpenter, L. R., Greenblatt, J. F., Baetz, K., and Buratowski, S. (2006). The Saccharomyces cerevisiae histone H2A variant Htz1 is acetylated by NuA4. *Genes & ...*.

Kidder, B. L., Hu, G., Yu, Z.-X., Liu, C., and Zhao, K. (2013). Extended self-renewal and accelerated reprogramming in the absence of Kdm5b. *Molecular and cellular biology*, 33(24):4793–4810.

Kim, D., Kim, C.-H., Moon, J.-I., Chung, Y.-G., Chang, M.-Y., Han, B.-S., Ko, S., Yang, E., Cha, K. Y., Lanza, R., and Kim, K.-S. (2009). Generation of Human Induced Pluripotent Stem Cells by Direct Delivery of Reprogramming Proteins. *Cell stem cell*, 4(6):472–476.

Kim, J. D., Faulk, C., and Kim, J. (2007). Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. *Nucleic acids research*, 35(10):3442–3452.

Kim, J. H., Kim, B., Cai, L., Choi, H. J., Ohgi, K. A., Tran, C., Chen, C., Chung, C. H., Huber, O., Rose, D. W., Sawyers, C. L., Rosenfeld, M. G., and Baek, S. H. (2005). Transcriptional regulation of a metastasis suppressor gene by Tip60 and beta-catenin complexes. *Nature*, 434(7035):921–926.

Kim, S.-I., Oceguera-Yanez, F., Hirohata, R., Linker, S., Okita, K., Yamada, Y., Yamamoto, T., Yamanaka, S., and Woltjen, K. (2015). KLF4 N-Terminal Variance Modulates Induced Reprogramming to Pluripotency. *Stem Cell Reports*, 4(4):727–743.

Kimura, A. and Horikoshi, M. (1998). Tip60 acetylates six lysines of a specific class in core histones in vitro. *Genes to cells : devoted to molecular & cellular mechanisms*, 3(12):789–800.

Kimura, A., Umehara, T., and Horikoshi, M. (2002). Chromosomal gradient of histone acetylation established by Sas2p and Sir2p functions as a shield against gene silencing. *Nature genetics*, 32(3):370–377.

Kleinsmith, L. J. and Pierce, G. B. (1964). Multipotentiality of Single Embryonal Carcinoma Cells. *Cancer research*, 24:1544–1551.

Ko, K., Tapia, N., Wu, G., Kim, J. B., Bravo, M. J. A., Sasse, P., Glaser, T., Ruau, D., Han, D. W., Greber, B., Hausdörfer, K., Sebastiano, V., Stehling, M., Fleischmann, B. K., Brüstle, O., Zenke, M., and Schöler, H. R. (2009). Induction of pluripotency in adult unipotent germline stem cells. *Cell stem cell*, 5(1):87–96.

Kobor, M. S., Venkatasubrahmanyam, S., and Meneghini, M. D. (2004). PLOS Biology: A Protein Complex Containing the Conserved Swi2/Snf2-Related ATPase Swr1p Deposits Histone Variant H2A.Z into Euchromatin. *PLoS biology*.

Kojima, Y., Kaufman-Francis, K., Studdert, J. B., Steiner, K. A., Power, M. D., Loebel, D. A. F., Jones, V., Hor, A., de Alencastro, G., Logan, G. J., Teber, E. T., Tam, O. H., Stutz, M. D., Alexander, I. E., Pickett, H. A., and Tam, P. P. L. (2014). The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell stem cell*, 14(1):107–120.

Korber, P. and Hörz, W. (2004). SWRred Not Shaken. *Cell*, 117(1):5–7.

Krijger, P. H. L., Di Stefano, B., de Wit, E., Limone, F., van Oevelen, C., de Laat, W., and Graf, T. (2016). Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell stem cell*, 18(5):597–610.

Krogan, N. J., Baetz, K., Keogh, M.-C., Datta, N., Sawa, C., Kwok, T. C. Y., Thompson, N. J., Davey, M. G., Pootoolal, J., Hughes, T. R., Emili, A., Buratowski, S., Hieter, P., and Greenblatt, J. F. (2004). Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. *Proceedings of the National Academy of Sciences of the United States of America*, 101(37):13513–13518.

Krogan, N. J., Keogh, M.-C., Datta, N., Sawa, C., Ryan, O. W., Ding, H., Haw, R. A., Pootoolal, J., Tong, A., Canadien, V., Richards, D. P., Wu, X., Emili, A., Hughes, T. R., Buratowski, S., and Greenblatt, J. F. (2003). A Snf2 Family ATPase Complex Required for Recruitment of the Histone H2A Variant Htz1. *Molecular cell*, 12(6):1565–1576.

Ku, M., Jaffe, J. D., Koche, R. P., Rheinbay, E., Endoh, M., Koseki, H., Carr, S. A., and Bernstein, B. E. (2012). H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome biology*, 13(10):R85.

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(6110):295–298.

Kusch, T., Florens, L., Macdonald, W. H., Swanson, S. K., Glaser, R. L., Yates, J. R., Abmayr, S. M., Washburn, M. P., and Workman, J. L. (2004). Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. *Science (New York, NY)*, 306(5704):2084–2087.

Kyttälä, A., Moraghebi, R., Valensisi, C., Kettunen, J., Andrus, C., Pasumarthy, K. K., Nakanishi, M., Nishimura, K., Ohtaka, M., Weltner, J., Van Handel, B., Parkkonen, O., Sinisalo, J., Jalanko, A., Hawkins, R. D., Woods, N.-B., Otonkoski, T., and Trokovic, R. (2016). Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. *Stem Cell Reports*, 6(2):200–212.

Labosky, P. A., Barlow, D. P., and Hogan, B. L. (1994). Mouse embryonic germ (EG) cell lines: transmission through the germline and differences in the methylation imprint of insulin-like growth factor 2 receptor (Igf2r) gene compared with embryonic stem (ES) cell lines. *Development*, 120(11):3197–3204.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Computational Biology*, 7(1):e1001057.

Langfelder, P., Zhang, B., and Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)*, 24(5):719–720.

Lee, K., Lau, Z. Z., Meredith, C., and Park, J. H. (2012). Decrease of p400 ATPase complex and loss of H2A.Z within the p21 promoter occur in senescent IMR-90 human fibroblasts. *Mechanisms of ageing and development*, 133(11-12):686–694.

Leitch, H. G., Blair, K., Mansfield, W., Ayetey, H., Humphreys, P., Nichols, J., Surani, M. A., and Smith, A. (2010). Embryonic germ cells from mice and rats exhibit properties consistent with a generic pluripotent ground state. *Development*, 137(14):2279–2287.

Leitch, H. G., Nichols, J., Humphreys, P., Mulas, C., Martello, G., Lee, C., Jones, K., Surani, M. A., and Smith, A. G. (2013). Rebuilding Pluripotency from Primordial Germ Cells. *Stem Cell Reports*, 1(1):66–78.

Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W., and Kaestner, K. H. (2012). Foxa2 and H2A.Z Mediate Nucleosome Depletion during Embryonic Stem Cell Differentiation. *Cell*, 151(7):1608–1616.

Lin, C.-J., Koh, F. M., Wong, P., Conti, M., and Ramalho-Santos, M. (2014). Hira-Mediated H3.3 Incorporation Is Required for DNA Replication and Ribosomal RNA Transcription in the Mouse Zygote. *Developmental cell*, 30(3):268–279.

Liu, X., Li, B., and GorovskyMA (1996). Essential and nonessential histone H2A variants in Tetrahymena thermophila. *Molecular and cellular biology*, 16(8):4305–4311.

Loizou, J. I., Oser, G., Shukla, V., Sawan, C., Murr, R., Wang, Z.-Q., Trumpp, A., and Herceg, Z. (2009). Histone acetyltransferase cofactor Trrap is essential for maintaining the hematopoietic stem/progenitor cell pool. *Journal of immunology (Baltimore, Md : 1950)*, 183(10):6422–6431.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550–48.

Lu, P. Y. T. L. Y. T., Lévesque, N., and Kobor, M. S. K. S. (2009). NuA4 and SWR1-C: two chromatin-modifying complexes with overlapping functions and componentsThis paper is one of a selection of papers published in this Special Issue, entitled 30th Annual International Asilomar Chromatin and Chromosomes Conference, and has undergone the Journal's usual peer review process. *Biochemistry and Cell Biology*, 87(5):799–815.

Lujan, E., Zunder, E. R., Ng, Y. H., Goronzy, I. N., Nolan, G. P., and Wernig, M. (2015). Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature*, 521(7552):352–356.

Ma, H., Morey, R., O'Neil, R. C., He, Y., Daughtry, B., Schultz, M. D., Hariharan, M., Nery, J. R., Castanon, R., Sabatini, K., Thiagarajan, R. D., Tachibana, M., Kang, E., Tippner-Hedges, R., Ahmed, R., Gutierrez, N. M., Van Dyken, C., Polat, A., Sugawara, A., Sparman, M., Gokhale, S., Amato, P., Wolf, D. P., Ecker, J. R., Laurent, L. C., and Mitalipov, S. (2014). Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature*, 511(7508):177–183.

Macfarlan, T. S., Gifford, W. D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S. E., Franco, L., Rosenfeld, M. G., Ren, B., and Pfaff, S. L. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes & development*, 25(6):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(7405):57–63.

Madigan, J. P., Chotkowski, H. L., and Glaser, R. L. (2002). DNA double-strand break-induced phosphorylation of Drosophila histone variant H2Av helps prevent radiation-induced apoptosis. *Nucleic acids research*, 30(17):3698–3705.

March-Díaz, R., García-Domínguez, M., Lozano-Juste, J., León, J., Florencio, F. J., and Reyes, J. C. (2007). Histone H2A.Z and homologues of components of the SWR1 complex are required to control immunity in Arabidopsis. *The Plant Journal*, 53(3):475–487.

Marques, M., Laflamme, L., Gervais, A. L., and Gaudreau, L. (2010). Reconciling the positive and negative roles of histone H2A.Z in gene transcription. *Epigenetics : official journal of the DNA Methylation Society*, 5(4):267–272.

Marques, M., Marques, M., Laflamme, L., Laflamme, L., Gaudreau, L., and Gaudreau, L. (2013). Estrogen receptor can selectively repress dioxin receptor-mediated gene expression by targeting DNA methylation. *Nucleic acids research*, 41(17):8094–8106.

Martin, G. R. (1980). Teratocarcinomas and mammalian embryogenesis. *Science (New York, NY)*.

Martin, G. R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638.

Martin, G. R. and Evans, M. J. (1975). Multiple differentiation of clonal teratocarcinoma stem cells following embryoid body formation in vitro. *Cell*, 6(4):467–474.

Martinato, F., Cesaroni, M., Amati, B., and Guccione, E. (2008). Analysis of Myc-induced histone modifications on target chromatin. *PloS one*, 3(11):e3650.

Matoba, S., Liu, Y., Lu, F., Iwabuchi, K. A., Shen, L., Inoue, A., and Zhang, Y. (2014). Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. *Cell*, 159(4):884–895.

Matsui, Y., Zsebo, K., and Hogan, B. L. (1992). Derivation of pluripotential embryonic stem cells from murine primordial germ cells in culture. *Cell*, 70(5):841–847.

Mehta, M., Kim, H. S., and Keogh, M. C. (2010). Sometimes one just isn't enough: do vertebrates contain an H2A. Z hyper-variant? *Journal of biology*.

Meneghini, M. D., Wu, M., and Madhani, H. D. (2003). Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin. *Cell*, 112(5):725–736.

Meregalli, M., Farini, A., and Torrente, Y. (2011). Stem Cell Therapy for Neuromuscular Diseases. InTech.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P. J., Brown, D. T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell*, 10(1):105–116.

Miao, J., Fan, Q., Cui, L., Li, J., Li, J., and Cui, L. (2006). The malaria parasite Plasmodium falciparum histones: organization, expression, and acetylation. *Gene*, 369:53–65.

Millar, C. B., Xu, F., Zhang, K., and Grunstein, M. (2006). Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. *Genes & development*, 20(6):711–722.

Mintz, B. and Illmensee, K. (1975). Normal genetically mosaic mice produced from malignant teratocarcinoma cells. *Proceedings of the National Academy of Sciences*, 72(9):3585–3589.

Mizuguchi, G., Shen, X., Landry, J., Wu, W. H., Sen, S., and Wu, C. (2004). ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science (New York, NY)*, 303(5656):343–348.

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(1):47–62.

Morillo-Huesca, M., Clemente-Ruiz, M., Andújar, E., and Prado, F. (2010). The SWR1 Histone Replacement Complex Causes Genetic Instability and Genome-Wide Transcription Misregulation in the Absence of H2A.Z. *PloS one*, 5(8):e12143.

Morrison, A. J. and Shen, X. (2009). Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. *Nature reviews Molecular cell biology*, 10(6):373–384.

Nagl, N. G., Wang, X., Patsialou, A., Van Scoy, M., and Moran, E. (2007). Distinct mammalian SWI/SNF chromatin remodeling complexes with opposing roles in cell-cycle control. *The EMBO journal*, 26(3):752–763.

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. *Proceedings of the National Academy of Sciences*, 90(18):8424–8428.

Nakatani, J., Tamada, K., Hatanaka, F., Ise, S., Ohta, H., Inoue, K., Tomonaga, S., Watanabe, Y., Chung, Y.-J., Banerjee, R., Iwamoto, K., Kato, T., Okazawa, M., Yamauchi, K., Tanda, K., Takao, K., Miyakawa, T., Bradley, A., and Takumi, T. (2009). Abnormal Behavior in a Chromosome- Engineered Mouse Model for Human 15q11-13 Duplication Seen in Autism. *Cell*, 137(7):1235–1246.

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(7):814–818.

Nichols, J. and Smith, A. (2009). Naive and Primed Pluripotent States. *Cell stem cell*, 4(6):487–492.

Nichols, J. and Smith, A. (2011). The origin and identity of embryonic stem cells. *Development*, 138(1):3–8.

Nichols, J. and Smith, A. (2012). Pluripotency in the Embryo and in Culture. *Cold Spring Harbor perspectives in biology*, 4(8):a008128–a008128.

Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R., and Rossant, J. (2005). Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation. *Cell*, 123(5):917–929.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, 448(7151):313–317.

Okita, K., Nakagawa, M., Hyenjong, H., Ichisaka, T., and Yamanaka, S. (2008). Generation of Mouse Induced Pluripotent Stem Cells Without Viral Vectors. *Science (New York, NY)*, 322(5903):949–953.

O'Malley, J., Skylaki, S., Iwabuchi, K. A., Chantzoura, E., Ruetz, T., Johnsson, A., Tomlinson, S. R., Linnarsson, S., and Kaji, K. (2013a). High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature*, 499(7456):88–91.

O'Malley, J., Skylaki, S., Iwabuchi, K. A., Chantzoura, E., Ruetz, T., Johnsson, A., Tomlinson, S. R., Linnarsson, S., and Kaji, K. (2013b). High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature*, 499(7456):88–91.

Papaioannou, V. E., Gardner, R. L., McBurney, M. W., Babinet, C., and Evans, M. J. (1978). Participation of cultured teratocarcinoma cells in mouse embryogenesis. *Journal of embryology and experimental morphology*, 44:93–104.

Papaioannou, V. E., McBurney, M. W., Gardner, R. L., and Evans, M. J. (1975). Fate of teratocarcinoma cells injected into early mouse embryos. *Nature*, 258(5530):70–73.

Papamichos-Chronakis, M., Watanabe, S., Rando, O. J., and Peterson, C. L. (2011). Global Regulation of H2A.Z Localization by the INO80 Chromatin-Remodeling Enzyme Is Essential for Genome Integrity. *Cell*, 144(2):200–213.

Pasque, V., Gillich, A., Garrett, N., and Gurdon, J. B. (2011). Histone variant macroH2A confers resistance to nuclear reprogramming. *The EMBO journal*, 30(12):2373–2387.

Paul, S., Pflieger, L., Dansithong, W., Figueroa, K. P., Gao, F., Coppola, G., and Pulst, S. M. (2016). Co-expression networks in generation of induced pluripotent stem cells. *Biology open*, 5(3):300–310.

Pera, M. F. (2014). In Search of Naivety. *Cell stem cell*, 15(5):543–545.

Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., Bar-Nur, O., Cheloufi, S., Stadtfeld, M., Figueroa, M. E., Robinton, D., Natesan, S., Melnick, A., Zhu, J., Ramaswamy, S., and Hochedlinger, K. (2012). A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell*, 151(7):1617–1632.

Polo, J. M., Liu, S., Figueroa, M. E., Kulalert, W., Eminli, S., Tan, K. Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., Natesan, S., Wagers, A. J., Melnick, A., Evans, T., and Hochedlinger, K. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nature biotechnology*, 28(8):848–855.

Prather, R. S., Sims, M. M., and First, N. L. (1990). Nuclear transplantation in the pig embryo: Nuclear swelling. *The Journal of experimental zoology*, 255(3):355–358.

Pyle, A. D., Lock, L. F., and Donovan, P. J. (2006). Neurotrophins mediate human embryonic stem cell survival. *Nature biotechnology*, 24(3):344–350.

Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A. A., Caspi, I., Krupalnik, V., Zerbib, M., Maza, I., Mor, N., Baran, D., Weinberger, L., Jaitin, D. A., Lara-Astiaso, D., Blecher-Gonen, R., Shipony, Z., Mukamel, Z., Hagai, T., Gilad, S., Amann-Zalcenstein, D., Tanay, A., Amit, I., Novershtern, N., and Hanna, J. H. (2013). Deterministic direct reprogramming of somatic cells to pluripotency. *Nature*, 502(7469):65–70.

Raisner, R. M., Raisner, R. M., Hartley, P. D., Hartley, P. D., Meneghini, M. D., Meneghini, M. D., Bao, M. Z., Bao, M. Z., Liu, C. L., Schreiber, S. L., Rando, O. J., and Madhani, H. D. (2005). Histone Variant H2A.Z Marks the 5´ Ends of Both Active and Inactive Genes in Euchromatin. *Cell*, 123(2):233–248.

Ralston, A. and Rossant, J. (2005). Genetic regulation of stem cell origins in the mouse embryo. *Clinical genetics*, 68(2):106–112.

Rangasamy, D. (2003). Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. *The EMBO journal*, 22(7):1599–1607.

Rangasamy, D. (2010). Histone variant H2A.Z can serve as a new target for breast cancer therapy. *Current medicinal chemistry*, 17(28):3155–3161.

Rangasamy, D., Greaves, I., and Tremethick, D. J. (2004). RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. *Nature structural & molecular biology*, 11(7):650–655.

Ren, Q. and Gorovsky, M. A. (2001). Histone H2A.Z acetylation modulates an essential charge patch. *Molecular cell*, 7(6):1329–1335.

Resnick, J. L., Bixler, L. S., Cheng, L., and Donovan, P. J. (1992). Long-term proliferation of mouse primordial germ cells in culture. *Nature*, 359(6395):550–551.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9309–9314.

Ridgway, P., Brown, K. D., Rangasamy, D., Svensson, U., and Tremethick, D. J. (2004). Unique Residues on the H2A.Z Containing Nucleosome Surface Are Important for Xenopus laevis Development . *Journal of Biological Chemistry*, 279(42):43815–43820.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature*, 323(6087):445–448.

Rossant, J. (2008). Stem Cells and Early Lineage Development. *Cell*, 132(4):527–531.

Rossant, J. (2015). Mouse and human blastocyst-derived stem cells: vive les differences. *Development*, 142(1):9–12.

Rossant, J. and McBurney, M. W. (1982). The developmental potential of a euploid male teratocarcinoma cell line after blastocyst injection. *Journal of embryology and experimental morphology*, 70(1):99–112.

Rouhani, F., Kumasaka, N., de Brito, M. C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. *PLoS genetics*, 10(6):e1004432.

Ruhl, Jingji Jin, Yong Cai, Selene Swanson, Laurence Florens, Michael P Washburn, Ronald C Conaway, Conaway, J. W., , and John C Chrivia (2006). Purification of a Human SRCAP Complex That Remodels Chromatin by Incorporating the Histone Variant H2A.Z into Nucleosomes†. *Biochemistry*, 45(17):5671–5677.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.-K., Beyer, T. A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J. L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell stem cell*, 7(1):64–77.

Santisteban, M. S., Kalashnikova, T., and Smith, M. M. (2000). Histone H2A.Z Regulates Transcription and Is Partially Redundant with Nucleosome Remodeling Complexes. *Cell*, 103(3):411–422.

Sarcinella, E., Zuzarte, P. C., Lau, P. N. I., Draker, R., and Cheung, P. (2007). Monoubiquitylation of H2A.Z Distinguishes Its Association with Euchromatin or Facultative Heterochromatin. *Molecular and cellular biology*, 27(18):6457–6468.

Sawan, C., Hernandez-Vargas, H., Murr, R., Lopez, F., Vaissière, T., Ghantous, A. Y., Cuenin, C., Imbert, J., Wang, Z.-Q., Ren, B., and Herceg, Z. (2013). Histone Acetyltransferase Cofactor Trrap Maintains Self-Renewal and Restricts Differentiation of Embryonic Stem Cells. *Stem cells (Dayton, Ohio)*, 31(5):979–991.

Schultz, R. M. (1993). Regulation of zygotic gene activation in the mouse. *BioEssays*, 15(8):531–538.

Schwarz, B. A., Bar-Nur, O., Silva, J. C. R., and Hochedlinger, K. (2014). Nanog is dispensable for the generation of induced pluripotent stem cells. *Current biology : CB*, 24(3):347–350.

Sevilla, A. and Binda, O. (2014). Post-translational modifications of the histone variant h2az. *Stem cell research*, 12(1):289–295.

Shahbazian, M. D. and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. *Annual Review of Biochemistry*, 76:75–100.

Shamblott, M. J., Axelman, J., Wang, S., Bugg, E. M., Littlefield, J. W., Donovan, P. J., Blumenthal, P. D., Huggins, G. R., and Gearhart, J. D. (1998). Derivation of pluripotent stem cells from cultured human primordial germ cells. *Proceedings of the National Academy of Sciences*, 95(23):13726–13731.

Sharma, U., Stefanova, D., and Holmes, S. G. (2013). Histone Variant H2A.Z Functions in Sister Chromatid Cohesion in Saccharomyces cerevisiae. *Molecular and cellular biology*, 33(17):3473–3481.

Shevinsky, L. H., Knowles, B. B., Damjanov, I., and Solter, D. (1982). Monoclonal antibody to murine embryos defines a stage-specific embryonic antigen expressed on mouse embryos and human teratocarcinoma cells. *Cell*, 30(3):697–705.

Shinagawa, T., Takagi, T., Tsukamoto, D., Tomaru, C., Huynh, L. M., Sivaraman, P., Kumarevel, T., Inoue, K., Nakato, R., Katou, Y., Sado, T., Takahashi, S., Ogura, A., Shirahige, K., and Ishii, S. (2014). Histone Variants Enriched in Oocytes Enhance Reprogramming to Induced Pluripotent Stem Cells. *Cell stem cell*, 14(2):217–227.

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(10):e253–11.

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(4):722–737.

Simeone, A., Acampora, D., Arcioni, L., Andrews, P. W., Boncinelli, E., and Mavilio, F. (1990). Sequential activation of HOX2 homeobox genes by retinoic acid in human embryonal carcinoma cells. *Nature*, 346(6286):763–766.

Skarnes, W. C., Rosen, B., West, A. P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A. O., Thomas, M., Harrow, J., Cox, T., Jackson, D., Severin, J., Biggs, P., Fu, J., Nefedov, M., de Jong, P. J., Stewart, A. F., and Bradley, A. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–342.

Smith, A. G. (2001). Embryo-Derived Stem Cells: Of Mice and Men. *Annual review of cell and developmental biology*, 17(1):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(6200):688–690.

Smithies, O. (2008). *Turning pages (Nobel lecture).*, volume 9.

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G. W., Cook, E. G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., Isacson, O., and Jaenisch, R. (2009). Parkinson's Disease Patient-Derived Induced Pluripotent Stem Cells Free of Viral Reprogramming Factors. *Cell*, 136(5):964–977.

Solter, D. (2006). From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nature reviews Genetics*, 7(4):319–327.

Solter, D., Skreb, N., and Damjanov, I. (1970). Extrauterine growth of mouse egg-cylinders results in malignant teratoma. *Nature*, 227(5257):503–504.

Sommer, C. A., Stadtfeld, M., Murphy, G. J., Hochedlinger, K., Kotton, D. N., and Mostoslavsky, G. (2009). Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem cells (Dayton, Ohio)*, 27(3):543–549.

Squatrito, M., Gorrini, C., and Amati, B. (2006). Tip60 in DNA damage response and growth control: many tricks in one HAT. *Trends in cell biology*, 16(9):433–442.

Stadtfeld, M., Nagaya, M., Utikal, J., Weir, G., and Hochedlinger, K. (2008). Induced pluripotent stem cells generated without viral integration. *Science (New York, NY)*, 322(5903):945–949.

Stevens, L. C. (1964). Experimental Production of Testicular teratomas in Mice. *Proceedings of the National Academy of Sciences*, 52:654–661.

Stevens, L. C. (1967). Origin of Testicular Teratomas From Primordial Germ Cells in Mice. *Journal of the National Cancer Institute*.

Stevens, L. C. (1968). The development of teratomas from intratesticular grafts of tubal mouse eggs. *Journal of embryology and experimental morphology*, 20(3):329–341.

Stevens, L. C. (1970). The development of transplantable teratocarcinomas from intratesticular grafts of pre- and postimplantation mouse embryos. *Developmental biology*, 21(3):364–382.

Stevens, L. C. (1973). A new inbred subline of mice (129-terSv) with a high incidence of spontaneous congenital testicular teratomas. *Journal of the National Cancer Institute*, 50(1):235–242.

Stevens, L. C. and Little, C. C. (1954). Spontaneous Testicular Teratomas in an Inbred Strain of Mice. *Proceedings of the National Academy of Sciences*, 40(11):1080–1087.

Stevens, L. C. and Varnum, D. S. (1974). The development of teratomas from parthenogenetically activated ovarian mouse eggs. *Developmental biology*, 37(2):369–380.

Stewart, T. A. and Mintz, B. (1981). Successive generations of mice produced from an established culture line of euploid teratocarcinoma cells. *Proceedings of the National Academy of Sciences*, 78(10):6314–6318.

Stewart, T. A. and Mintz, B. (1982). Recurrent germ-line transmission of the teratocarcinoma genome from the METT-1 culture line to progeny in vivo. *The Journal of experimental zoology*, 224(3):465–469.

Strickland, S. and Mahdavi, V. (1978). The induction of differentiation in teratocarcinoma stem cells by retinoic acid. *Cell*, 15(2):393–403.

Strumpf, D., Mao, C.-A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F., and Rossant, J. (2005). Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*, 132(9):2093–2102.

Stuart, H. T., van Oosten, A. L., Radzisheuskaya, A., Martello, G., Miller, A., Dietmann, S., Nichols, J., and Silva, J. C. R. (2014). NANOG Amplifies STAT3 Activation and They Synergistically Induce the Naive Pluripotent Program. *Current biology : CB*, 24(3):340–346.

Subramanian, V., Mazumder, A., Surface, L. E., Butty, V. L., Fields, P. A., Alwan, A., Torrey, L., Thai, K. K., Levine, S. S., Bathe, M., and Boyer, L. A. (2013a). H2A.Z Acidic Patch Couples Chromatin Dynamics to Regulation of Gene Expression Programs during ESC Differentiation. *PLoS genetics*, 9(8):e1003725.

Subramanian, V., Mazumder, A., Surface, L. E., Butty, V. L., Fields, P. A., Alwan, A., Torrey, L., Thai, K. K., Levine, S. S., Bathe, M., and Boyer, L. A. (2013b). H2A.Z Acidic Patch Couples Chromatin Dynamics to Regulation of Gene Expression Programs during ESC Differentiation. *PLoS genetics*, 9(8):e1003725–21.

Suka, N., Luo, K., and Grunstein, M. (2002). Sir2p and Sas2p opposingly regulate acetylation of yeast histone H4 lysine16 and spreading of heterochromatin. *Nature genetics*, 32(3):378–383.

Svotelis, A., Gévry, N., and Gaudreau, L. (2009). Regulation of gene expression and cellular proliferation by histone H2A.Z . *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 87(1):179–188.

Svotelis, A., Gévry, N., Grondin, G., and Gaudreau, L. (2010). H2A.Z overexpression promotes cellular proliferation of breast cancer cells. *Cell cycle (Georgetown, Tex)*, 9(2):364–370.

Taapken, S. M., Nisler, B. S., Newton, M. A., Sampsell-Barron, T. L., Leonhard, K. A., McIntire, E. M., and Montgomery, K. D. (2011). Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nature biotechnology*, 29(4):313–314.

Tachibana, M., Amato, P., Sparman, M., Gutierrez, N. M., Tippner-Hedges, R., Ma, H., Kang, E., Fulati, A., Lee, H.-S., Sritanaudomchai, H., Masterson, K., Larson, J., Eaton, D., Sadler-Fredd, K., Battaglia, D., Lee, D., Wu, D., Jensen, J., Patton, P., Gokhale, S., Stouffer, R. L., Wolf, D., and Mitalipov, S. (2013). Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell*, 153(6):1228–1238.

Tada, T., Tada, M., Hilton, K., Barton, S. C., Sado, T., Takagi, N., and Surani, M. A. (1998). Epigenotype switching of imprintable loci in embryonic germ cells. *Development Genes and Evolution*, 207(8):551–561.

Tahmasebi, S., Jafarnejad, S. M., Tam, I. S., Gonatopoulos-Pournatzis, T., Matta-Camacho, E., Tsukumo, Y., Yanagiya, A., Li, W., Atlasi, Y., Caron, M., Braunschweig, U., Pearl, D., Khoutorsky, A., Gkogkas, C. G., Nadon, R., Bourque, G., Yang, X.-J., Tian, B., Stunnenberg, H. G., Yamanaka, Y., Blencowe, B. J., Giguère, V., and Sonenberg, N. (2016). Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proceedings of the National Academy of Sciences*, 113(44):12360–12367.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Sasaki, A., Yamamoto, M., Nakamura, M., Sutou, K., Osafune, K., and Yamanaka, S. (2014). Induction of pluripotency in human somatic cells via a transient state resembling primitive streak-like mesendoderm. *Nature communications*, 5:ncomms4678.

Takahashi, K. and Yamanaka, S. (2006a). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.

Takahashi, K. and Yamanaka, S. (2006b). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., Reik, W., Bertone, P., and Smith, A. G. (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell*, 158(6):1254–1269.

Talbert, P. B. and Henikoff, S. (2010). Histone variants — ancient wrap artists of the epigenome. *Nature reviews Molecular cell biology*, 11(4):264–275.

Taubert, S., Gorrini, C., Frank, S. R., Parisi, T., Fuchs, M., Chan, H.-M., Livingston, D. M., and Amati, B. (2004). E2F-dependent histone acetylation and recruitment of the Tip60 acetyltransferase complex to chromatin in late G1. *Molecular and cellular biology*, 24(10):4546–4556.

Teranishi, T., Tanaka, M., Kimoto, S., Ono, Y., Miyakoshi, K., Kono, T., and Yoshimura, Y. (2004). Rapid replacement of somatic linker histones with the oocyte-specific linker histone H1foo in nuclear transfer. *Developmental biology*, 266(1):76–86.

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. G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, 448(7150):196–199.

Thatcher, T. H. and Gorovsky, M. A. (1994). Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic acids research*, 22(2):174–179.

Theunissen, T. W., Costa, Y., Radzisheuskaya, A., van Oosten, A. L., Lavial, F., Pain, B., Castro, L. F. C., and Silva, J. C. R. (2011). Reprogramming capacity of Nanog is functionally conserved in vertebrates and resides in a unique homeodomain. *Development*, 138(22):4853–4865.

Theunissen, T. W., Powell, B. E., Wang, H., Mitalipova, M., Faddah, D. A., Reddy, J., Fan, Z. P., Maetzel, D., Ganz, K., Shi, L., Lungjangwa, T., Imsoonthornruksa, S., Stelzer, Y., Rangarajan, S., D'Alessio, A., Zhang, J., Gao, Q., Dawlaty, M. M., Young, R. A., Gray, N. S., and Jaenisch, R. (2014a). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. *Cell stem cell*, 15(4):471–487.

Theunissen, T. W., Powell, B. E., Wang, H., Mitalipova, M., Faddah, D. A., Reddy, J., Fan, Z. P., Maetzel, D., Ganz, K., Shi, L., Lungjangwa, T., Imsoonthornruksa, S., Stelzer, Y., Rangarajan, S., D'Alessio, A., Zhang, J., Gao, Q., Dawlaty, M. M., Young, R. A., Gray, N. S., and Jaenisch, R. (2014b). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. *Stem Cell*, pages 1–46.

Thompson, S., Clarke, A. R., Pow, A. M., Hooper, M. L., and Melton, D. W. (1989). Germ line transmission and expression of a corrected HPRT gene produced by gene targeting in embryonic stem cells. *Cell*, 56(2):313–321.

Thomson, J. A. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science (New York, NY)*, 282(5391):1145–1147.

Tiemann, U., Marthaler, A. G., Adachi, K., Wu, G., Fischedick, G. U. L., Araúzo-Bravo, M. J., Schöler, H. R., and Tapia, N. (2014). Counteracting Activities of OCT4 and KLF4 during Reprogramming to Pluripotency. *Stem Cell Reports*, 2(3):351–365.

Trelle, M. B., Salcedo-Amaya, A. M., Cohen, A. M., Stunnenberg, H. G., and Jensen, O. N. (2009). Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite Plasmodium falciparum. *Journal of proteome research*, 8(7):3439–3450.

Ueda, A., Akagi, T., and Yokota, T. (2017). GA-Binding Protein Alpha Is Involved in the Survival of Mouse Embryonic Stem Cells. *Stem cells (Dayton, Ohio)*, 306:391.

Umemoto, T., Yamato, M., Ishihara, J., Shiratsuchi, Y., Utsumi, M., Morita, Y., Tsukui, H., Terasawa, M., Shibata, T., Nishida, K., Kobayashi, Y., Petrich, B. G., Nakauchi, H., Eto, K., and Okano, T. (2012). Integrin-αvβ3 regulates thrombopoietin-mediated maintenance of hematopoietic stem cells. *Blood*, 119(1):83–94.

Valamehr, B., Robinson, M., Abujarour, R., Rezner, B., Vranceanu, F., Le, T., Medcalf, A., Lee, T. T., Fitch, M., Robbins, D., and Flynn, P. (2014). Platform for Induction and Maintenance of Transgene-free hiPSCs Resembling Ground State Pluripotent Stem Cells. *Stem Cell Reports*, 2(3):366–381.

Valdes-Mora, F., Song, J. Z., Statham, A. L., Strbenac, D., Robinson, M. D., Nair, S. S., Patterson, K. I., Tremethick, D. J., Stirzaker, C., and Clark, S. J. (2012). Acetylation of H2A.Z is a key epigenetic modification associated with gene deregulation and epigenetic remodeling in cancer. *Genome research*, 22(2):307–321.

van Daal, A. and Elgin, S. C. (1992). A histone variant, H2AvD, is essential in Drosophila melanogaster. *Molecular biology of the cell*, 3(6):593–602.

van den Berg, D. L. C., 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(4):369–381.

Vidal, S. E., Amlani, B., Chen, T., Tsirigos, A., and Stadtfeld, M. (2014). Combinatorial Modulation of Signaling Pathways Reveals Cell-Type-Specific Requirements for Highly Efficient and Synchronous iPSC Reprogramming. *Stem Cell Reports*, 3(4):574–584.

Waddington, C. H. (1957). *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.* London: George Allen & Unwin, Ltd.

Wakayama, S., Kohda, T., Obokata, H., Tokoro, M., Li, C., Terashita, Y., Mizutani, E., Nguyen, V. T., Kishigami, S., Ishino, F., and Wakayama, T. (2013). Successful Serial Recloning in the Mouse over Multiple Generations. *Cell stem cell*, 12(3):293–297.

Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R., and Yanagimachi, R. (1998). Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei . *Nature*, 394(6691):369–374.

Wang, F., Han, J., Wang, L., Jing, Y., Zhu, Z., Hui, D., Wang, Z., Wang, Y., Dong, Y., and Tan, T. (2017). CCCTC-Binding Factor Transcriptionally Targets Wdr5 to Mediate Somatic Cell Reprogramming. *Stem cells and development*, 26(10):743–750.

Wang, J., Qiao, M., He, Q., Shi, R., Loh, S. J. H., Stanton, L. W., and Wu, M. (2015). Pluripotency Activity of Nanog Requires Biochemical Stabilization by Variant Histone Protein H2A.Z. *Stem cells (Dayton, Ohio)*, 33(7):2126–2134.

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G. G., Chen, W., Lorincz, M. C., Ivics, Z., Hurst, L. D., and Izsvák, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531):405–409.

Wang, W., Wang, W., Yang, J., Yang, J., Liu, H., Liu, H., Lu, D., Lu, D., Chen, X., Chen, X., Zenonos, Z., Zenonos, Z., Campos, L. S., Campos, L. S., Rad, R., Rad, R., Guo, G., Guo, G., Zhang, S., Zhang, S., Bradley, A., Bradley, A., Liu, P., and Liu, P. (2011). Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1. *Proceedings of the National Academy of Sciences*, 108(45):18283–18288.

Ware, C. B., Nelson, A. M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E. C., Jimenez-Caliani, A. J., Deng, X., Cavanaugh, C., Cook, S., Tesar, P. J., Okada, J., Margaretha, L., Sperber, H., Choi, M., Blau, C. A., Treuting, P. M., Hawkins, R. D., Cirulli, V., and Ruohola-Baker, H. (2014). Derivation of naive human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(12):4484–4489.

Warren, L., Manos, P. D., Ahfeldt, T., Loh, Y.-H., Li, H., Lau, F., Ebina, W., Mandal, P. K., Smith, Z. D., Meissner, A., Daley, G. Q., Brack, A. S., Collins, J. J., Cowan, C., Schlaeger, T. M., and Rossi, D. J. (2010). Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. *Cell stem cell*, 7(5):618–630.

Weismann, A., Parker, W. N., and Ronnfeldt, H. (1894). The Germ-Plasm: A Theory of Heredity. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 23:433.

Wenkert, D. and Allis, C. D. (1984). Timing of the appearance of macronuclear-specific histone variant hv1 and gene expression in developing new macronuclei of Tetrahymena thermophila. *The Journal of cell biology*, 98(6):2107–2117.

West, M. H. P. and Bonner, W. M. (1980). Histone 2A, a heteromorphous family of eight protein species. *Biochemistry*, 19(14):3238–3245.

Whittle, C. M., McClinic, K. N., Ercan, S., Zhang, X., Green, R. D., Kelly, W. G., and Lieb, J. D. (2008). The genomic distribution and function of histone variant HTZ-1 during C. elegans embryogenesis. *PLoS genetics*, 4(9):e1000187.

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(6200):684–687.

Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J., and Campbell, K. H. (1997). Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385(6619):810–813.

Woltjen, K., Michael, I. P., Mohseni, P., Desai, R., Mileikovsky, M., Hämäläinen, R., Cowling, R., Wang, W., Liu, P., Gertsenstein, M., Kaji, K., Sung, H.-K., and Nagy, A. (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*, 458(7239):766–770.

Wong, M. M., Wong, M. M., Cox, L. K., Cox, L. K., Chrivia, J. C., and Chrivia, J. C. (2007). The Chromatin Remodeling Protein, SRCAP, Is Critical for Deposition of the Histone Variant H2A.Z at Promoters. *Journal of Biological Chemistry*, 282(36):26132–26139.

Wu, G., Lei, L., and Schöler, H. R. (2017). Totipotency in the mouse. *Journal of Molecular Medicine*, 95(7):687–694.

Wu, J., Okamura, D., Li, M., Suzuki, K., Luo, C., Ma, L., and He, Y. (2015). An alternative pluripotent state confers interspecies chimaeric competency. *Nature*.

Wu, R. S., Tsai, S., and Bonner, W. M. (1982). Patterns of histone variant synthesis can distinguish G0 from G1 cells. *Cell*, 31(2 Pt 1):367–374.

Wu, S., Shi, Y., Mulligan, P., Gay, F., Landry, J., Liu, H., Lu, J., Qi, H. H., Wang, W., Nickoloff, J. A., Wu, C., and Shi, Y. (2007). A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair. *Nature structural & molecular biology*, 14(12):1165–1172.

Wu, W.-H., Alami, S., Luk, E., Wu, C.-H., Sen, S., Mizuguchi, G., Wei, D., and Wu, C. (2005). Swc2 is a widely conserved H2AZ-binding module essential for ATP-dependent histone exchange. *Nature structural & molecular biology*, 12(12):1064–1071.

Wu, W.-H., Wu, C.-H., Ladurner, A., Mizuguchi, G., Wei, D., Xiao, H., Luk, E., Ranjan, A., and Wu, C. (2009). N terminus of Swr1 binds to histone H2AZ and provides a platform for subunit assembly in the chromatin remodeling complex. *Journal of Biological Chemistry*, 284(10):6200–6207.

Xiao, H., Chung, J., Kao, H.-Y., and Yang, Y.-C. (2003). Tip60 is a co-repressor for STAT3. *Journal of Biological Chemistry*, 278(13):11197–11204.

Xu, Y., Ayrapetov, M. K., Xu, C., Gursoy-Yuzugullu, O., Hu, Y., and Price, B. D. (2012). Histone H2A.Z Controls a Critical Chromatin Remodeling Step Required for DNA Double-Strand Break Repair. *Molecular cell*, 48(5):723–733.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J.-y., Horvath, S., and Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNAsequencing. *Nature*, 500(7464):593–597.

Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. *Nature*, 460(7251):49–52.

Yamanaka, S. (2012). Induced Pluripotent Stem Cells: Past, Present, and Future. *Stem Cell*, 10(6):678–684.

Yi, H., Sardesai, N., Fujinuma, T., Chan, C.-W., Veena, and Gelvin, S. B. (2006). Constitutive expression exposes functional redundancy between the Arabidopsis histone H2A gene HTA1 and other H2A gene family members. *The Plant cell*, 18(7):1575–1589.

Ying, Q.-L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP Induction of Id Proteins Suppresses Differentiation and Sustains Embryonic Stem Cell Self-Renewal in Collaboration with STAT3. *Cell*, 115(3):281–292.

Ying, Q.-L. and Smith, A. G. (2003). Defined conditions for neural commitment and differentiation. *Methods in enzymology*, 365:327–341.

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(7194):519–523.

Young, M. A., Larson, D. E., Sun, C.-W., George, D. R., Ding, L., Miller, C. A., Lin, L., Pawlik, K. M., Chen, K., Fan, X., Schmidt, H., Kalicki-Veizer, J., Cook, L. L., Swift, G. W., Demeter, R. T., Wendl, M. C., Sands, M. S., Mardis, E. R., Wilson, R. K., Townes, T. M., and Ley, T. J. (2012). Background Mutations in Parental Cells Account for Most of the Genetic Heterogeneity of Induced Pluripotent Stem Cells. *Cell stem cell*, 10(5):570–582.

Youngren, K. K., Coveney, D., Peng, X., Bhattacharya, C., Schmidt, L. S., Nickerson, M. L., Lamb, B. T., Deng, J. M., Behringer, R. R., Capel, B., Rubin, E. M., Nadeau, J. H., and Matin, A. (2005). The Ter mutation in the dead end gene causes germ cell loss and testicular germ cell tumours. *Nature*, 435(7040):360–364.

Yu, C., Zhang, Y.-L., Pan, W.-W., Li, X.-M., Wang, Z.-W., Ge, Z.-J., Zhou, J.-J., Cang, Y., Tong, C., Sun, Q.-Y., and Fan, H.-Y. (2013). CRL4 complex regulates mammalian oocyte survival and reprogramming by activation of TET proteins. *Science (New York, NY)*, 342(6165):1518–1521.

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. *Science (New York, NY)*, 324(5928):797–801.

Yu, S., Ma, H., Ow, J. R., Goh, Z., Chiang, R., Yang, H., Loh, Y.-H., and Wu, Q. (2016). Zfp553 Is Essential for Maintenance and Acquisition of Pluripotency. *Stem cells and development*, 25(1):55–67.

Yusa, K., Rad, R., Takeda, J., and Bradley, A. (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nature methods*, 6(5):363–369.

Yusa, K., Zhou, L., Li, M. A., Bradley, A., and Craig, N. L. (2011). A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences*, 108(4):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., Wersto, R. P., and Ko, M. S. H. (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature*, 464(7290):858–863.

Zhang, B. and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article17.

Zhang, H., Richardson, D. O., Roberts, D. N., Utley, R., Erdjument-Bromage, H., Tempst, P., Côté, J., and Cairns, B. R. (2004). The Yaf9 component of the SWR1 and NuA4 complexes is required for proper gene expression, histone H4 acetylation, and Htz1 replacement near telomeres. *Molecular and cellular biology*, 24(21):9424–9436.

Zhang, Y. H., Mei, S. Q., Peng, X. W., Zuo, B., Lei, M. G., Xiong, Y. Z., and Li, F. E. (2009). Molecular cloning and polymorphism of the porcine H2AFZ gene. *Animal : an international journal of animal bioscience*, 3(6):779–782.

Zhao, Y., Yang, Z., Wu, J., Wu, R., Keshipeddy, S. K., Wright, D., and Wang, L. (2017). High-mobility-group protein 2 regulated by microRNA-127 and small heterodimer partner modulates pluripotency of mouse embryonic stem cells and liver tumor initiating cells. *Hepatology Communications*, 15:496.

Zhou, H., Wu, S., Joo, J. Y., Zhu, S., Han, D. W., Lin, T., Trauger, S., Bien, G., Yao, S., Zhu, Y., Siuzdak, G., Schöler, H. R., Duan, L., and Ding, S. (2009). Generation of induced pluripotent stem cells using recombinant proteins. *Cell stem cell*, 4(5):381–384.

Zlatanova, J. and Thakar, A. (2008). H2A.Z: View from the Top. *Structure*, 16(2):166–179.

Zofall, M., Fischer, T., Zhang, K., Zhou, M., Cui, B., Veenstra, T. D., and Grewal, S. I. S. (2009). Histone H2A.Z cooperates with RNAi and heterochromatin factors to suppress antisense RNAs. *Nature*, 461(7262):419–422.

Zucchi, I., Mento, E., Kuznetsov, V. A., Scotti, M., Valsecchi, V., Simionati, B., Vicinanza, E., Valle, G., Pilotti, S., Reinbold, R., Vezzoni, P., Albertini, A., and Dulbecco, R. (2004). Gene expression profiles of epithelial cells microscopically isolated from a breast-invasive ductal carcinoma and a nodal metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52):18147–18152.

Supplementary Analysis

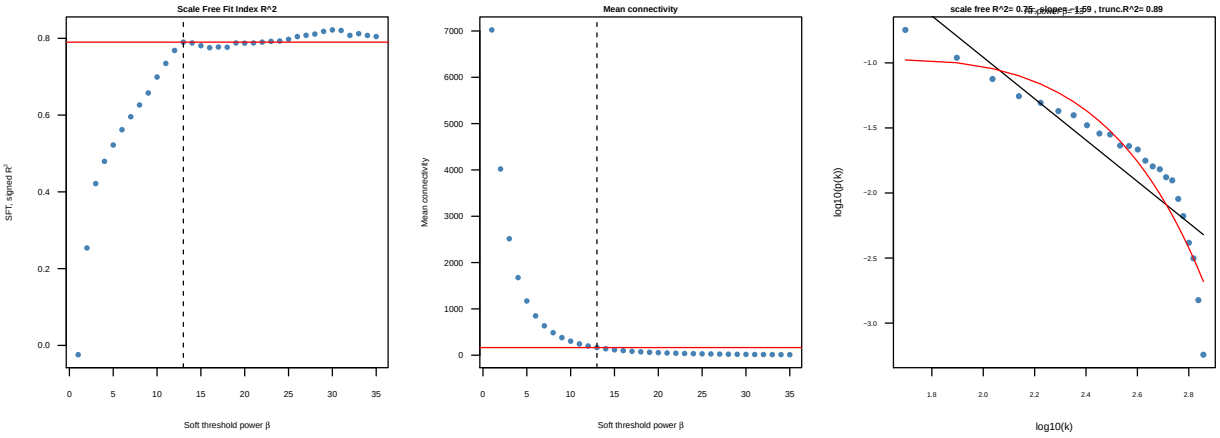
Effect of Sample Replicates in Network Structure

Fabio Amaral

4 Jan 2018

A robust Weighted Gene Correlation Network Analysis (WGCNA) is dependent on the number of samples used. According to the WGCNA package FAQ (https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html), correlations on fewer than 20 samples will incur in too much noise which consequently impacts its biological meaningfulness. Therefore, in order to assess the effect of sample replicates I selected only one biological replicate (*rep1) from each type from the Chantzoura et al 2013 data set resulting in exactly 20 samples. Information about this data set can be found in Chapter 4, Table 4.1.

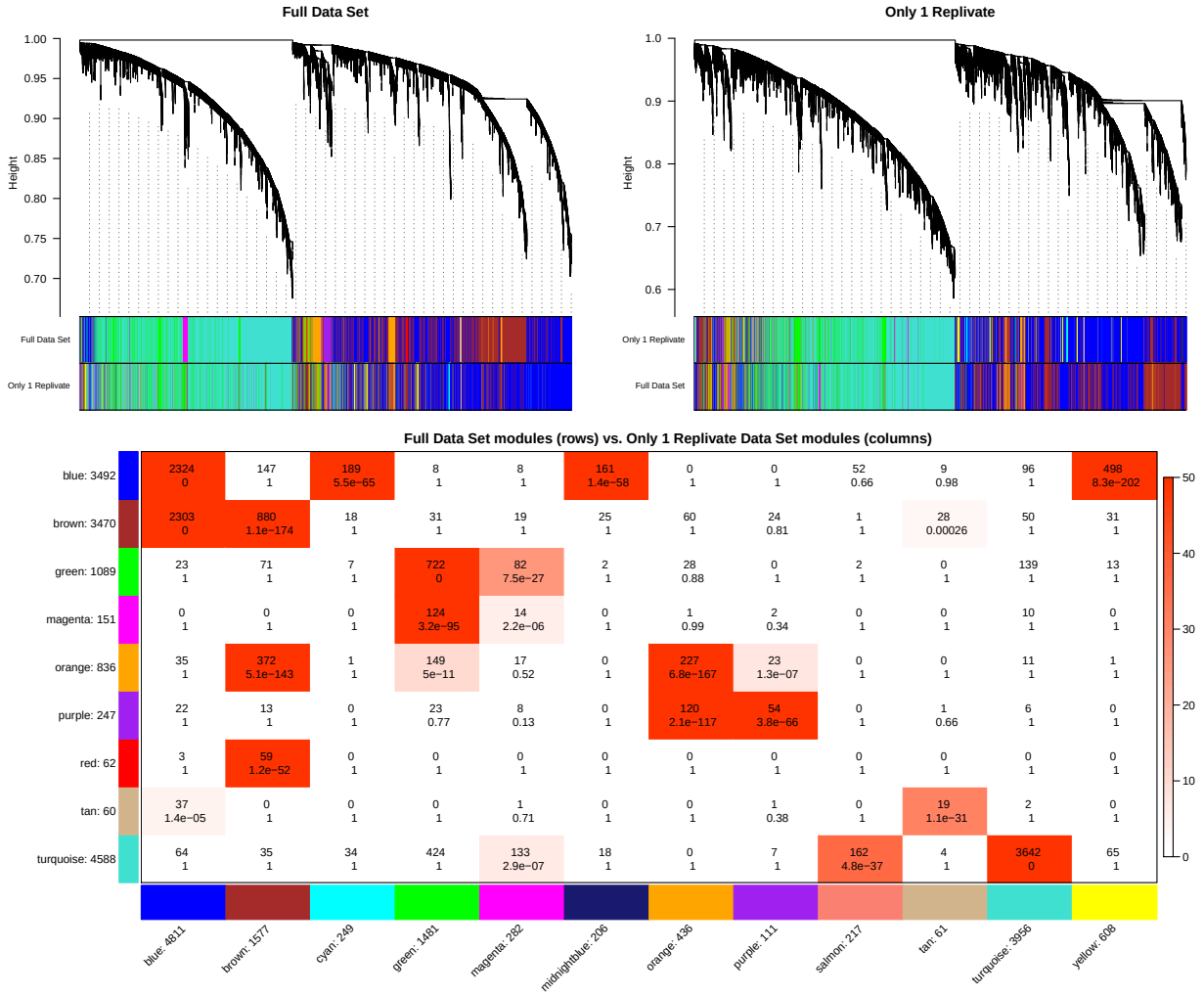
The downstream processing for the data subset with one replicate was done exactly as described in Chapter 4 except for the selection of the soft thresholding power β. While the chosen power β for the full data set was 12 the subset data set required a power β of 13 to more closely meet a scale free topology criteria as illustrated below.



Suplementary Figure 1: **Selection of β for the Chantzoura data subset with one biologival replicate for each cell type** :A β is selected at the point where the scale free fit index R^2 is close to saturation and the mean connectivity is close to the minimum. The β of 13 was chosen as the soft threshold value which causes the network topology to approximate a scalefree topology the most.

##

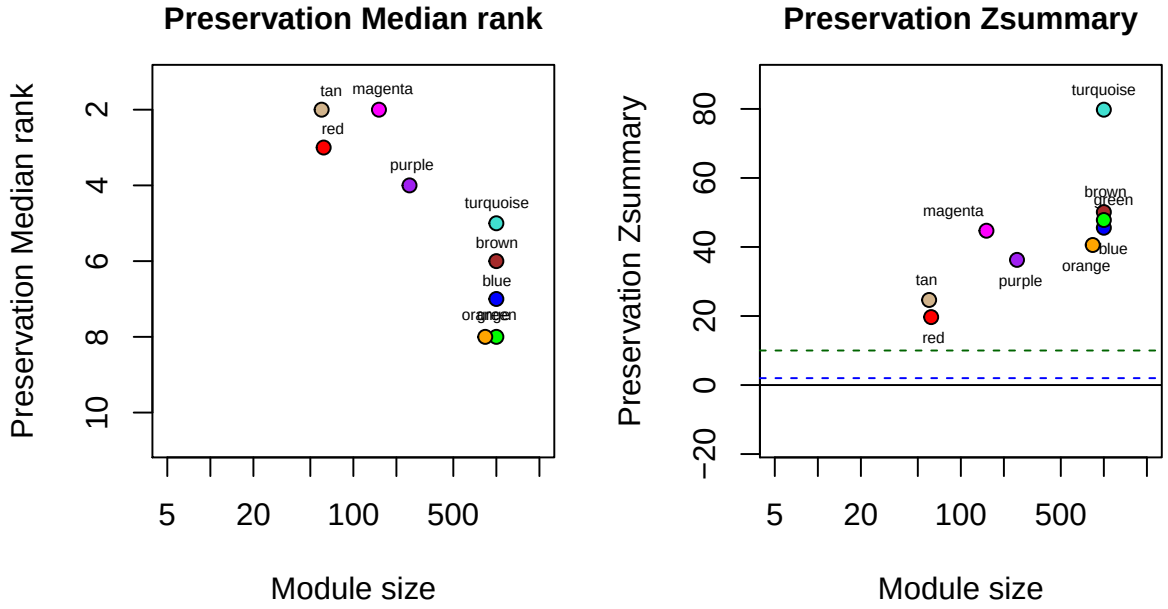
In order to compare the network module definitions when using the full data set or the data subset with one replicate per cell type each network dendrogramme was treated independently and resulting gene memberships were contrasted between the two analyses. The top two figures from the panel of Suplementary Figure 2 illustrate such dedrogrammes with the resulting color coded modules memberships imediately below. The second color coded modules memberships serve for a rough comparison between the two networks in question. As we can see there is good overlap between the colors from top and bottom colored network membreships regardless if the full data set or the subset was used for the module definition. The cross-tabulation in Suplementary Figure 2 further corroborate this notion of a well preserved network structure although the network created with the subset data resulted in more modules possibly as a consequense of noise resulted from using fewer samples.



Suplementary Figure 2: **Cross-tabulation based comparison of modules (defined as clusters) in Full Data Set and Data Subset (1 replicate per cell type)**: (Top) Hierarchical clustering trees (dendrogram) of genes based on full data set and data subset co-expression network. Each "leaf"" (short vertical line) corresponds to one gene. The module membership is indicated by the color rows below the dendrogram defined by cutting brantches of dendrograms (Left: Full data set and Right: Data Subset). (Bottom) Cross-tabulation of Full Data set modules (rows) and Data Subset modules (columns). Each row and column is labeled by the corresponding module color and the total number of genes in the module. The values inside the table correspond to the count of genes in the intersection of corresponding row and column modules. The red pallette color coding of the table represent the Fisher exact test p value, according to the color legend on the right.

Finally, to more systematically evaluate the preservation of the defined modules between using full and subset data I resorted to the calculation of the Preservation Median Rank and Preservation Zsummary as previously described in chapter 5 and using 100 permutations in order to calculate levels of significance. The Suplementary Figure 3 below illustrate the results of such analyses. The Preservation Median Rank shows that the small tan, magenta and red modules defined with the full data set were the most preserved modules. However the Preservation Zsummary results show these modules as some of the least preserved modules. This is aparently contradiction is not unexpected as the Zsummary calculation is heavily influenced the by the size (number of genes) of the modules analysized and interpretaion of these two complementary measurements of module preservation should be performed in conjunction (Langfelder et al. 2011). Nevertheless, all modules are extremely preserved (Zsummary >= 20 indicates high preservation) and these results demonstrate that

the number of replicates has not significantly influenced in the network structures whether using full or subset data sets.



Suplementary Figure 3: ** Chantzoura data sets (Full vs Subset) module preservation statistics.** All modules are extremely well preserved. The preservation median rank is illustrated on the left and the preservation Zsummary on the right. All modules showed a very high preservation Zsummary scores indicating that the subset data set was still able to capture very similar network structure when compared to using the full data set.

Conclusion

Given that there is a significant overlap between gene module memberships and high preservation network statistics between analyses with full and subeset (1 replicate per cell type) I conclude that the number of replicates was not responsible for driving the network module definition in the main analyses performed throughout this thesis.

Publications

nature COMMUNICATIONS

ARTICLE

Received 7 Dec 2015 | Accepted 5 Aug 2016 | Published 26 Sep 2016

DOI: 10.1038/ncomms12846 OPEN

Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd

Zichen Wang[1], Caroline D. Monteiro[1], Kathleen M. Jagodnik[1,2,3], Nicolas F. Fernandez[1], Gregory W. Gundersen[1], Andrew D. Rouillard[1], Sherry L. Jenkins[1], Axel S. Feldmann[1], Kevin S. Hu[1], Michael G. McDermott[1], Qiaonan Duan[1], Neil R. Clark[1], Matthew R. Jones[1], Yan Kou[1], Troy Goff[1], Holly Woodland[4], Fabio M.R. Amaral[5], Gregory L. Szeto[6,7,8,9], Oliver Fuchs[10], Sophia M. Schüssler-Fiorenza Rose[11,12], Shvetank Sharma[13], Uwe Schwartz[14], Xabier Bengoetxea Bausela[15], Maciej Szymkiewicz[16], Vasileios Maroulis[17], Anton Salykin[18], Carolina M. Barra[19], Candice D. Kruth[20], Nicholas J. Bongio[21], Vaibhav Mathur[22], Radmila D. Todoric[23], Udi E. Rubin[24], Apostolos Malatras[25], Carl T. Fulp[26], John A. Galindo[27], Ruta Motiejunaite[28], Christoph Jüschke[29], Philip C. Dishuck[30], Katharina Lahl[31], Mohieddin Jafari[32,33], Sara Aibar[34], Apostolos Zaravinos[35,36], Linda H. Steenhuizen[37], Lindsey R. Allison[38], Pablo Gamallo[39], Fernando de Andres Segura[40], Tyler Dae Devlin[41], Vicente Pérez-García[42] & Avi Ma'ayan[1]

Gene expression data are accumulating exponentially in public repositories. Reanalysis and integration of themed collections from these studies may provide new insights, but requires further human curation. Here we report a crowdsourcing project to annotate and reanalyse a large number of gene expression profiles from Gene Expression Omnibus (GEO). Through a massive open online course on Coursera, over 70 participants from over 25 countries identify and annotate 2,460 single-gene perturbation signatures, 839 disease versus normal signatures, and 906 drug perturbation signatures. All these signatures are unique and are manually validated for quality. Global analysis of these signatures confirms known associations and identifies novel associations between genes, diseases and drugs. The manually curated signatures are used as a training set to develop classifiers for extracting similar signatures from the entire GEO repository. We develop a web portal to serve these signatures for query, download and visualization.

[1] Department of Pharmacological Sciences, BD2K-LINCS Data Coordination and Integration Center, Illuminating the Druggable Genome Knowledge Management Center, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place Box 1215, New York, New York 10029, USA. [2] Fluid Physics and Transport Processes Branch, NASA Glenn Research Center, 21000 Brookpark Rd, Cleveland, Ohio 44135, USA. [3] Center for Space Medicine, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA. [4] Daylesford, the Fairway, Weybridge, Surrey KT13 0RZ, UK. [5] School of Biosciences, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Leicestershire LE12 5RD, UK. [6] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [7] David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [8] Department of Materials Science & Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [9] The Ragon Institute of MGH, MIT, and Harvard, 400 Technology Square, Cambridge, Massachusetts 02139, USA. [10] Paediatric Allergology and Pulmonology, Dr von Hauner University Children's Hospital, Ludwig-Maximilians-University of Munich, Member of the German Centre for Lung Research (DZL), Lindwurmstrasse 4, Munich 80337, Germany. [11] Spinal Cord Injury Service, Veteran Affairs Palo Alto Health Care System, Palo Alto, California 94304, USA. [12] Department of Neurosurgery, Stanford School of Medicine, Stanford, California 94304, USA. [13] Department of Research, Institute of Liver & Biliary Sciences, D1, Vasant Kunj, New Delhi 110070, India. [14] Department of Biochemistry III, University of Regensburg, Universitätsstrasse 31, Regensburg 93053, Germany. [15] Department of Pharmacology and Toxicology, University of Navarra, Pamplona, Irunlarrea 1, Pamplona 31008, Spain. [16] Warsaw School of Information Technology under the auspices of the Polish Academy of Sciences, 6 Newelska St, Warsaw 01-447, Poland. [17] Plomariou 1 St, 15126 Athens, Greece. [18] Department of Biology, Faculty of Medicine, Masaryk University, Brno 625 00, Czech Republic. [19] IMIM-Hospital Del Mar, PRBB Barcelona, Dr Aiguader, Barcelona 88.08003, Spain. [20] 85 Hailey Ln, Apt C-11, Strasburg, Virginia 22657, USA. [21] Department of Biology, Shenandoah University, 1460 University Dr Winchester, Winchester, Virginia 22601, USA. [22] IBM India Pvt Ltd., Bengaluru 560045, India. [23] Dr Aleksandra Sijacica 20, Backa Topola 24300, Serbia. [24] Department of Biological Sciences, 600 Fairchild Center, Mail Code 2402, Columbia University, New York, New York 10032, USA. [25] Center for Research in Myology, Sorbonne Universités, UPMC Univ Paris 06, INSERM UMRS975, CNRS FRE3617, 47 Boulevard de l'hôpital, Paris 75013, France. [26] 13-1, Higashi 4-chome Shibuya-ku, Tokyo 150-0011, Japan. [27] Department of Biology and Institute of Genetics, Universidad Nacional de Colombia, Bogota, Cr. 30 # 45-08, Colombia. [28] Center for Interdisciplinary Cardiovascular Sciences, Brigham and Women's Hospital, 3 Blackfan Circle, Boston, Massachusetts 02115, USA. [29] Department of Human Genetics, Faculty of Medicine and Health Sciences, University of Oldenburg, Ammerländer Heerstrasse 114-118, Oldenburg 26129, Germany. [30] 2312 40th ST NW #2, Washington DC 20007, USA. [31] Technical University of Denmark, National Veterinary Institute, Bülowsvej 27 Building 2-3, Frederiksberg C 1870, Denmark. [32] Protein Chemistry and Proteomics Unit, Biotechnology Research Center, Pasteur Institute of Iran, No. 358, 12th Farwardin Ave, Jomhhoori St, Tehran 13164, Iran. [33] School of Biological Sciences, Institute for Researches in Fundamental Sciences, Niavaran Square, P.O.Box, Tehran 19395-5746, Iran. [34] University of Salamanca, Salamanca, Madrid 37008, Spain. [35] Division of Clinical Immunology, Department of Laboratory Medicine, Karolinska Institute, Alfred Nobels Allé 8, level 7, Stockholm SE141 86, Sweden. [36] Department of Life Sciences, School of Sciences, European University Cyprus, 6 Diogenes Str. Engomi, P.O.Box 22006, Nicosia 1516, Cyprus. [37] Anna Blamansingel 216, Amsterdam 102 SW, Netherlands. [38] 7300 Brompton #6024, Houston, Texas 77025, USA. [39] Aligustre 30 1-C, Madrid 28039, Spain. [40] CICAB, Clinical Research Centre, Extremadura University Hospital, Elvas Av., s/n. 06006 Badajoz 06006, Spain. [41] 69 Brown Street, Box 8278, Providence, Rhode Island 02912, USA. [42] Consejo Superior de Investigaciones Científicas, Centro Nacional de Biotecnología, Department of Immunology and Oncology, c/Darwin, 3 Madrid 28049, Spain. Correspondence and requests for materials should be addressed to A.M. (email: avi.maayan@mssm.edu).

Omics repositories such as the NCBI Gene Expression Omnibus (GEO)[1] and EBI ArrayExpress[2] accumulate and serve gene expression data from thousands of studies. It is clear that these data contain much more information than what has typically been extracted from each individual dataset for the accompanying initial publication. However, currently, performing integrative analysis of large collections of gene expression studies to obtain a global integrated view of cellular regulation requires a significant data wrangling effort, that is, manually unifying data formats, adding metadata and converting the data to be more machine readable.

Due to high cost, gene expression profiling data are typically produced on a small scale, in targeted studies that are diverse with respect to tissue or cell type, genetic or chemical perturbation, disease model, expression assay platform and model organism. When submitted into public repositories such as GEO, the requirement for metadata annotation is minimal. Lack of standards for extensive metadata collection, and the diversity of individual studies, prohibits the easy reuse and integration of this type of data.

One of the advantages of carefully annotating studies from databases such as GEO is the potential for developing a signature search engine that operates at the data level. Tools such as SIGNATURE[3], SPIED[4], Cell Montage[5], ProfileChaser[6], ExpressionBlast[7] and SEEK[8] automatically attempt to compute differentially expressed signatures from GEO to provide a signature search engine at the data level. However, these tools are prone to mistakes because they automatically select the control and perturbation samples, as well as other aspects of signature generation and annotation, without relying on an extensive high-quality gold standard, which is needed for training better-quality classifiers.

Manual extraction of collections of gene expression signatures from GEO has been demonstrated to be highly useful. It was applied for drug repurposing[9], suggesting novel drugs for many diseases[10], and explaining mechanisms of action for many approved drugs[11]. Several efforts have attempted to further annotate datasets from GEO manually; one example is Gene Expression data Mining Toward Relevant Network Discovery (GEM-TREND)[12]. The disadvantage of manual curation is that it does not scale up to cover the thousands of studies currently available. For similar challenges, crowdsourcing projects have been developed as a potential solution to overcome this obstacle.

Crowdsourcing projects fall into two categories: microtasks and megatasks[13,14]. Microtasks consist of relatively trivial tasks that require a large number of participants; for example, extracting features from images of cells[15]. Crowdsourcing microtask projects in biomedical research have been established to improve automated mining of biomedical text for annotating diseases[16], curation of gene-mutation relations[17], identifying relationships between drugs and side-effects[18], drugs and their indications[19], as well as annotation of microRNA functions[20]. These efforts produce large collections of high-quality datasets that can be further utilized by algorithms that can extract new knowledge from already-published data that require better annotation, cleaning and reprocessing.

When computing gene expression signatures, the computational method used to identify the differentially expressed genes (DEGs) has a significant impact on the results. Using several benchmarks, including matching expression changes after transcription factor perturbations with ChIP-seq data, we previously showed that a method we developed called the Characteristic Direction (CD) significantly improves the prioritization of differentially expressed genes[21] when compared with several commonly applied methods such as fold change, *T*-test or ANOVA, SAM[22], *limma*[23] or DESeq[24].

In this study, we present the results of a crowdsourcing microtask project implemented to annotate and extract gene expression signatures from GEO. Our analysis of the crowdsourced gene expression signatures demonstrates that our collection of signatures is of high quality and can be used to recover prior knowledge, as well as discover new knowledge, about associations between drugs, genes and diseases. We also develop a web portal for users to visually identify associations between signatures, download the signatures for further computational analyses, and search the collections of gene expression signatures created for this project with their own signatures or by keywords. To scale up the collection of signatures for the three themes: disease, drug and gene perturbation, we use the manually extracted signature collections as a gold standard to train classifiers that automatically extract signatures from GEO.

Results

Crowdsourcing gene expression signatures. The crowdsourcing challenge we designed followed several steps and consisted of several components and processes (Fig. 1). First, participants were asked to identify GEO studies in which single-gene or -drug perturbations were applied to mammalian cells, or in which normal versus diseased tissues were compared. After identifying relevant studies, participants extracted metadata from the studies and computed differential expression using GEO2Enrichr[25], a Chrome extension we developed that makes the signature extraction process easy for non-experts. Extracted signatures were stored in a local database and sanitized by automated filters and manual inspection for improving accuracy and quality. The cleaned database of extracted signatures was used to visualize and analyse these signatures on the CRowd Extracted Expression of Differential Signatures (CREEDS) web portal. To scale up the collections, the human-extracted signatures were used as a gold standard for training machine learning classifiers for automated signature extraction. To date, the manual component of the signature database contains 3,100 submissions for single-gene perturbations, covering 1,186 genes from 1,635 studies; 1,081 disease signature submissions covering 450 diseases from 748 studies; as well as 1,238 submissions for drug perturbations covering 343 drugs from 443 studies (Supplementary Fig. 1a). After sanitizing the collections of signatures, a total of 2,177; 828 and 1,221 unique and valid signatures remained in the CREEDS database for single-gene perturbations, disease signatures, and drug perturbation signatures, respectively. The automated expansion of the signatures resulted in an additional set of 8,620 single-gene, 1,430 disease and 4,295 single-drug signatures extracted from 2,543 GEO studies.

We observe a skewed distribution with a long tail for the number of submissions per contributor (Supplementary Fig. 1b). A few enthusiastic curators contributed many more signatures than most others. The median number of signatures submitted per person was 16. We found no significant correlation between the number of signatures submitted per user and the quality of submissions (Supplementary Fig. 1c, Spearman's $\rho = -0.08$, P value $= 0.42$). The leaderboard generally incentivized volunteers to submit more gene expression signatures. We found a significant negative correlation (Spearman's $\rho = -0.64$, P value $< 8.0e^{-51}$) between the scaled ranks of contributors and the number of newly submitted studies per day (Supplementary Fig. 1d). This suggests that highly ranked curators were inclined to continue to submit more.

Quality improvement of crowdsourced gene expression signatures. To improve the quality of the gene expression signatures derived from thousands of GEO studies, we first checked for batch effects.

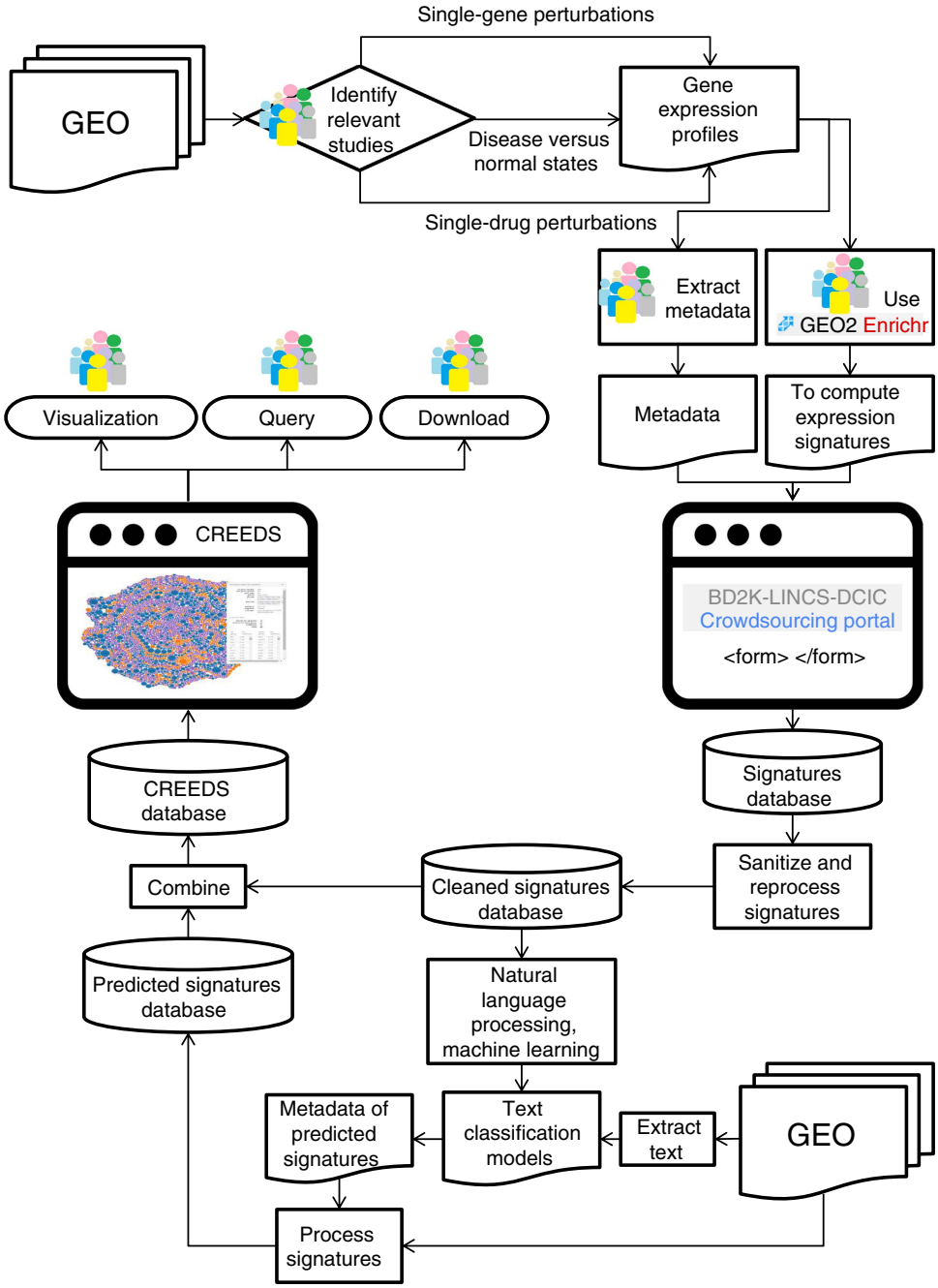


Figure 1 | Workflow of the crowdsourcing project. Participants identify relevant studies from GEO and then extract gene expression signatures using GEO2Enrichr. Participants also add metadata to each signature. Submitted signatures were manually reviewed and then used to scale up the collections with machine learning methods. All signatures are served on the CRowd Extracted Expression of Differential Signatures (CREEDS) web portal.

To achieve this, we obtained the 'scan date' from the raw microarray data files as an indicator of a potential source for batch effects. We then estimated the magnitude of such batch effect using principal variation component analysis[26,27]. We estimate that batch effects on average account for ~18.7% of the variance in the gene expression dataset collections, whereas the perturbation versus control on average accounts for ~16.7% of the variance (Supplementary Fig. 2a).

To correct for these batch effects, we applied the surrogate variable analysis (SVA)[28] algorithm and generated new signatures using both the CD and *limma* methods to call the DEGs. To benchmark the quality of these signatures with or without the batch correction, we used collections of genes that are expected to be differentially expressed: direct protein interactions for gene perturbation, disease-gene associations for disease signatures, and targets of drugs for the drug-induced signatures. We observe that the batch correction improves the signal and quality of signatures (Fig. 2). We also found that the CD method outperformed *limma* in ranking the expected DEGs with these benchmarks.

Comparing the collections with other similar resources. Next, we compared the collection of the crowdsourced gene expression signatures with MSigDB[29], which contains 8 collections of gene sets. The collection C2 has curated gene sets extracted manually from tables and figures within publications. We compared the

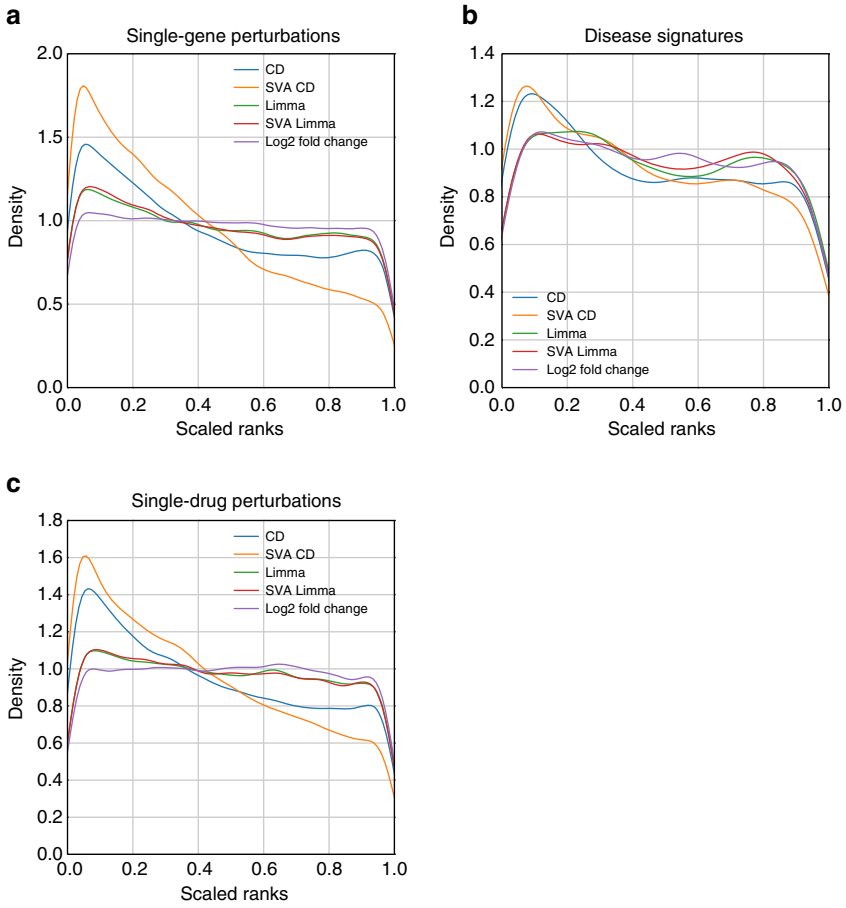


Figure 2 | Batch effect correction influence on the quality of gene expression signatures. Line plots show the probability density distribution of the scaled ranks of expected DEGs in gene expression signatures from the three collections: (**a**) single-gene perturbations, (**b**) disease signatures, and (**c**) single-drug perturbations. The colours indicate which algorithm was used to call the differentially expressed genes: Characteristic Direction (CD), *limma*, or fold change; and whether batch effect correction was applied with surrogate variable analysis (SVA).

Chemical and Genetic Perturbations (CGP) subset within C2 from the latest version of MSigDB (v5.1) with our collections of signatures. The CGP subset has 3,396 gene sets, 33% of which have GEO identifiers (GSE) (Supplementary Fig. 3a). We first compared the overlapping GSEs and found that our collection covers 2,066 microarray studies, whereas the CGP subset covers 361 microarray studies with 54 shared studies (Supplementary Fig. 3b). Breaking down the overlap into the three collections, the shared GSEs with MSigDB are 31, 21 and 7 for the gene, disease and drug perturbations, respectively (Supplementary Fig. 3b). To compare the concordance of the gene-set for the 31 shared gene perturbations, we plotted the cumulative distribution from uniform distribution of the scaled ranks of the genes from our collection and those matching from MSigDB, and found that these gene sets are significantly similar (Supplementary Fig. 3c). Overall, we find that the MSigDB signatures overlap significantly with matched crowd-generated signatures, with only a few exceptions (Supplementary Fig. 3d, Supplementary Table 1). The discrepancies were due to a figure from He *et al.*[30] that only reported genes related to the cell-cycle as opposed to all DEGs; the Sagiv *et al.*[31] study reported DEGs in both siRNA knockdown and mAb treatment, whereas the DEGs in our database were derived from knockdown versus control only; and the gene sets curated from Soucek *et al.*[32] by MSigDB do not match the original figure from that paper. However, overall, our analysis shows strong agreement between the matched signatures in both databases.

Assessment of signature associations within each collection. We next asked whether signature similarity within and across the three collections can recover prior knowledge and discover novel connections. To globally assess associations between signatures within each collection, we used various methods to compute similarity between all pairs of signatures, and compared ranked signature associations with prior knowledge. Our results show that all of the three signature collections recover prior knowledge associations between genes, drugs and diseases (Supplementary Tables 2–4), and these associations are more discernable when computing differential expression with the CD method (Fig. 3). For example, individual independent studies that perturbed *Prkag3* by either knockout or gain-of-function mutation were identified as opposing signatures[33] (Supplementary Table 2). An example that emerged from comparing disease signatures was the high similarity between hypercholesterolaemia and hepatocellular carcinoma signatures (Supplementary Table 3). It was shown that cholesterol metabolism is indeed deregulated in hypercholesterolaemia and hepatocellular carcinoma[34,35]. There are some top-ranked drug pairs that induce similar gene expression changes. For instance, the gene expression signatures for diethylstilbestrol, estradiol and tamoxifen from independent studies are very similar (Supplementary Table 4). The confirmation with prior knowledge associations suggests that we can predict novel associations with these data. In other words, top-ranked associations or top-ranked opposing signatures between drugs, diseases or genes that do not have literature

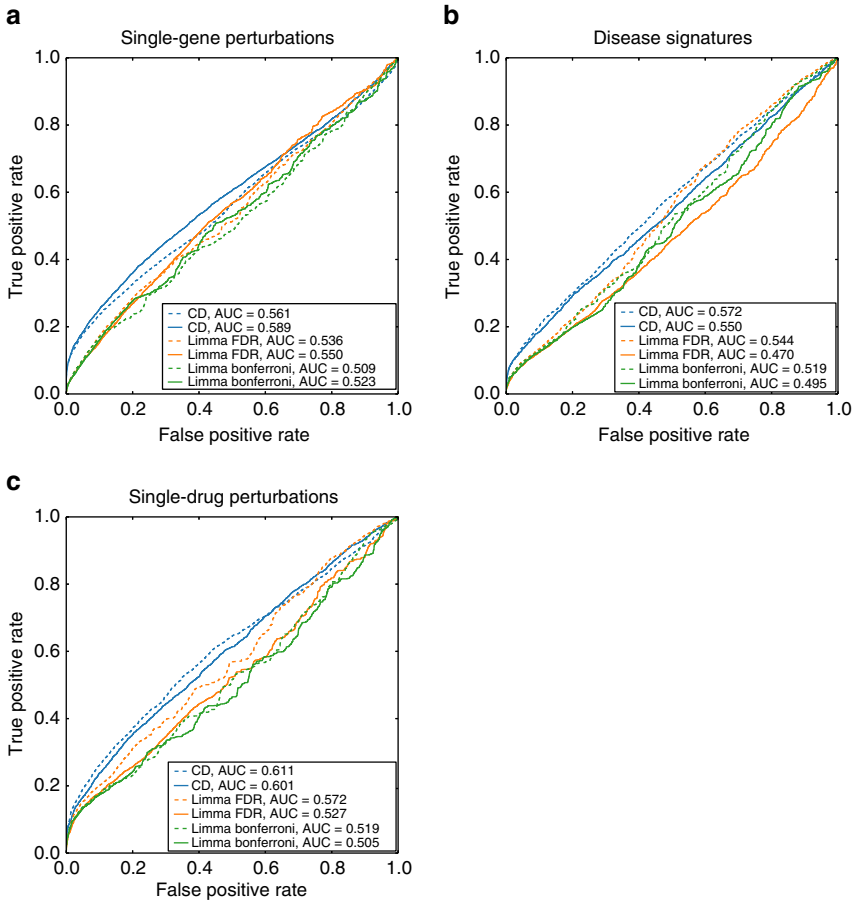


Figure 3 | Benchmarking signature connections with prior knowledge. Signed Jaccard index and absolute Jaccard index are used to measure the similarity between signatures, and plotted in dashed and solid lines, respectively. Different methods for identifying differentially expressed genes include: the Characteristic Direction (CD), *limma* with Benjamini–Hochberg (BH) correction, and *limma* with Bonferroni correction. These are plotted in blue, orange and green, respectively. ROC curves are plotted for (**a**) recovering the same perturbed genes; (**b**) recovering similar diseases; and (**c**) recovering drugs with similar chemical structure.

support should be considered as high-quality predictions. Given the observation that drugs with highly similar chemical structure induce slightly more similar gene expression signatures than expected by chance (Fig. 3c), we further investigated whether the correlation between chemical similarity and gene expression signature similarity also applied to drugs pairs with lower chemical similarity scores. By binning the signed Jaccard index by Tanimoto coefficients, we found no correlation between lower chemical similarity and gene expression signature similarity (Supplementary Fig. 4), suggesting that partial chemical similarity is not predictive of expression similarity.

Signature associations across the three collections. Using the signed Jaccard index, we computed an adjacency matrix for all possible pairs of signatures from the three collections (Fig. 4a) and observed many clusters. These clusters are heterogeneous, containing connections between genes, diseases and drugs. We highlight a few of these clusters (Fig. 4c,d), while others can be explored using the interactive clustergram or packed circles plot on the CREEDS web portal. In the first cluster that we chose to highlight, imatinib, a small molecule that is known to be a tyrosine kinase inhibitor[36], has signatures that were generated from multiple cell lines, including K562 leukaemia cell line (GSE1922), chronic myelogenous leukaemia (CML) CD34 + cells (GSE12211) and three other CML cell lines (KU-812, KCL-22, JURL-MK1) (GSE24493), which cluster together with knockdown signatures of *NRAS* in melanoma cell lines (GSE12445) (Fig. 4b). This strongly suggests that *NRAS* is targeted by imatinib. Although *NRAS* is currently not considered a direct target of imatinib, a recent study showed that melanoma patients with *NRAS* mutations are resistant to imatinib therapy[37]. This raises the possibility that the wild-type form of *NRAS* is at least a key downstream effector of imatinib.

In the second cluster that we chose to highlight, multiple myelodysplastic syndrome (MDS) signatures from CD34 + cells (GSE4619, GSE19429) and *ERBB2* overexpression signature from MCF10A cells (GSE14990) cluster together (Fig. 4c), suggesting that the up-regulation of *ERBB2* may have a role in MDS. Indeed, it was shown that *ERBB2* amplification is present in 35% of a cohort of MDS patients[38]. In the third example, endometrial cancer signatures (GSE17025) are shown to cluster with estradiol signatures derived from MCF7 cells from multiple independent studies (GSE4668, GSE11352, GSE53394), as well as *MIR34A* overexpression signature from HCT116 cells (GSE7754), *PPARG* overexpression signature from NIH-3T3 cells (GSE2192), and IGF1 stimulation signature from MCF7 cells (GSE7561) (Fig. 4d). Estradiol has been shown to increase the risk for endometrial cancer[39,40] and was previously discovered in a meta-analysis study of this disease[41]. Insulin-like growth factor 1 (*IGF1*) and its receptor *IGF1R* are known to be indirectly activated by estradiol[42–44]. Downstream of the *IGF1R* receptor

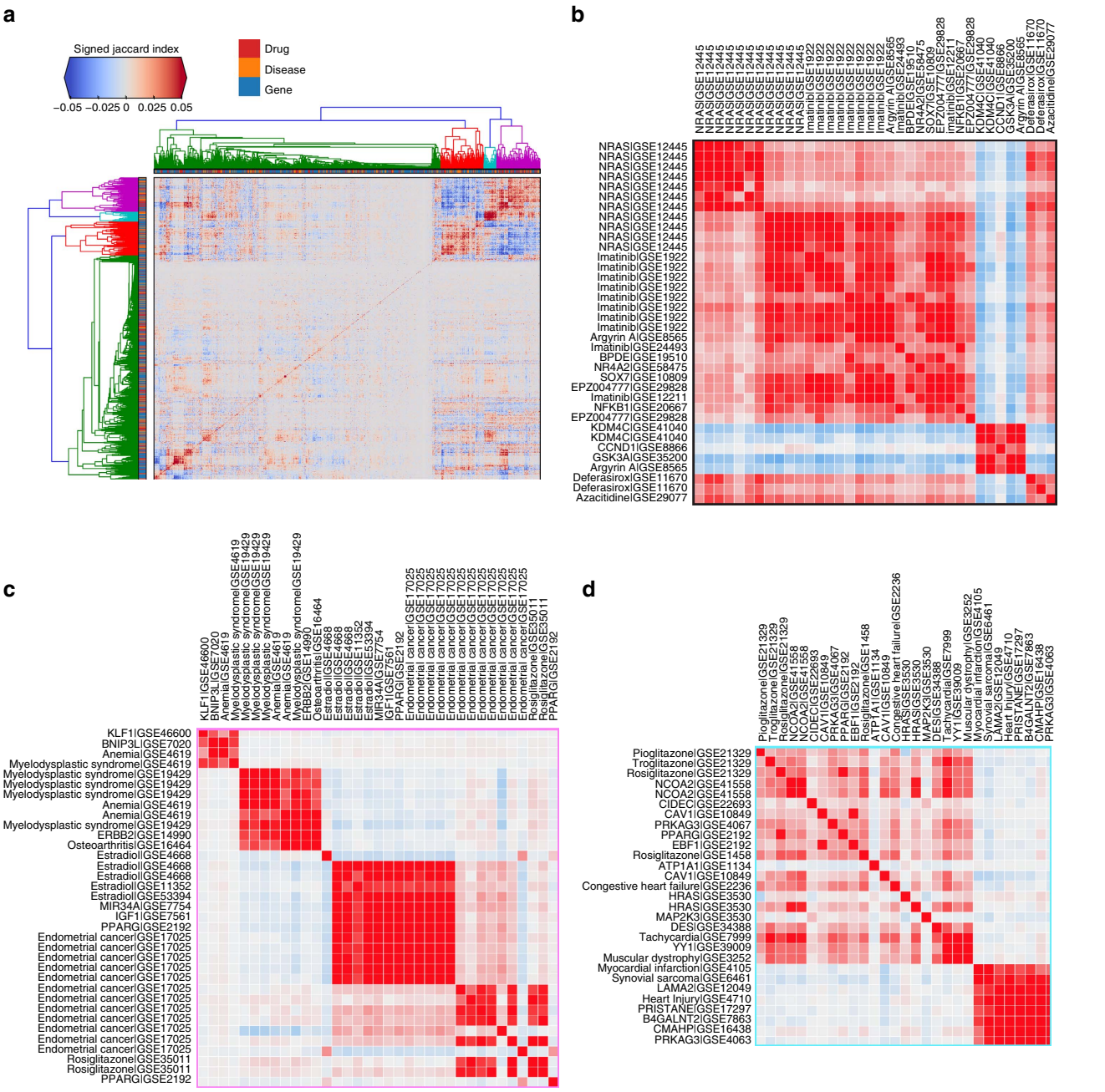


Figure 4 | Hierarchical clustering of the adjacency matrix of all gene expression signatures and selected clusters. (**a**) The entire adjacency matrix of all signatures. (**b**–**d**) Three selected zoomed-in views of clusters from the adjacency matrix displayed in (**a**).

phosphoinositide kinase 3 (*PI3K*), the mammalian target of rapamycin (*mTOR*) and MAPK signalling promote protein synthesis, cell growth, and cell proliferation, potentially driving the progression of endometrial cancer[45,46]. Peroxisome proliferator-activated receptor gamma (*PPARG*) has also been shown to induce the development of multiple types of cancers[47], and it is known to play a role downstream of adiponectin during insulin resistance[48], which is a significant risk factor for endometrial cancer[49]. The fourth cluster contains a *YY1* knockout (GSE39009) signature produced in mice soleus, and an autosomal muscular dystrophy signature from a mouse model sourced from the diaphragm (GSE3252). This association suggests that *YY1* may be disrupted in muscular dystrophy tissues. Literature supports that almost all facioscapulohumeral muscular dystrophy patients carry deletions of repetitive elements (D4Z4) that contain binding sites for *YY1*[50,51]. All of the aforementioned examples are just a small portion of the signature connections our integrative analysis offers. These examples illustrate how novel associations between diseases, genes and drugs can be discovered through a crowdsourcing project.

Identifying drug mimickers. To further demonstrate the utility of the crowdsourced gene expression signatures of drug perturbations, we queried these signatures against the database of drug or other small molecule compound signatures derived from the LINCS L1000 dataset. We then recorded the ranks of the matched drugs out of >30,000 LINCS L1000 signatures and found that

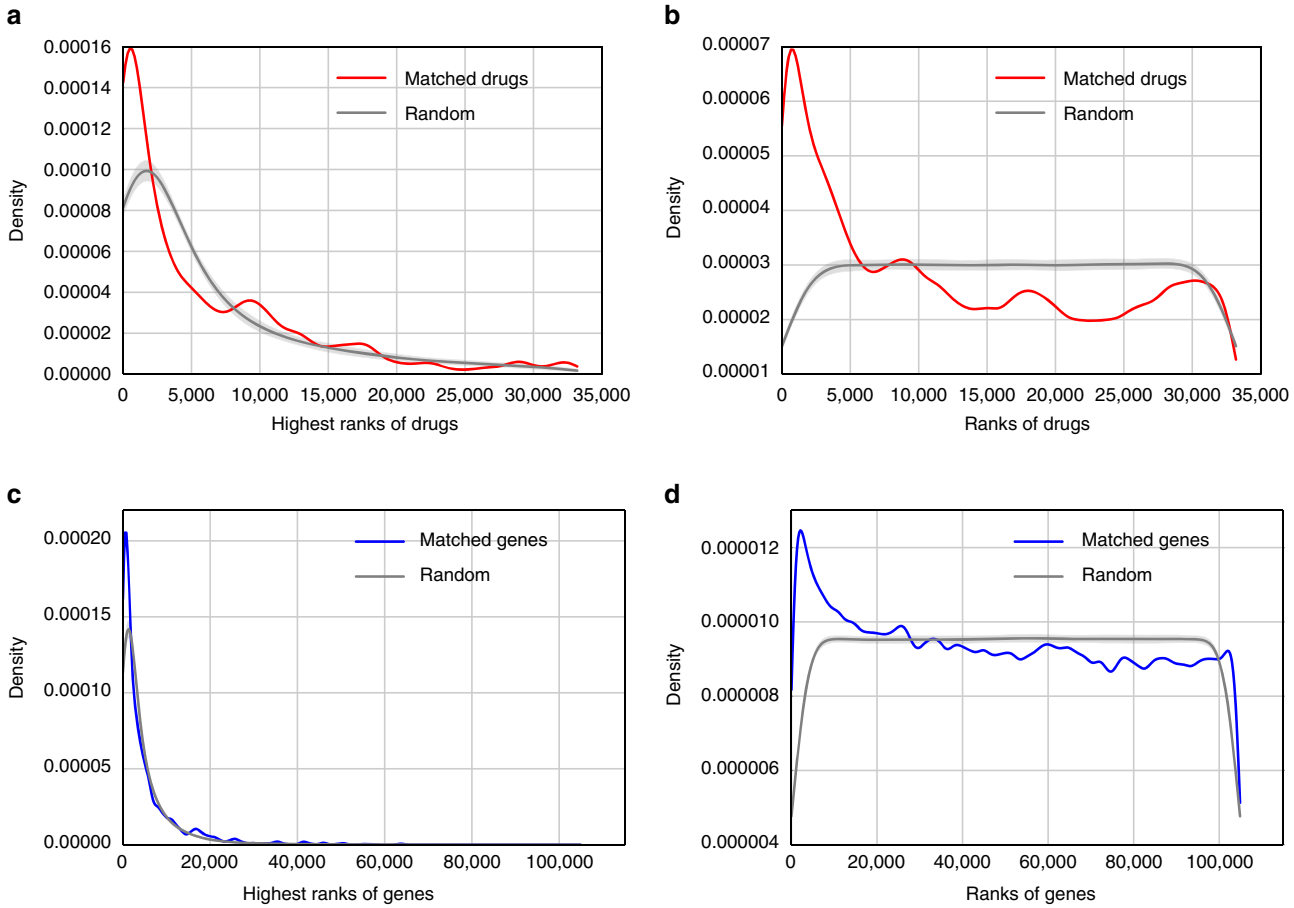


Figure 5 | Distributions of the ranks of matched perturbations between signatures from CREEDS and the LINCS L1000 dataset. The highest ranks (**a**,**c**), and all ranks (**b**,**d**) of matched drugs (**a**,**b**) and matched genes (**c**,**d**) are presented. Drug perturbation signatures from CREEDS were queried against ~30,000 significant drug perturbation signatures from the LINCS L1000 dataset; whereas gene perturbation signatures from CREEDS were queried against ~110,000 gene perturbation signatures from the LINCS L1000 dataset.

many crowdsourced drug perturbation signatures are significantly highly ranked (Rank sum P value $<4.8e^{-24}$) (Fig. 5a,b, Table 1). Similarly, the results can also be reproduced when querying the drug perturbation signatures against >6,000 signatures from the Connectivity Map dataset[52] (Supplementary Fig. 5). We additionally queried the gene perturbation signatures against 109,000 shRNA knockdown and over-expression profiles from the LINCS L1000 data and found similar consistency (Fig. 5c,d). These results suggest that some drugs induce similar transcriptional changes in small-scale studies, when compared with results from large-scale studies such as LINCS L1000 and the original Connectivity Map. This means that we can identify potential mimickers using the LINCS L1000 dataset for drugs whose signatures are highly similar between the LINCS L1000 dataset and the GEO studies. Interestingly, we found that dexamethasone signatures in the LINCS L1000 dataset were ranked in the top 10 using dexamethasone-induced gene expression signatures from three independent GEO studies: GSE34313, GSE7683 and GSE54608 (Supplementary Table 5). The three studies treated dexamethasone in different cell types: human airway smooth muscle cells, mice primary chondrocytes, and in a human oviductal cell line, suggesting that the effect of this glucocorticoid agonist is robust across mammalian cells. Among the top-ranked potential mimickers of dexamethasone, flumetasone and betamethasome are both corticosteroids indicated for inflammation, confirming that the approach is able to identify drugs with similar physiological effects. Moreover, we found a small molecule compound 5,6-epoxycholesterol (BRD-K61480498) with gene expression profiles highly similar to that of dexamethasone. 5,6-epoxycholesterol also has a similar chemical structure, but unknown anti-inflammatory effects. As such, it is an example of a strong candidate for further experimental validation.

Web portal to visualize and query the signatures database. To provide easier access to the three collections of the gene expression signatures for knowledge reuse and exploration, we developed a web portal (Supplementary Fig. 6). This portal visualizes all of the signatures in a packed circles layout in which similar signatures are closer to each other. Furthermore, the portal has interactive heatmaps of hierarchically clustered matrices of all signatures. The web portal is available at: http://amp.pharm.mssm.edu/creeds. The portal also has a search engine that enables users to search by text or by providing lists of up and down DEGs. Since DEGs for the gene expression profiles in the CREEDS database were computed with the CD method, which is not a standard method, we tested whether signatures computed via other methods would produce similar results. We found that most signatures computed by fold change or *limma* are ranked similarly (Supplementary Fig. 7). However, some signatures were not ranked as expected. The CD is a multivariate method, whereas fold change and *limma* are univariate; a gene can be identified as significantly differentially expressed by a

Table 1 | Top hits for drug signatures extracted from GEO queried against drug perturbations from the LINCS L1000 dataset processed using the Characteristic Direction method.

Drug name	PubChem ID	GEO Accession	organism	GEO platform	Rank
Dexamethasone	5743	GSE34313	human	GPL6480	1
Doxorubicin	31703	GSE58074	human	GPL10558	1
Azacitidine	9444	GSE29077	human	GPL571	1
Azacitidine	9444	GSE29077	human	GPL571	1
Azacitidine	9444	GSE29077	human	GPL571	1
Lapatinib	208908	GSE38376	human	GPL6947	2
Methylprednisolone	6741	GSE490	rat	GPL85	2
Lapatinib	208908	GSE38376	human	GPL6947	2
Dexamethasone	5743	GSE54608	human	GPL10558	3
Lapatinib	208908	GSE38376	human	GPL6947	3
Tretinoin	444795	GSE1588	mouse	GPL81	3
Methylprednisolone	6741	GSE490	rat	GPL85	3
Tretinoin	444795	GSE32161	human	GPL570	3
Methylprednisolone	6741	GSE490	rat	GPL85	3
Methylprednisolone	6741	GSE490	rat	GPL85	4
Trichostatin A	444732	GSE1437	mouse	GPL81	4
Dexamethasone	5743	GSE7683	mouse	GPL1261	5
Cycloheximide	6197	GSE8597	human	GPL570	5
Methylprednisolone	6741	GSE490	rat	GPL85	6
Sorafenib	216239	GSE39192	human	GPL6947	7
Vemurafenib	42611257	GSE37441	human	GPL10558	8
Methylprednisolone	6741	GSE490	rat	GPL85	10
Curcumin	969516	GSE10896	human	GPL570	14
Curcumin	969516	GSE10896	human	GPL570	15
Vemurafenib	42611257	GSE37441	human	GPL10558	15
Lapatinib	208908	GSE38376	human	GPL6947	16
Methylprednisolone	6741	GSE490	rat	GPL85	17
Tretinoin	444795	GSE1588	mouse	GPL81	20
Vemurafenib	42611257	GSE42872	human	GPL6244	23
Azacitidine	9444	GSE29077	human	GPL571	24
Troglitazone	5591	GSE21329	rat	GPL341	31
Decitabine	451668	GSE29077	human	GPL571	36
Vemurafenib	42611257	GSE37441	human	GPL10558	36
Thapsigargin	446378	GSE19519	human	GPL570	37
Methylprednisolone	6741	GSE490	rat	GPL85	48

univariate method but may not contribute to the joint expression changes of large sets of genes.

Finally, to scale up the three collections of signatures, we developed machine learning classifiers that use the manually curated signatures as a training set. The classification task was divided into two parts: (1) classify whether a GEO dataset is likely to contain gene, disease or drug signatures, and (2) label the samples as control and perturbation. The features for the classifiers were extracted from the text associated with the each GEO study in our manually curated collection as well as from all currently available studies on GEO where genome-wide expression was assessed by microarrays to profile human, mouse or rat cells and tissues. Overall, we observe that various classifiers perform very well (Supplementary Fig. 8).

We next asked whether we have collected a sufficient number of manually curated studies or whether more manual curation could improve the performance of the classifiers. We see, for example, that Naïve Bayesian classifiers no longer improve once ~1,000 annotated studies are used for each collection category (Supplementary Figs 9–13). With these machine learning classifiers, we automatically identified a large collection of additional signatures for the three collections. In total, this process enabled us to add 8,620 gene; 4,295 drug and 1,430 disease automatically extracted signatures. Each signature carries a *P*-value for confidence, and all these signatures are available for download and search on the CREEDS web portal.

Discussion

Gene expression profiling is arguably the most common type of omic data. The resource we developed for this project can be combined with transcriptomics profiling projects such as Genotype-Tissue Expression[53], the Cancer Genome Atlas[54], the Cancer Cell-Line Encyclopaedia[55], and the Library of Integrated Network-based Cellular Signatures (LINCS). Here we show, for example, how combining drug perturbation signatures collected from GEO with the LINCS L1000 data can be used to identify potential novel drug mimickers.

The manually extracted and cleaned signatures were proven to be useful as a training set that enabled us to scale up the three collections of signatures using machine learning. However, we are aware that the quality of the automatically generated signatures is not as good as the signatures created by the human annotators. One solution to improve the process is to intelligently integrate machine learning with crowdsourcing by using active learning. With active learning, unlabelled instances are presented to human annotators with suggestions; this allows the classifiers to be improved dynamically while reducing the effort required of the curators[56]. Active learning methods have been shown to achieve improved performance in similar settings[57,58].

This project highlights the commitment of citizen scientists to spare their time in pursuit of a common goal that can advance science and medicine. Indeed, we show how this collective effort was used to identify novel relationships between genes, drugs and diseases. While we highlighted several top predictions that

emerged from our analysis, many more hypotheses can be formed by interacting with the CREEDS portal at: http://amp.pharm.mssm.edu/creeds.

Methods

Extracting gene expression signatures from GEO by the crowd. Three crowdsourcing microtasks were established to collect gene expression signatures from GEO. These are: single-gene perturbations, comparison between diseased and normal tissues, and single-drug perturbations. These three types of signatures were extracted using the Google Chrome extension GEO2Enrichr[25] and submitted through the BD2K-LINCS-DCIC Crowdsourcing Portal at: http://www.maayanlab.net/crowdsourcing/. These crowdsourcing tasks were open to all participants, but a significant majority of the contributors were students from the massive open online course Network Analysis in Systems Biology 2015 (NASB2015) offered on the Coursera platform. These participants were given detailed instructions for finding, labelling, and extracting gene expression profiles from GEO. Participation was strictly voluntary, and was not required for completion of any parts of the course. Participants were not provided with a list of predefined gene expression profiles; instead, they were encouraged to find diverse, yet relevant, gene expression studies from GEO. Briefly, contributors first had to locate relevant GEO studies fitting into one of the three themes, and then select the perturbation and control samples (GSMs) from GEO series (GSE) or GEO datasets (GDS). Only gene expression studies from selected species of mammals (human, mouse and rat) were considered valid. Participants were also asked to submit additional metadata about the cell or tissue type, and gene, disease or drug used in each experiment and associate these with common published identifiers. Standard names of genes, diseases, and drugs were provided as autocomplete options in the submission forms, created from controlled vocabularies: HGNC for genes[59], disease names from the Disease Ontology[60] and drug names from DrugBank[61]. To incentivize participants, a real-time leaderboard was developed to display the number of submissions from each user, and modest prizes were promised to the top ten contributors (custom T-shirt and headphones). Additionally, co-authorship on the published research resulting from these crowdsourcing tasks was promised to contributors of a minimum of 15 valid entries.

Sanitization of the crowdsourced gene expression signatures. Multiple steps of quality control filters were applied to improve the collection of the gene expression signatures extracted by the crowd. We first performed integrity checks using the association between GEO studies (GSE or GDS) and samples within these studies (GSMs) by re-processing all the collected gene expression signatures based on the metadata supplied by the curators. Signatures in which GSMs did not match their GSE or GDS, as well as signatures with the same GSMs in the control and perturbation groups, were automatically detected and removed. The next filter was applied only to the single-gene perturbation collection. We checked whether gene symbols submitted by the curators are valid HGNC gene symbols, removing all entries with invalid genes. The next filter was semi-automatic: we corrected signatures in which the control and perturbation samples were switched. Our final filter was to manually check if the submitted signatures agree with the descriptions associated with the original GEO studies. After applying each of these filters, we recorded the number of invalid submissions by curators and removed the submissions from any curators who had submitted more than 10% invalid signatures. As a result, $\sim$20% of all the submissions were removed from the final collections.

Evaluation of batch effects. To obtain batch information from each study, we retrieved the 'scan date' from the raw microarray CEL files and assumed that the experiments were performed on the same dates that were listed within the experimental batch. We then quantified the batch effect using principal variation component analysis[26,27], which attributes the variation in the gene expression data to known sources such as batches and experimental conditions. Batch effects were corrected using the surrogate variable analysis (SVA) algorithm[28] implemented in R[62] with default parameters.

Construction of expected DEGs from prior knowledge. To generate lists of expected DEGs for the three collections of signatures for benchmarking, we used: (1) the known direct physical interactors of the protein product of a gene from a consolidated protein–protein interaction network we assembled for a previous study[63]; (2) a consolidated collection of manually-curated disease-gene associations from the DISEASES resource[64]; and (3) known drug targets from DrugBank v4.3[61].

Measuring similarity between signatures. To compare signatures, we abstracted signatures to sets of up- and down-regulated genes. The signed Jaccard index for two signatures S_i and S_j is defined as:

$$SJ(S_i, S_j) = \frac{J\left(S_i^{up}, S_j^{up}\right) + J\left(S_i^{down}, S_j^{down}\right) - J\left(S_i^{up}, S_j^{down}\right) - J(S_i^{down}, S_j^{up})}{2}$$

where $S^{up}$and S^{down} denote the up- and down-regulated gene sets, respectively. The signed Jaccard index considers the direction when comparing a pair of gene expression signatures. It has a range of $[-1, 1]$ where 1 represents identical signatures, and -1 represents signatures of reverse effect, whereas 0 represents unrelated signatures.

Signature pairs from different GEO studies were ranked based on the signed Jaccard index. Prior knowledge from various resources about known connections between genes, diseases and drugs was used to examine whether signature similarity can be used to recover known associations between genes, drugs and diseases. Specifically, pairs of diseases were connected through the Disease Ontology[60], and pairs of drugs were connected by the drugs' molecular structure fingerprints and considered similar if the Tanimoto coefficient was >0.9. Structural fingerprints were computed with the extended-connectivity fingerprints ECFP4[65]. To score the predictions of associations between genes, drugs and diseases, receiver operating characteristic (ROC) curves were plotted and the area under the ROC curve (AUC) was calculated. DeLong's test[66] was performed to compare the difference between ROC curves.

Natural language processing of text from GEO series. The text from each GEO series including title, summary, and keywords were extracted and processed separately. Text was first tokenized into words that were then lemmatized using the WordNet Lemmatizer[67] and stemmed using the Porter stemming algorithm[68]. Term frequency-inverse document frequency (TF-IDF)[69] was used to convert stems of both unigrams and bigrams into numerical values that measure the importance of an n-gram to a document in the context of the collection of documents. Truncated singular value decomposition was used to reduce dimensionality of the TF-IDF matrices to capture at least 10% of the variance. To visualize the GEO studies in the textural feature space, t-Distributed Stochastic Neighbour Embedding[70] was used to reduce the dimensionality of the matrices from the truncated singular value decomposition. To classify whether a GEO series contains a disease signature, three textural feature matrices representing the title, summary and keywords were used to train and test a classifier. To measure the performance of the classification, three-fold cross-validation was applied to calculate the area under the ROC curve, area under the precision-recall curve, Matthew's correlation coefficient and F1 score. Classifiers from the scikit-learn[71] package were tested including: random forest[72], extra trees[73], support vector classifier and the XGBoost implementation of gradient boosting machines[74]. Hyperparameters of the classifiers were optimized using grid search.

Classifying control versus treatment samples based on text. We formulate the problem of classifying GEO samples as a binary classification problem. This means that we aim to learn from text-derived features whether a sample is part of the control or treatment group. Features were extracted from the following text fields associated with each GEO sample: title, description, characteristics and source name. These text elements were tokenized and converted to binary vectors representing the presence or absence of tokens for each sample. The classifier we used for solving this problem is a Bagging[75] of 20 multinomial Bernoulli Naïve Bayesian[69] classifiers after probability calibration with isotonic regression[76]. To measure the performance of the classifier, 10-fold cross-validation was applied to calculate area under the ROC curve, area under the precision-recall curve, Matthew's correlation coefficient and F1 score.

Development of the CREEDS web portal. A web portal was developed for visualizing and querying the collections of the gene expression signatures. Relationships between all signatures are visualized using the D3.js pack layout and D3.js clustergrammer. Clustergrammer is a visualization tool we developed starting with the open-source code example for the matrix co-occurrence visualization on the D3.js website. All data and metadata of the signatures are stored in a MongoDB database. The portal uses the Python Flask framework. Signed Jaccard index was implemented to query signatures in which users input up or down gene lists into two separate text boxes. The text signature search option queries the metadata text of all signatures in the database. RESTful application programming interface (API) endpoints were also developed to enable users to programmatically query and search the CREEDS database.

Automatic extraction of gene expression signatures from GEO. To automatically extract gene expression signatures from GEO, we first applied the gradient boosting machines classifier (described above) to predict the categories of all GEO studies ($n=31,905$) performed in human, mouse or rat using microarrays. The classifier utilized the title, summary and keywords from each study. After this step, we selected the studies that were predicted to be gene, disease or drug perturbations with a probability threshold greater than $P>0.9$. We then applied the Naive Bayesian-based classifiers described above to predict the probability of whether samples associated with these studies have controls based on the sample titles. Next, we computed the pairwise Manhattan distance between the samples based on features extracted from sample descriptive terms, and then used the DBSCAN[77] algorithm with minimum samples set of 2 to perform clustering on the distance matrix between samples to identify clusters of semantically similar samples. We removed any clusters with large standard deviation ($P>0.2$) to reduce instances of mixture between control and perturbation samples. To determine whether a cluster of samples is a control group or a perturbation group, we chose

the average probability $P>0.7$ and $P<0.3$ from the Naive Bayesian-based classifier as control group and treatment group, respectively. Next, we enumerated every pair of valid control groups and perturbation groups within each study as metadata for valid predicted gene expression signatures.

To properly label the terms associated with each predicted signature, we used the API of BeCAS[78] to tag biological entities from the text associated with each study, as well as the text associated with the samples, including: genes, cell or tissue, disease, and drug or other small molecule chemical; and then recorded these term counts for a final decision of which terms we should use to label each signature. To process the gene expression data of the predicted gene expression signatures, we first used SVA[28] to correct the batch effect as described above, and then applied the CD algorithm[21] to compute differential expression.

Data availability. All extracted and processed signatures with their accession numbers and other metadata are freely available for download from the CREEDS portal at: http://amp.pharm.mssm.edu/creeds. The CREEDS portal also provides the data through API. Users can search the data by submitting their own signatures for analysis. The site also provides two modes of visualization of all signatures. Accession codes for top hits for drug signatures extracted from GEO queried against drug perturbations can be found in Table 1.

References

1. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41,** D991–D995 (2013).
2. Rustici, G. *et al.* ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41,** D987–D990 (2013).
3. Chang, J. *et al.* SIGNATURE: A workbench for gene expression signature analysis. *BMC Bioinformatics* **12,** 443 (2011).
4. Williams, G. A searchable cross-platform gene expression database reveals connections between drug treatments and disease. *BMC Genom.* **13,** 12 (2012).
5. Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H. & Horton, P. CellMontage: similar expression profile search server. *Bioinformatics* **23,** 3103–3104 (2007).
6. Engreitz, J. M. *et al.* ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics* **27,** 3317–3318 (2011).
7. Zinman, G. E., Naiman, S., Kanfi, Y., Cohen, H. & Bar-Joseph, Z. ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods* **10,** 925–926 (2013).
8. Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods* **12,** 211–214 (2015).
9. Dudley, J. T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **3,** 96ra76–96ra76 (2011).
10. Hu, G. & Agarwal, P. Human disease-drug network based on genomic expression profiles. *PLoS ONE* **4,** e6536 (2009).
11. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci.* **107,** 14621–14626 (2010).
12. Feng, C. *et al.* GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genom.* **10,** 411 (2009).
13. Good, B. M. & Su, A. I. Crowdsourcing for bioinformatics. *Bioinformatics* **29,** 1925–1933 (2013).
14. Khare, R., Good, B. M., Leaman, R., Su, A. I. & Lu, Z. Crowdsourcing in biomedicine: challenges and opportunities. *Brief. Bioinf.* **17,** 23–32 (2015).
15. Candido dos Reis, F. J. *et al.* Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMed.* **2,** 681–689 (2015).
16. Benjamin, M. G., Max, N., Chunlei, W. U. & Andrew, I. S. in *Biocomputing 2015* 282–293 (World Scientific, 2014).
17. Burger, J. D. *et al.* Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. *Database* **2014,** bau094 (2014).
18. Gottlieb, A., Hoehndorf, R., Dumontier, M. & Altman, R. B. Ranking adverse drug reactions with crowdsourcing. *J. Med. Internet Res.* **17,** e80 (2015).
19. Khare, R. *et al.* Scaling drug indication curation through crowdsourcing. *Database* **2015,** bav016 (2015).
20. Vergoulis, T. *et al.* mirPub: a database for searching microRNA publications. *Bioinformatics* **31,** 1502–1504 (2015).
21. Clark, N. *et al.* The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinf.* **15,** 79 (2014).
22. Storey, J. D. & Tibshirani, R. in *The analysis of gene expression data*, 272–290 (Springer, 2003).
23. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
24. Anders, S. Analysing RNA-Seq data with the DESeq package. *Mol. Biol.* **43,** 1–17 (2010).
25. Gundersen, G. W. *et al.* GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions. *Bioinformatics* **31,** 3060–3062 (2015).
26. Li, J., Bushel, P. R., Chu, T.-M. & Wolfinger, R. D. in *Batch Effects and Noise in Microarray Experiments*, 141–154 (John Wiley & Sons, Ltd, 2009).
27. Boedigheimer, M. J. *et al.* Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genom.* **9,** 1–16 (2008).
28. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** e161 (2007).
29. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).
30. He, X. C. *et al.* PTEN-deficient intestinal stem cells initiate intestinal polyposis. *Nat. Genet.* **39,** 189–198 (2007).
31. Sagiv, E. *et al.* Targeting CD24 for treatment of colorectal and pancreatic cancer by monoclonal antibodies or small interfering RNA. *Cancer Res.* **68,** 2803–2812 (2008).
32. Soucek, L. *et al.* Mast cells are required for angiogenesis and macroscopic expansion of Myc-induced pancreatic islet tumors. *Nat. Med.* **13,** 1211–1218 (2007).
33. Nilsson, E. C. *et al.* Opposite transcriptional regulation in skeletal muscle of AMP-activated protein kinase γ3 R225Q transgenic versus knock-out mice. *J. Biol. Chem.* **281,** 7244–7252 (2006).
34. Hwang, S. J. *et al.* Hypercholesterolaemia in patients with hepatocellular carcinoma. *J. Gastroenterol. Hepatol.* **7,** 491–496 (1992).
35. Sohda, T. *et al.* Reduced expression of low-density lipoprotein receptor in hepatocellular carcinoma with paraneoplastic hypercholesterolemia. *J. Gastroenterol. Hepatol.* **23,** e153–e156 (2008).
36. Savage, D. G. & Antman, K. H. Imatinib mesylate—a new oral targeted therapy. *N. Engl. J. Med.* **346,** 683–693 (2002).
37. Hodi, F. S. *et al.* Imatinib for melanomas harboring mutationally activated or amplified kit arising on mucosal, acral, and chronically sun-damaged skin. *J. Clin. Oncol.* **31,** 3182–3190 (2013).
38. Martínez-Ramírez, A. *et al.* Analysis of myelodysplastic syndromes with complex karyotypes by high-resolution comparative genomic hybridization and subtelomeric CGH array. *Genes Chromosomes Cancer* **42,** 287–298 (2005).
39. Antunes, C. M. F. *et al.* Endometrial cancer and estrogen use. *N. Engl. J. Med.* **300,** 9–13 (1979).
40. Weiderpass, E. *et al.* Risk of endometrial cancer following estrogen replacement with and without progestins. *J. Natl Cancer Inst.* **91,** 1131–1137 (1999).
41. Grady, D., Gebretsadik, T., Kerlikowske, K., Ernster, V. & Petitti, D. Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstet. Gynecol.* **85,** 304–313 (1995).
42. Kahlert, S. *et al.* Estrogen receptor α rapidly activates the IGF-1 receptor pathway. *J. Biol. Chem.* **275,** 18447–18453 (2000).
43. Song, R. X. *et al.* The role of Shc and insulin-like growth factor 1 receptor in mediating the translocation of estrogen receptor α to the plasma membrane. *Proc. Natl Acad. Sci. USA* **101,** 2076–2081 (2004).
44. Sirianni, R. *et al.* Targeting estrogen receptor-α reduces adrenocortical cancer (ACC) cell growth *in Vitro* and *in Vivo*: potential therapeutic role of selective estrogen receptor modulators (SERMs) for ACC treatment. *J. Clin. Endocrinol. Metab.* **97,** E2238–E2250 (2012).
45. Pollak, M. Insulin and insulin-like growth factor signalling in neoplasia. *Nat. Rev. Cancer* **8,** 915–928 (2008).
46. Schmandt, R. E., Iglesias, D. A., Co, N. N. & Lu, K. H. Understanding obesity and endometrial cancer risk: opportunities for prevention. *Am. J. Obstet. Gynecol.* **205,** 518–525 (2011).
47. Michalik, L., Desvergne, B. & Wahli, W. Peroxisome-proliferator-activated receptors and cancers: complex stories. *Nat. Rev. Cancer* **4,** 61–70 (2004).
48. Tsuchida, A. *et al.* Peroxisome proliferator-activated receptor (PPAR)α activation increases adiponectin receptors and reduces obesity-related inflammation in adipose tissue: comparison of activation of PPARα, PPARγ, and their combination. *Diabetes* **54,** 3358–3370 (2005).
49. Mu, N., Zhu, Y., Wang, Y., Zhang, H. & Xue, F. Insulin resistance: a significant risk factor of endometrial cancer. *Gynecol. Oncol.* **125,** 751–757 (2012).
50. Tupler, R. & Gabellini, D. Molecular basis of facioscapulohumeral muscular dystrophy. *CMLS Cell Mol. Life Sci.* **61,** 557–566 (2004).
51. Tawil, R. & Van Der Maarel, S. M. Facioscapulohumeral muscular dystrophy. *Muscle Nerve* **34,** 1–15 (2006).
52. Lamb, J. *et al.* The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
53. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).
54. The Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
55. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).
56. Settles, B. Active learning literature survey. *University of Wisconsin, Madison* **52,** 11 (2010).

57. Yan, Y., Fung, G. M., Rosales, R. & Dy, J. G. in *Proceedings of the 28th international conference on machine learning* (ICML-11). *Active learning from crowds*. 1161–1168 (2011).
58. Mozafari, B., Sarkar, P., Franklin, M., Jordan, M. & Madden, S. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proc. VLDB Endow.* **8,** 125–136 (2014).
59. Gray, K. A. *et al.* Genenames. org: the HGNC resources in 2013. *Nucleic acids Res.* **41,** D1071–D1078 (2012).
60. Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43,** D545–D552 (2015).
61. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–D1097 (2014).
62. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).
63. Wang, Z., Clark, N. & Ma'ayan, A. Dynamics of the discovery process of protein-protein interactions from low content studies. *BMC Syst. Biol.* **9,** 26 (2015).
64. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: text mining and data integration of disease–gene associations. *Methods* **74,** 83–89 (2015).
65. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50,** 742–754 (2010).
66. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44,** 837–845 (1988).
67. Fellbaum, C. *WordNet* (Wiley Online Library, 1998).
68. Van Rijsbergen, C. J., Robertson, S. E. & Porter, M. F. *New models in probabilistic information retrieval*. (Computer Laboratory, University of Cambridge, 1980).
69. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to information retrieval* Vol. 1 (Cambridge university press Cambridge, 2008).
70. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 85 (2008).
71. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
72. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32 (2001).
73. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63,** 3–42 (2006).
74. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29,** 1189–1232 (2001).
75. Breiman, L. Bagging predictors. *Mach. Learn.* **24,** 123–140 (1996).
76. Zadrozny, B. & Elkan, C. *in ICML,* vol. 1, 609–616 (Citeseer, 2001).
77. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, **96,** 226–231 (1996).
78. Nunes, T., Campos, D., Matos, S. & Oliveira, J. L. BeCAS: biomedical concept recognition services and visualization. *Bioinformatics* **29,** 1915–1916 (2013).

Acknowledgements

This work is supported by NIH grants: R01GM098316, U54HL127624 and U54CA189201 to A.M.

Author contributions

Z.W. and A.M. developed the crowdsourcing portal. Z.W., G.W.G., N.F.F. and A.M. developed the CREEDS web portal. A.M., Z.W. and K.M.J. wrote the paper. A.M., Z.W., N.R.C., S.L.J., M.G.M., A.D.R., G.W.G., Q.D., Y.K. and A.S.F. contributed relevant materials to the Coursera course. M.R.J. and M.G.M. performed systems administration tasks to set up the web server environment. G.W.G developed the tool used to annotate and extract signatures. Z.W. and C.D.M. reviewed entries for quality. All other authors not mentioned above and C.D.M., K.M.J., A.D.R., A.S.F., Z.W. and A.M. contributed to the crowdsourcing signature extraction process by submitting signatures to the database.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Wang, Z. *et al.* Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* 7:12846 doi: 10.1038/ncomms12846 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016