NANOGP1 AS A MODEL TO STUDY THE CONSEQUENCES OF GENE DUPLICATIONS ON HUMAN PLURIPOTENCY AND DEVELOPMENT

Katsiaryna Maskalenka

Churchill College

University of Cambridge
Babraham Institute

This dissertation is submitted for the degree of Doctor of Philosophy

December 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit for the Biology Degree Committee.

NANOGP1 AS A MODEL TO STUDY THE CONSEQUENCES OF GENE DUPLICATIONS ON HUMAN PLURIPOTENCY AND DEVELOPMENT

Katsiaryna Maskalenka

Abstract

Gene duplication events play an important role in genome evolution; they can also create developmental strategies that differ between species. However, the functional contribution of duplicated genes in early human development and pluripotency is poorly understood. To address this knowledge gap, I investigated *NANOGP1*, which is a duplicated pseudogene of a key pluripotency factor called *NANOG*. *NANOGP1* was chosen as a model for studying gene duplication in human pluripotency for several reasons. Firstly, *NANOGP1* is an evolutionarily conserved duplicate in Hominidae that appears to have an intact coding sequence. The pseudogene is currently annotated as non-protein-coding, although no functional assays have been performed to test this. Secondly, upon investigating the expression of pseudogenes in human naïve pluripotent stem cells (PSCs), I found that *NANOGP1* is among the top 1% of the highest expressed pseudogenes. Because high expression levels of *NANOG* are crucial for maintaining human pluripotency, I hypothesised that a duplicated copy of this important developmental regulator could have similar properties and might contribute to the regulation of human pluripotency.

Gene expression profiling revealed that *NANOG* and *NANOGP1* have overlapping but distinct expression patterns, both in human embryos and in PSC states. *NANOGP1* is highly expressed in naïve pluripotent cells but is significantly downregulated in primed pluripotent cells, while *NANOG* expression levels do not differ to the same extent between the two pluripotent states. RNA splicing analysis predicted that *NANOGP1* encodes a protein with an intact homeodomain and transactivation domain, but lacking part of the N-terminus. The divergent N-terminus is the main structural difference between NANOG and NANOGP1 and was therefore used in this study to distinguish between the two genes. Using CRISPR/Cas12a-mediated gene editing in naïve PSCs, I introduced an epitope tag at the start of the predicted protein sequence, and this enabled me to demonstrate for the first time that endogenous *NANOGP1* encodes an expressed protein.

The ability to be translated into the stable protein raised the possibility that *NANOGP1* could have a functional role. To test this, I performed a series of assays and established that at least two key functional properties are conserved between *NANOG* and *NANOGP1*: gene autorepression, and the ability to promote primed-to-naïve PSC reprogramming. Alongside this, however, downregulating *NANOGP1* expression using inducible CRISPRi in naïve PSCs did not lead to a differentiation phenotype,

which is in contrast to *NANOG* loss of function. Finally, using ChIP-seq, I showed that NANOGP1 shared a subset of chromatin binding sites with NANOG, and also, surprisingly, had a small number of unique, NANOG-independent sites particularly at the promoters of neural-associated genes.

Overall, I conclude that *NANOGP1*, a previously overlooked duplicated copy of *NANOG*, is an expressed, protein-coding transcription factor in human naïve PSCs. Most of the CDS, and several of the functional properties, are conserved, implying that *NANOGP1* could be supporting or cooperating with its ancestral gene copy in stabilising pluripotency. At the same time, differences in the N-terminal of the CDS, binding occupancy, and distinct expression patterns, could potentially contribute to functional diversification. These differences could have significant evolutionary consequences for creating species-specific developmental strategies, such as novel cell type-specific activity, expanded protein interaction networks and interplay with signalling pathways. Collectively, these potential new properties might extend functional potential and, hence, could encourage diversification of developmental mechanisms. Taken together, my work has demonstrated that *NANOG/NANOGP1* duplication serves as a paradigm for exploring how pseudogenes could support their ancestral copies, as well as expand the evolutionary potential of conserved developmental programmes.

Table of Acknowledgements

Initial training in techniques and laboratory practice and subsequent mentoring:
Peter Rugg-Gunn
Adam Bendall
Aled Parry
Andrew Malcolm
Anne Segonds-Pichon
Asif Nakhuda
Rachael Walker
Charlene Fabian
Clara Novo
Claudia Semprich
Katarzyna Wojdyla
Maria Rostovskaya
Rebecca Roberts
Sarah Elderkin
Simon Andrews
Simon Walker
Stephen Bevan

Data obtained from a technical service provider:
Babraham Bioinformatics Facility
Babraham Flow Cytometry
Babraham Sequencing Facility

Data produced jointly:
Adam Bendall
Christel Krueger
Katie Mullholland

Data/materials provided by someone else:
Peter Rugg-Gunn
Amanda Collier
Aylwyn Scally
Felix Krueger
Gökberk Alagöz
Maria Rostovskaya

Table of Contents

List of Figures

List of Tables

Abbreviations

2i/5i Two/five inhibitors

ATAC Assay for transposase-accessible chromatin

cDNA Complementary DNA

CDS Coding sequence

ChIP Chromatin immunoprecipitation

ChIP-seq ChIP sequencing

Chiron CHIR99021

crRNA crispr RNA

Cs Carnegie stage

DNA Deoxyribonucleic acid

dNTP Deoxyribonucleotide triphosphate

Dpf Ddays post fertilisation

EpiSC Epiblast stem cell

ESC Embryonic stem cell

EST Expressed sequence tag

FACS Fluorescence-activated cell sorting

FBS Foetal bovine serum

FGF Fibroblast growth factor

HD Homeodomain

HDR Homology directed repair

hPSC Human pluripotent stem cell

ICM Inner cell mass

iPSC Induced pluripotent stem cell

IVF *In vitro* fertilisation

KD Knockdown

KO Knockout

lncRNA Long non-coding RNA

MAPQ Mapping Quality

mAU Milli Absorbance units

mESC Mouse embryonic stem cell

Mya million years ago

mRNA messenger RNA

NGS Next-generation sequencing

NHEJ Non-homologous end joining

NHP Non-human primate

NK2 *NANOG*, *KLF2*

OE Overexpression

OR Olfactory receptor

ORF Open reading frame

PAM Protospacer adjacent motif

PBS Phosphate-buffered saline

PCA Principal component analysis

PCR Polymerase chain reaction

PD03 PD0325901

PE Primitive endoderm

PGC Primordial germ cell

FPKM Fragments per kilobase of transcript per million

POI Protein of interest

PSC Pluripotent stem cell

QC Quality control

RNA Ribonucleic acid

RNA-seq RNA sequencing

RPKM Reads per kilobase of transcript per million mapped reads

RPM Reads per million

RT-qPCR Real-time quantitative PCR

scRNA-seq Single cell RNA sequencing

SD Segmental duplication

sgRNA Single guide RNA

ssODN Single-stranded oligodeoxynucleotide

TE Trophectoderm

tracrRNA - trans-activating crispr RNA

TSS Transcription start site

UTR Untranslated region

WR Tryptophan-rich region

1 Introduction

1.1 Human development from zygote to gastrula

1.1.1 History of studying human early embryo development

Human pre-implantation embryos were first characterised by Arthur Hertig and John Rock, who obtained the study material from volunteers that had undergone hysterectomy (Hertig et al., 1954; Hertig et al., 1956). The two scientists studied and described pre-implantation developmental states, beginning with the 2-cell stage and ending with early gastrulating embryos. After that, replacing hysterectomy as a main form of sample collection, *in vitro* fertilisation (IVF) became the main source of material for researching human pre-implantation development. In the 1970s, embryo research was mostly focused on culturing and observing embryos *in vitro* to better understand their developmental progression, which in return was also instrumental in developing methods for the IVF technology (Edwards et al., 1970; Edwards et al., 1980; Lopata, 1980; Lopata et al., 1978; Steptoe et al., 1971; Steptoe et al., 1980). In the following years, when the morphological stages of the embryo development became more understood, scientists started to prioritise improving the IVF efficiency and enabling long-term storage of human embryos, therefore advancing such techniques as IVF human embryo transfer and human embryo cryopreservation (Gardner et al., 2000; Leeton et al., 1982; Saito et al., 2000; Testart et al., 1988; van Royen et al., 1999). One of the biggest researchers of early human development in 1980-2000 was Dr. Arunachalam Sathananthan, an electron microscopist, who captured and visualised ultrastructural morphology of the human oocyte fertilisation and subsequent embryo development. His work demonstrated the effects of culture and cryopreservation on the embryo functions, as well as described functional competence of parental gametes (Sathananthan, 1984; Sathananthan, 1990; Sathananthan, 1993; Sathananthan, 1994; Sathananthan, 1997; Sathananthan, 1998; Sathananthan and Trounson, 1985; Sathananthan and Trounson, 1989; Sathananthan et al., 1982; Sathananthan et al., 1986; Sathananthan et al., 1990; Sathananthan et al., 1999; Sathananthan. A et al., 1993). More recently, following the line of research established by Dr. Sathananthan, Wong and colleagues captured the first seven days of human embryo development using time-lapse imaging and *in vitro* embryo culture (Wong et al., 2010). This study demonstrated that it is already possible to predict whether the embryo would develop into a functioning blastocyst at the 4-cell stage, prior to embryonic genome activation, which connected successful embryo survival with the influence of maternal and/or paternal factors. Interestingly, human pre-implantation embryogenesis was also found to be notably inefficient both *in vitro* and *in vivo,* with more than 50% of the embryos failing to form a blastocyst (French et al., 2010; Gardner et al., 2000; Hertig et al., 1954).

Studying later stages of embryo development, involving cytotrophoblast proliferation, differentiation and embryo implantation, was more limited due to ethical and technical reasons. The focus was shifted towards investigating maternal contribution to placenta, such as an early study by Khong and Robertson who, like Hertig and Rock, also used postpartum hysterectomy material (Khong and Robertson, 1987). Therefore, studies focusing on mechanisms of human embryo invasion and implantation were mostly comparative and hypothetical, such as those described in Norwitz et al., 2001 and Moffett and Loke, 2006.

For the first-time, the transcriptome of the pre-implantation human embryo was described in 2004 (Dobson et al., 2004). More recently, in the past 10 years, several research groups performed complex single-cell RNA-sequencing (scRNA-seq) analysis of early human embryos in order to deepen our understanding of lineage marker expression, establish lineage segregation patterns and identify human-specific developmental milestones (Blakeley et al., 2015; Liu et al., 2019; Meistermann et al., 2021; Molè et al., 2021; Okamoto et al., 2011; Petropoulos et al., 2016; Vallot et al., 2017; Xiang et al., 2020; Yan et al., 2013). All of these studies were limited by the 14-day rule, first proposed in the UK in 1979 and then adopted by many other jurisdictions, which prohibited all experimental manipulations and embryo culture past this developmental timepoint (Cavaliere, 2017). However, in May 2021, the International Society for Stem Cell Research (ISSCR) announced new guidelines proposing that research on embryos after 14 days of development could be permitted if local policies and regulation permit, provided that each project proposal is assessed by a suitable science and ethical review. Following the new legislation, a landmark study to first examine human development at a later stage was published in November 2021 (Tyser et al., 2021). This publication contained a single-cell transcriptome study of a gastrulating embryo, the age of which was 16-19 days post fertilisation (dpf).

In summary, in the past 70 years, human embryo research has progressed considerably, starting from simple description of embryo morphology to complex single-cell molecular studies. In combination with recent modifications of ethical guidelines, such significant technological advancements have a high potential of transforming the human embryology field – and the way scientists address biological questions that are beyond our understanding.

1.1.2 Human development from zygote to gastrula: timeline, milestones and comparison to mouse and primate embryo development

Historically, human embryo research has been ethically limited and mostly comparative, using other model organisms and *in vitro* stem cell systems to infer knowledge about early human development. Two major non-human model groups of species were rodents and non-human primates (NHPs), and while mouse has been the predominant system for mammalian biology, primate

development has been studied less well. Among NHPs, the biomedical field has been mostly focused on the four following species: *Pan troglodytes* (chimpanzee), *Macaca fascicularis* (cynomolgus monkey/crab-eating macaque), *Callithrix jacchus* (common marmoset) and *Macaca mulatta* (rhesus macaque), with macaque being the most well researched (Johnsen et al., 2012).

This section compares early human development to that of mouse and NHPs, focusing on the need for further research of primate models and the limitations of studying mouse development for inferring about human development.

Human embryo development (gestation weeks 1–8, which precede the foetal period) is divided into 23 morphological stages called Carnegie stages (O'Rahilly and Müller, 2010), summarised in Table 1.1.

Table 1.1 Key stages of human embryonic development. The table contains nine selected developmental steps, while other stages are omitted for simplicity. These key stages are also compared between human, macaque and mouse development further in the text and **Figure 1.1**.

Carnegie stage	**Developmental milestone(s)**
Carnegie stage 1 (cs1)	Fertilisation and formation of zygote
Carnegie stage 2 (cs2)	Formation of blastomeres; morula stage
Carnegie stage 3 (cs3)	Segregation into hypoblast, inner cell mass (ICM) and trophectoderm; blastocyst stage
Carnegie stage 4 (cs4)	Formation of syncytiotrophoblast, cytotrophoblast; beginning of implantation
Carnegie stage 5 (cs5)	Formation of bilaminar germ disk; complete implantation and formation of secondary yolk sac
Carnegie stage 6 (cs6)	Formation of primitive streak and primordial germ cells (PGCs)
Carnegie stage 10 (cs10)	Embryo has 4-12 somites; neural fold formation; lateral folding of the embryo
Carnegie stage 14 (cs14)	Formation of ophthalmic vesicle and endolymphatic canal, cerebellar and visible ocular primordium, and sixth pharyngeal arch
Carnegie stage 23 (cs23)	Head is ~50% of the embryo's size; formation of the external genital primordium, chin and nasal pit

Key developmental steps, such as formation of a morula, blastocyst and gastrulation, are mostly similar between humans, mice and NHPs, however their development strategies exhibit several fundamental differences, described below.

Developmental timeline. Human embryo development is often described as protracted due to its significantly extended timeline compared to other key model species (Figure 1.1).

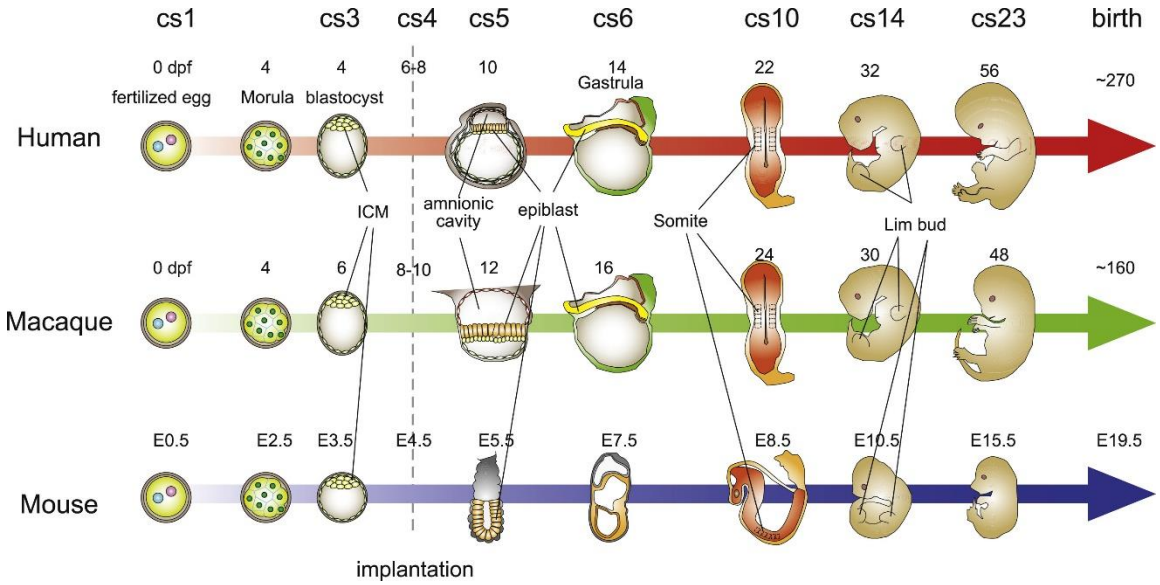


Figure 1.1 Diagram showing comparison of human, macaque and mouse embryonic development (reproduced from Nakamura et al., 2021). Stand-alone numbers indicate the number of days. Cs – Carnegie stage, dpf – days post fertilisation, E0.5 – E19.5 – mouse developmental stages, where values indicate number of days post conception.

Mouse embryos reach the morula stage ~2 dpf, whereas humans and macaques accomplish this by ~4 dpf. Another significant milestone, implantation, occurs in mouse on ~4 dpf, while for human this stage is observed at ~7-8 dfp and for macaque – at ~9-10 dpf (Finn and McLaren, 1967; Hertig et al., 1959; Heuser and Streeter, 1941a; Niakan et al., 2012; Norwitz et al., 2001). On the whole, the embryonic period of human development reaches the final, cs23 stage by 56 dpf, while similar progress takes 48 dpf and 16 dpf for macaque and mouse, respectively (Kaufman, 1992; Nakamura et al., 2021; O'Rahilly and Müller, 2010). Moreover, the human gestation period also takes longer than for the others, ~270 days, while that of macaque and mouse lasts ~160 and ~20 days, respectively. Collectively, this demonstrates that the timing for both embryonic and foetal periods varies significantly among major model species, including distinct developmental stage length among primates as well.

Protracted timeline and cellular heterogeneity. Longer embryo and foetal periods in humans could have potential consequences in organogenesis, as discussed in Nakamura et al., 2021, which used germ cell development as an example. Germ cell development in mouse and human is characterised by the presence of identical stages: PGC formation, migration and mitotic proliferation in genital ridges (Culty, 2009; de Felici, 2013; Felix, 1911; Fuss, 1911; Fuss, 1912; Ginsburg et al., 1990; Hilscher et al., 1974; Molyneaux et al., 2001; Politzer, 1930; Politzer, 1933; Saitou et al., 2002; Seki et

al., 2007; Speed, 1982; Tam and Snow, 1981; Witschi, 1948). Then, male PGCs enter mitotic arrest and differentiate into spermatogonia while female PGCs enter meiosis and become oocytes (Culty, 2009; Edson et al., 2009; Kurilo, 1981). In mouse development, these processes are strictly synchronised with the developmental timeline. In humans, however, they not only take longer, but are also highly asynchronous. For instance, human female PGCs arrive at the genital ridges for mitotic proliferation at ~35 dpf and then enter meiosis only two months later, around ~100 dpf. However, mitotically proliferating female PGCs can still be observed at much later stages, at ~180 dpf (Kurilo, 1981; Li et al., 2017) demonstrating that PGCs and oocytes co-exist in the developing ovaries for a rather prolonged period of time, potentially allowing multiple currently unknown cellular interactions to occur. This cellular heterogeneity demonstrates how different early human ovary development is from that of mouse, and simultaneously indicates that such diversification could also take place in other developing lineages and tissues.

Morphology and implantation. In addition to timing and cellular heterogeneity, rodent and primate early embryonic development exhibits crucial morphological differences. In primates, the early embryo forms the amniotic cavity prior to gastrulation, using epiblast cells only. After that, the primate trophoblast actively invades the maternal uterus, while the epiblast proliferates and develops into a flat embryonic disc (Heuser and Streeter, 1941b; Nakamura et al., 2016; O'Rahilly and Müller, 2010; Rossant and Tam, 2017). In mouse, the morphology of this process is drastically different. Mouse epiblast and extraembryonic cells contribute to the formation of the amniotic cavity, which takes place after gastrulation. Following that, epiblast extends to the distal side of the embryo, leading to formation of a cup shape embryo, as opposed to a flat disc in primates (Bedzhov and Zernicka-Goetz, 2014; Gardner, 1978; Kaufman, 1992). In conclusion, morphological changes of the embryo, upon a defining developmental step, implantation, are considerably different between rodents and primates.

Epigenetic regulation of early development. Mouse and human early development have distinct mechanisms of X-chromosome dosage compensation. Firstly, human embryo does not exhibit paternal X-chromosome silencing, observed in mouse (Okamoto et al., 2011). In contrast, expression from both X-chromosomes undergoes gradual dampening (Petropoulos et al., 2016). Additionally, in mouse, expression of long non-coding RNA (lncRNA) *Xist* initiates X-chromosome silencing (in *cis*), while human *XIST* is expressed from both X-chromosomes and does not trigger X-inactivation (Okamoto et al., 2011; Vallot et al., 2017). Together these facts highlight a major epigenetic difference between mouse and human early development.

In summary, differences in the developmental timeline, synchronicity, embryo morphology and early epigenetic mechanisms lead to a conclusion that developmental mechanisms studied in mouse are not always appropriate for inferring details of early human development, while NHP

models are more likely to provide accurate insight into processes that we cannot study in human directly.

1.1.3 Gene duplication in human and non-human primate genomes

Human evolution and development have been driven by the emergence of segmental duplications (SDs). SDs can lead to an increase in gene copy number, which is considered to possess a higher potential for species evolution than other mutations, such as single-nucleotide polymorphisms and segmental indels (Marques-Bonet et al., 2009b). In humans, over 5% of the genome consists of SDs, with more than 90% identity shared between the ancestral and the duplicated copies (Bailey, 2002; Marques-Bonet et al., 2009b). The percentage of duplicated regions in humans is remarkably high compared to Old World monkeys, such as macaques, where only 1.5% of the genome consists of SDs (Marques-Bonet et al., 2009b). Indeed, a sudden burst of gene duplication events occurred following the divergence of apes from Old World monkeys, and these SDs account for ~80% of modern human-specific duplications (Marques-Bonet et al., 2009a). Importantly, higher abundance of gene duplications is not limited to humans but can also be observed in other Great Apes, which also involves such species as gorilla, orangutan, chimpanzee and bonobo (Hahn et al., 2007). Molecular mechanisms for the increased overall amount of gene duplication are not fully understood and require further investigation.

1.1.4 Human embryo pluripotency: key regulators and comparison to other species

In development, cellular fate transitions are regulated by a finely tuned transcription factor network. Expression of major pluripotency and lineage-specific transcription factors is mostly conserved among human, NHP and rodent embryos (Blakeley et al., 2015; Boroviak et al., 2015; Boroviak et al., 2018; Meistermann et al., 2021; Molè et al., 2021; Nakamura et al., 2016; Petropoulos et al., 2016; Stirparo et al., 2018; Xiang et al., 2020; Yan et al., 2013). Moreover, orthologues of core pluripotency-associated factors, such as *Oct4/Pou5f1* and *Nanog,* are present in embryo development in other vertebrates as well, such as amphibia, birds and fish, with some variations in their function and expression patterns (Dixon et al., 2010; Lavial et al., 2007; Tapia et al., 2012; Theunissen et al., 2011b).

In mouse, the early pluripotent compartment exhibits expression of such core pluripotency genes as *Oct4*, *Nanog* and *Sox2* (Avilion et al., 2003; Dietrich and Hiiragi, 2007; Palmieri et al., 1994; Rosner et al., 1990). Recent studies of non-murine organisms, however, revealed that lineage marker

localisation and molecular regulation in other mammals often go against the commonly accepted mouse-focused paradigm, as described in the following paragraphs.

In mouse blastocysts, OCT4 and SOX2 proteins are initially expressed in all cells of the ICM, and only later become restricted to the epiblast and excluded form primitive endoderm (Avilion et al., 2003; Grabarek et al., 2012). Division of uniform mouse ICM into epiblast and primitive endoderm is marked by ICM-specific NANOG and GATA6, which are co-expressed at first and then resolve into a 'salt and pepper' expression pattern – NANOG in the epiblast and GATA6 in the primitive endoderm (Chazaud et al., 2006; Plusa et al., 2008; Rossant et al., 2003). As a result of these studies, *Oct4*, *Nanog* and *Sox2* became known as ICM and epiblast-specific markers, while *Gata6* along with *Gata4* and *Sox17* among others (Artus et al., 2011; Morris et al., 2010) represented primitive endoderm. Interestingly, in humans, NANOG and SOX2 proteins are expressed in the ICM but can also be detected in some trophectoderm cells (Cauffman et al., 2009). Human OCT4 expression pattern also differs from that of mouse and does not get restricted to the epiblast but expands to all cells of the embryo, including the extraembryonic trophectoderm lineage as well (Cauffman et al., 2006; Niakan and Eggan, 2013). Similar to OCT4, human GATA6 has an extended expression pattern and can be detected not only in primitive endoderm, but also in some trophectoderm cells (Deglincerti et al., 2016a; Roode et al., 2012). Such extended gene expression patterns were observed in other non-rodent model organisms as well, such as GATA6 in bovine and NHP embryos (Boroviak et al., 2015; Kuijk et al., 2012; Nakamura et al., 2016) and OCT4 in the rabbit and NHP (Cauffman et al., 2009; Harvey et al., 2009).

In certain cases, developmental markers are expressed both in mouse and human embryo, but in different compartments. For instance, human *ESRRB* expression is restricted to primitive endoderm and trophectoderm, instead of epiblast where its expression is observed in mouse (Blakeley et al., 2015).

Some mouse transcription factors are not expressed in early human embryo development at all. For example, one of the main pluripotency-associated transcription factors *Klf2*, is expressed in the mouse pre-implantation epiblast, while in human and NHP early embryos its expression is absent. Human and NHP embryos, however, exhibit expression of *KLF17* (Blakeley et al., 2015; Nakamura et al., 2016), structurally highly similar to both mouse and human *KLF2* (Yamane et al., 2018), therefore potentially utilising similar mechanisms as *Klf2* in mouse.

Some early developmental markers are present both in mouse and human, but with different onset of expression. For instance, in mouse, *Elf5* and *Cdx2* are involved in regulation of early trophectoderm cells (Donnison et al., 2005; Strumpf et al., 2005). In humans, they are also expressed in this extraembryonic lineage, but at later stages: *ELF5* - in the cytotrophoblast cells (Hemberger et al., 2010) and *CDX2* - after the blastocyst cavitation stage (Chen et al., 2009; Niakan and Eggan, 2013).

Finally, in addition to divergent expression patterns of key transcription factors in non-rodent species, their overall blastocyst development seemingly exhibits more plasticity than mouse. In cynomolgus monkey and human embryos, trophectoderm, epiblast and primitive endoderm were shown to develop simultaneously during the blastocyst stage (Meistermann et al., 2021; Nakamura et al., 2016; Petropoulos et al., 2016; Stirparo et al., 2018), whereas in mouse these compartments form in two sequential steps (reviewed in Niakan et al., 2012a). Moreover, if human blastocysts are disaggregated, both inner (ICM) and outer (trophectoderm) cells can re-form blastocysts that contain new correctly-positioned ICM and trophectoderm compartments (de Paepe et al., 2013), proving the plastic state of the blastocysts. Another more recent study also demonstrated that human naïve hPSCs can generate trophoblast cells directly (Dong et al., 2020).

In summary, despite mouse being commonly considered as the primary model organism for mammalian development, its developmental programmes are not universal, and other non-murine vertebrates share more between each other than with rodents.

So far, mouse has been the scientific model of choice due to less strict ethical regulations, greater accessibility, shorter gestation period than other higher animals and, certainly, the power of mouse genetics. Complementing these approaches, laboratories have also started designing non-murine embryo research strategies that could be just as accessible and would help advance our understanding of human early development. For instance, recent improvement of the human embryo culture allows investigating implantation-like stages *ex vivo* (Deglincerti et al., 2016a; Shahbazi et al., 2016), while other laboratories step away from working with embryos completely and focus on developing artificial systems replicating early developmental programmes, such as human gastruloids, blastoids, micropatterns and even so-called 'artificial embryos' (Deglincerti et al., 2016b; Fan et al., 2021; Kagawa et al., 2021; Lee et al., 2009; Liu et al., 2021; Manfrin et al., 2019; Minn et al., 2020; Moris et al., 2020; Paik et al., 2012; Sozen et al., 2018; Sozen et al., 2021; Tewary et al., 2017; Tewary et al., 2019; Warmflash et al., 2014; Yanagida et al., 2021; Yu et al., 2021). All of these different approaches help further investigation of human development from different angles and are already relatively widespread. In the near future, knowledge obtained in human and NHP research could match the amount of data gathered using the mouse model.

1.2 Human pluripotency and stem cells

1.2.1 Discovery and derivation of conventional human pluripotent stem cells

Pluripotent stem cells (PSCs) represent an *in vitro* model for studying early embryo development. The first mammalian embryonic stem cell (ESC) cultures were derived from a mouse

pre-implantation embryo in 1981 (Evans and Kaufman, 1981; Martin, 1981). Derivation of PSCs from human blastocysts (human PSCs, or hPSCs) occurred much later, in 1998, and was performed by James Thomson and colleagues (Thomson, 1998). In their experiment, embryos were cultured *in vitro* and, upon reaching the blastocyst stage, were used for deriving five hPSCs lines, namely H1, H7, H9, H13, and H14. Prior to this breakthrough, James Thomson's research group also established the first NHP PSCs for rhesus macaque (Thomson et al., 1995) and marmoset (Thomson et al., 1996).

Methods of demonstrating pluripotent properties of conventional hPSCs and, thus, validating their phenotype, differed from that of mouse stem cell culture. For instance, classic murine *in vivo* assays, such as tetraploid aggregation and blastocyst chimera formation (Brinster, 1974; Nagy et al., 1990) are unfeasible in human due to ethical reasons. This was overcome in several human/non-human animal chimera experiments. Thomson *et al*. (Thomson, 1998) tested the ability of hPSCs to form teratomas in mice, while Kurosawa (Kurosawa, 2007) demonstrated the capability of hPSCs to differentiate into all three germ layers, via embryoid body *in vitro* assays. Later, Mascetti and Pedersen (Mascetti and Pedersen, 2016a) explored whether hPSCs are able to integrate into mouse embryo and contribute to its development.

Conventional hPSC derivation techniques have been refined continuously. For instance, the optimal source of hPSCs was found to be the late blastocyst, more specifically, late epiblast cells exhibiting high levels of OCT4 expression (Chen et al., 2009), while plating whole intact blastocysts resulted in expansion of trophectoderm, interfering with hPSC derivation (Niakan and Eggan, 2013).

Despite all the achievements in derivation optimisation studies, conventional H9 (female) and H1 (male) hPSC lines, created by Thomson and colleagues in 1998 (Thomson et al., 1998), have remained the most frequently used cell lines worldwide, each reported in more than 46% and 23% of studies published in 2008 - 2016, respectively, according to Guhr et al., 2018.

1.2.2 Primed and naïve pluripotency in mouse and human

1.2.2.1 Primed and naive pluripotency in mouse

Mouse and human PSCs (hPSCs) can be captured and maintained *in vitro* in primed and naïve states, which represent *in vitro* counterparts of developmentally distinct pluripotent cells existing *in vivo*. In the early studies, mouse embryonic stem cells (mESCs) were derived from the pre-implantation epiblast in serum-based medium supplemented with Leukaemia Inhibitory Factor (LIF), defined here as serum/LIF. Another sub-type of pluripotent cells, currently referred to as mouse 'naive' state, was discovered in 2008 (Nichols et al., 2009; Ying et al., 2008). Similar to the serum/LIF cultures, naive ESCs can also be derived from the mouse pre-implantation epiblast (more specifically,

E3.5), although using a chemically-defined culture medium, 2iLIF (2i = 2 inhibitors). In 2iLIF, serum is replaced with inhibitors of Mitogen-Activated Protein Kinase Kinase/Extracellular Signal-Regulated Kinase (MEK/ERK) and Glycogen Synthase Kinase-3 (GSK3) pathways, using the small molecules PD0325901 (or, PD03) and CHIR99021 (or, Chiron), respectively. Serum/LIF ESCs can be adjusted to grow in the defined 2iLIF medium over the course of one or two passages, which results in their acquisition of the naive state (Tosolini and Jouneau, 2015) and induces such changes as core pluripotency protein binding reorganisation and global epigenetic reprogramming (Galonska et al., 2015). Mouse ESCs express mouse pluripotency markers, self-renew, contribute to blastocyst chimeras and have the potential to differentiate into all three embryonic lineages (Boroviak et al., 2015; Du et al., 2019; Ghimire et al., 2018; Nagy et al., 1993; Stewart, 1993).

Since the chemical composition of serum is animal-dependent, serum/LIF medium conditions exhibit batch-to-batch variation. Because of this, serum/LIF mESCs are heterogeneous and likely represent a more prolonged continuum of pluripotency, while naïve 2iLIF mESCs represent a more homogeneous cell culture (Marks et al., 2012; Wray et al., 2010), similar to the E4.5 epiblast (Boroviak et al., 2015).

In 2007, one more stem cell type/state was discovered, mouse epiblast stem cells (EpiSCs), this time derived from the post-implantation epiblast (Brons et al., 2007; Tesar et al., 2007). EpiSCs represent the 'primed' pluripotent state, transcriptionally resembling anterior primitive streak, which makes them more developmentally progressed than the naïve and serum/LIF cultures (Kojima et al., 2014; Tsakiridis et al., 2014). Despite their origin and distinct morphology, primed EpiSCs are also capable of self-renewal and differentiating towards the three germ lineages (Brons et al., 2007; Tesar et al., 2007). Notably, genetically unmodified EpiSCs are only capable to form chimeras when introduced into the post-implantation embryo (Brons et al., 2007; Huang et al., 2012; Kojima et al., 2014; Mascetti and Pedersen, 2016b; Tesar et al., 2007). As a reflection of the developmental dissimilarity, PSC states also differ morphologically when observed under a microscope: homogeneous naïve culture contains dome-shaped colonies, more heterogeneous serum/LIF colonies are flat, and primed EpiSC colonies are large, flat and have cobblestone-like structure (Figure 1.2).

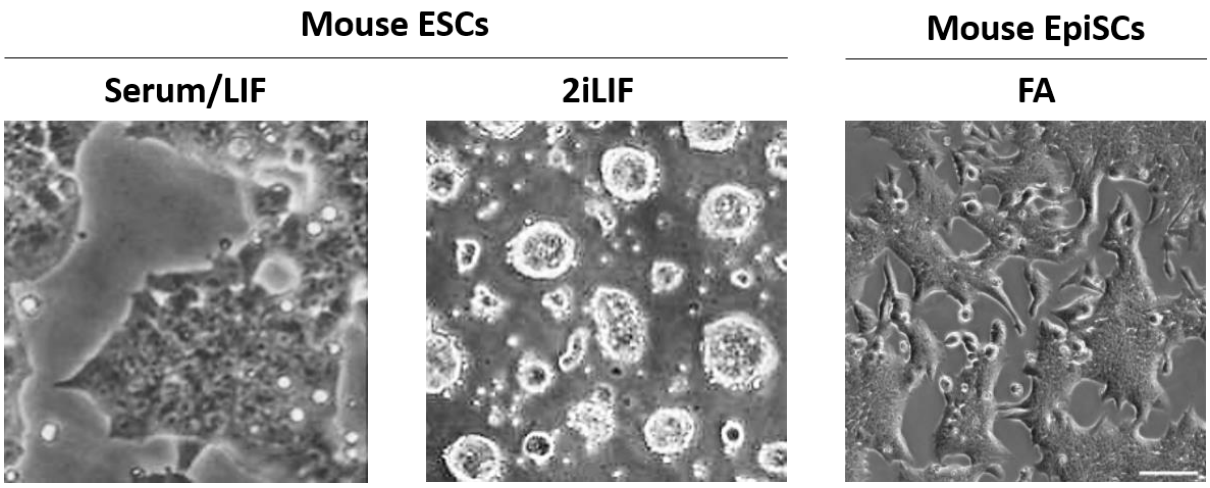


Figure 1.2 Bright field images showing morphology of mouse ESC and EpiSC cultures. Mouse ESCs – adapted from Tosolini and Jouneau, 2015 (no image scale given). Mouse EpiSCs – adapted from Stuart, 2019. Scale, 100 µm

1.2.2.2 Human primed pluripotency: differences and similarities with the primed pluripotency in mouse

Human PSCs, obtained using a conventional method in 1998, represent the primed state, despite being derived from the pre-implantation epiblast: they have similar morphology to mouse EpiSCs and require a similar culture medium to be maintained in the undifferentiated state (Beattie et al., 2005; James et al., 2005; Vallier et al., 2005; Wang et al., 2005; Xu et al., 2005). Mouse serum/LIF culture medium components are not suitable for maintaining conventional hPSCs in the undifferentiated state (Dahéron et al., 2004; Gerami-Naini et al., 2004; Humphrey, 2004; Xu et al., 2002; Xu et al., 2005). These studies demonstrated that with help of BMP proteins, provided by serum, conventionally derived hPSCs achieve extraembryonic differentiation, while LIF signalling does not have any effect on the self-renewal of hPSCs. Notably, in the latter case, hPSCs did respond to the LIF binding, as it induced downstream phosphorylation as well as nuclear translocation, although without any self-renewal-related effect (Dahéron et al., 2004; Humphrey, 2004). Defined 2iLIF medium also failed to maintain conventionally-derived hPSCs in their pluripotent state (Gafni et al., 2013; Hanna et al., 2010; Theunissen et al., 2014; Ware et al., 2014). Additional evidence for conventional hPSCs and mouse EpiSCs similarity was shown in the interspecies chimera experiment: hPSCs could only contribute to mouse post-implantation epiblast and did not display this ability when they were introduced into pre-implantation embryos (James et al., 2006; Masaki et al., 2015; Mascetti and Pedersen, 2016a). Finally, conventional hPSCs were classified as having a highly similar profile to the late post-implantation epiblast (Nakamura et al., 2016). In addition to that, one of the latest reports regarding the *in vivo* comparison of the conventional hPSC culture was provided by Tyser and

colleagues, who demonstrated that primed hPSCs are transcriptionally similar to the gastrula stage post-implantation epiblast in humans (Tyser et al., 2021).

Curiously, despite their similarities, human and mouse primed pluripotent cell cultures are not identical, as reviewed in Weinberger et al., 2016. Indeed, in addition to the common 'primed' properties, conventional hPSCs exhibit several 'naïve-like' features, positioning them 'in between' the naïve and primed murine states. For example, conventional hPSCs upregulate expression of cell adhesion protein E-cadherin, similar to mouse ESCs but not N-cadherin, expressed at a high level in mouse EpiSCs (Gafni et al., 2013). Moreover, conventional hPSCs express *REX1*, an early epiblast marker, and not *FGF5*, a marker of the late epiblast, in contrast to the mouse EpiSC, positive for *Fgf5* and not for *Rex1* (Chia et al., 2010; Gafni et al., 2013). Also, hPSC deoxyribonucleic acid (DNA) methylation pattern resembles that of mouse ESC grown in serum/LIF medium conditions and not that of EpiSCs (Hackett et al., 2013; Shipony et al., 2014). Another important evidence pointing at the intermediate positioning of primed hPSCs in relation to murine ESCs and EpiSCs is localisation of TFE3, whose orthologue mouse TFE3 is regulating exit from the pluripotent state in mouse ESCs. In naïve mouse ESCs, TFE3 protein exhibits strictly nuclear localisation while in EpiSCs it is present only in cytoplasm (Betschinger et al., 2013). Primed hPSCs also express TFE3, but its localisation cannot be linked to a single cellular compartment as it is detected both in the nucleus and cytoplasm (Gafni et al., 2013). Collectively, these findings support the hypothesis that primed hPSCs are not completely identical to the primed state in mouse, but instead likely to be positioned 'in between' the naïve and primed developmental states.

The dissimilarity between mouse ESCs and conventional primed hPSCs has been puzzling for many researchers, since both cell cultures were originally derived from the same developmental stage and under similar medium conditions. Therefore, it was not originally confirmed whether the naïve hPSC state, similar to the naïve 2iLIF mESC, existed at all (Evans and Kaufman, 1981; Martin, 1981; Thomson, 1998). To a certain extent, this could be explained by the fundamental difference between mouse and human early embryo development, which involves different developmental timelines, regulation of pluripotency and morphology, allowing to suggest that hPSC cultures required a different approach as well. This motivated the scientific community to develop novel, more elaborate culture medium recipes, that would help capture and investigate hPSC states.

1.2.2.3 History of deriving naïve human pluripotent stem cells

Initial attempts to derive human naïve cells *in vitro* were described in various studies (Hanna et al., 2010, Chan et al., 2013, Gafni et al., 2013, Valamehr et al., 2014 and Ware et al., 2014). However, none of the cell lines produced in the latter four studies were truly naive (Theunissen et al., 2014) or,

as in the study by Hanna and colleagues in 2010, transgene-independent. After a long search for experimental conditions that would enable deriving bona fide naïve hPSCs, a breakthrough was made in 2014, when this was achieved in 5i/L/A/F (5i = 5 inhibitors) reprogramming cell culture medium by Theunissen and colleagues (Theunissen et al., 2014), as well as by Takashima and colleagues who induced naïve pluripotency by overexpressing *NANOG* and *KLF2* transgenes in 2iLIF conditions, further replacing them with a PKC inhibitor, Go6983 (Takashima et al., 2014). Later, human naïve cells were also obtained using an updated version of 5i/L/A/F, 5i/L/A, which did not require FGF2 (Theunissen et al., 2016), or from pre-implantation ICM in t2iLGo medium conditions (Guo et al., 2016), as well as via other reprogramming methods (see Section 1.2.3.2 for more naïve hPSC derivation examples).

Unlike conventional primed cells, naïve hPSCs were able to contribute to mouse pre-implantation epiblast (Takashima et al., 2014), proving their similarity with the mouse ESCs, as well as their transcriptional overlap with the human morula and early epiblast (Huang et al., 2014; Theunissen et al., 2016). Due to these advancements, today we have evidence that the same developmental stage, pre- or early post-implantation one, can be used for deriving primed and naïve cells, both in mouse and human, and the outcome depends mostly on the medium conditions used in that experiment.

In conclusion, today we hypothesise that naïve and primed stem cells, mouse and human, do not reflect existence of two fixed developmental states but rather, represent a dynamic continuum of pluripotency. This way, naïve and primed cell cultures represent snapshots of two developmental stages that can be isolated and maintained *in vitro*. Additionally, another pluripotent cell type was recently derived in both human and mouse models, representing a formative state (Section 1.2.4) which reinforces the developmental continuum hypothesis. However, whether it is possible to maintain and propagate 'every' embryonic developmental state to recreate full progression of development *in vitro* is still unknown and awaits further investigation.

1.2.3 Molecular regulation of human pluripotent stem cells

Pluripotency depends on a balanced interplay of internal and external signalling molecules. Internal signalling, as implied by the name, is already present within a cell, while external signalling has to be provided by the culture medium, 'recapitulating' a signalling environment that the cell would have been exposed to *in vivo*. Signalling pathways involved in establishing a pluripotent state, as well

as their activators and inhibitors used in primed, naïve and reprogramming medium recipes are summarised in Figure 1.3.

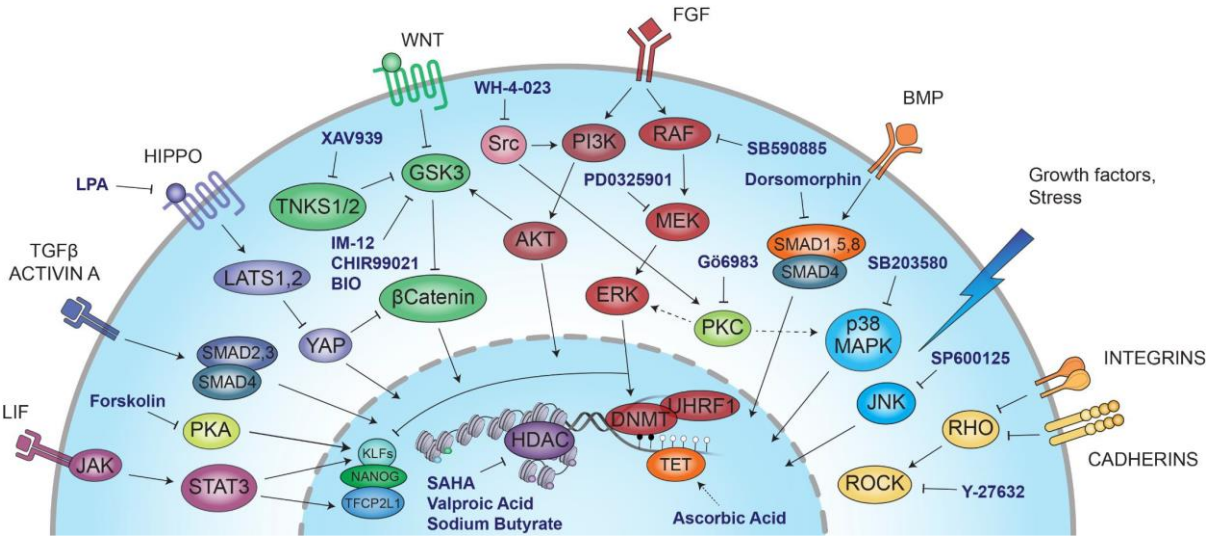


Figure 1.3 Diagram showing key pluripotency signalling pathways as well as their inhibitors/activators, used in culture medium for the PSC maintenance (reproduced from Collier and Rugg-Gunn, 2018)**.** See Section 1.2.3.1-1.2.3.2, describing the activity of selected components.

1.2.3.1 Primed pluripotent stem cell culture components and signalling

Primed hPSC culture medium conditions are very similar to that of primed mouse EpiSC culture (Beattie et al., 2005; James et al., 2005; Vallier et al., 2005; Wang et al., 2005; Xu et al., 2005). A key component, ensuring maintenance of the primed state in hPSCs, is the Fibroblast Growth Factor (FGF) pathway, and inhibiting its receptor causes complete loss of primed pluripotency (Vallier et al., 2005). Binding of FGF2 ligand to FGF receptor 1 activates its downstream targets Rapidly Accelerated Fibrosarcoma (RAF)/MEK/ERK and PI3K/AKT (Nakashima and Omasa, 2016), which, in turn, are responsible for maintaining the undifferentiated cell state and ensuring cell survival and proliferation (Li et al., 2007). Additionally, FGF2 causes feeder cells to secrete Inhibinβ-B, Gremlin 1 and FGF7, which also participate in maintaining the primed pluripotency (Diecke et al., 2008). Moreover, as per the model proposed by Bendall and colleagues, pluripotent hPSCs are capable of spontaneous transformation into fibroblast-like cells which, in their turn and in response to FGF protein present in the culture medium, produce IGF, TGF and other signalling molecules (Bendall et al., 2007). This way, hPSCs were hypothesised to enable continuous maintenance of the primed pluripotency via paracrine signalling. However, FGF signalling is not sufficient to maintain primed pluripotency on its own and requires Transforming Growth Factor (TGF)-β/ACTIVIN/NODAL pathway to be active as well, which is ensured by adding ACTIVIN A protein the culture medium (James et al., 2005; Vallier et al., 2005). Similar to the FGF pathway, inhibition of TGF signalling causes loss of primed pluripotency (Smith et al., 2008; Vallier et al., 2005)

In summary, to ensure maintenance of the pluripotent state, human primed hPSCs and mouse EpiSCs utilise similar external and internal signalling mechanisms, with key components being active FGF and TGF-β pathways.

1.2.3.2 Naïve pluripotent stem cell culture components and reprogramming methods

For 15 years since the discovery of primed hPSC culture, it had not been clear whether a naïve pluripotent state existed: mouse ESC embryo derivation methods resulted in establishing primed hPSC culture (see Section 1.2.1), and culturing primed hPSCs in mouse defined naive medium conditions (2iLIF), resulted in cell differentiation instead of pluripotency acquisition (Gafni et al., 2013; Hanna et al., 2010; Theunissen et al., 2014; Ware et al., 2014). However, in last eight years, several combinations of supplements and inhibitors were discovered that facilitated direct derivation of naïve-like hPSCs from the pre-implantation embryo (Gafni et al., 2013; Guo et al., 2016; Theunissen et al., 2014; Ware et al., 2014), as well as induction of the naïve fate in primed and somatic cells via reprogramming (Carter et al., 2016; Chan et al., 2013; Duggal et al., 2015; Gafni et al., 2013; Guo et al., 2016; Theunissen et al., 2014; Ware et al., 2014).

Naïve hPSCs, derived and maintained using these methods share distinct naïve-like features, such as dome/ball-shaped colonies, naïve transcriptional profiles, low levels of DNA methylation and increase in X-chromosome activation. However, some of these derivation and maintenance protocols required the presence of FGF2, which is a more common component of the primed cell culture or even resulted in upregulation of lineage-specific markers along with the pluripotency factors (Chan et al., 2013). Moreover, some naïve-like cultures (Gafni et al., 2013; Ware et al., 2014) produced naïve hPSCs that would not transcriptionally cluster with mouse naïve cultures. In other words, the naïve-like cell lines obtained by methods summarised in Table 1.2.
were not identical and, despite all having some naïve features, likely represented a spectrum, or variations, of naïve-like pluripotency.

Table 1.2 Summary of cell culture medium recipes used for deriving naïve hPSC (reproduced from (Collier, 2019)**.** Base Medium: (1) N2B27; (2) KnockOut-DMEM+N2B27; (3) TeSR™1 (4) DMEM+20%

KSR. ✓ - component added to medium. Conditions: ✓/X - component added to medium (optional). * - short-term induction/addition to the medium. ¥ - two optional components. '-'- not specified.

Chemical inhibitor, growth factor	Target effect	OKK+ 2iL (Hanna et al., 2010)	NHSM (Gafni et al., 2013)	3iL (Chan et al., 2013)	HDACi 2i+F (Ware et al., 2014)	5iLA(F) (Theunissen et al., 2014; Theunissen et al., 2016)	NK2 t2iL+PKCi (Takashima et al., 2014)	2iL+ STAT3 (Chen et al., 2015)	2iL+ F,FK,AA (Duggal et al., 2015)	2il+ FK,YAP (Qin et al., 2016)	HDACi t2iL+PKCi (Guo et al., 2017)
LIF	LIF signalling	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PD0325901	MEK inhibition	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CHIR99021	GSK3 inhibition	✓	✓		✓		✓	✓	✓	✓	✓
IM-12	GSK3 inhibition					✓					
BIO	GSK3 inhibition			✓							
Gö6983	PKC inhibition		✓				✓				✓
Y-27632	ROCK inhibition		✓			✓					✓
WH-4-023	SRC inhibition					✓					
SB590885	RAF inhibition					✓					
SP600125	JNK inhibition		✓								
SB203580	p38/MAPK inhibition		✓								
TGFβ	TGFβ signalling			✓							
Activin A	TGFβ signalling					✓					
FGF	FGF signalling			✓	✓	✓/X			✓		
SAHA	HDAC inhibition				✓*						
Sodium Butyrate	HDAC inhibition				✓*						✓*¥
Valproic Acid	HDAC inhibition										✓*¥
Dorsomorphin	BMP inhibition			✓							
Forskolin	PKA inhibition								✓	✓	
Ascorbic Acid	Demethylation								✓		
Lysophosphatidic acid	HIPPO inhibition									✓	
Base Medium		(1)	(2)	(3)	(4)	(1)	(1)	(1)(4)¥	(4)	(1)(3)	(1)
O_2 Level		20%	20% or 5%	-	5%	5%	-	5%	5%	-	5%
Transgenes		*OCT4 KLF2 KLF4*					*NANOG* KLF2**	*STAT3**			

All human naïve medium recipes, described in Table 1.2, contained three signalling components: an inhibitor of MEK/ERK signalling, an inhibitor of GSK3, and an activator of the Janus Kinase - Signal Transducer and Activator of Transcription (JAK-STAT) signalling, repeating the defined 2iLIF medium composition used in mouse naïve ESC culture. Additionally, each study proposed their own unique list of supplementary inhibitors, activators and/or transgenes, which were hypothesised to be beneficial in achieving and improving naïve properties of the newly discovered stem cell state. Supplementary components involved, in different combinations, inhibitors of SRC, BMP, RAF, c-Jun N-terminal kinase (JNK), HIPPO and p38/Mitogen-Activated Protein Kinase (MAPK), TGF-β and FGF activators, histone deacetylase (HDAC) and Protein Kinase A (PKA) inhibitors, as well as pluripotency factor transgenes (Table 1.2). The following paragraphs will detail the signalling components in the formulations PXGL and t2iLGo, which are considered to be among the most efficient and robust, while being FGF/ACTIVIN independent and allowing for feeder-free and transgene-free culture.

The t2iLGo recipe was first described in Takashima et al., 2014. It resembles mouse defined medium 2iLIF the most, as it contains MEK/ERK inhibitor PD03, GSK3 inhibitor Chiron (although at a lower, titrated concentration), and JAK-STAT activator LIF, all facilitating naïve self-renewal and pluripotency maintenance (Takashima et al., 2014). The only exception is a new additional component, Protein Kinase C inhibitor (PKCi) Go6983. Go6983 was shown to be able to block differentiation in mouse ESCs (Dutta et al., 2011) and, remarkably, was able to have the same effect in the naïve human

culture as well. Go6983 provides additional indirect inhibition for MEK/ERK and p38/MAPK pathways by negatively affecting the activity of the PKC protein. Without Go6983, human naïve colonies in 2iLIF degenerate, proving the necessity of this component in addition to MEK/ERK and GSK3 inhibition, and STAT activation (Takashima et al., 2014). Boroviak and Nichols also speculate that Go6983 might prevent cellular polarisation to keep naïve cells apolar, causing the round cells to form dome-shaped colonies (Boroviak and Nichols, 2017).

Similar to t2iLGo, PXGL (Bredenkamp et al., 2019b; Rostovskaya et al., 2019) also contains PD03, Go6983 and LIF. The key difference between these two medium recipes is that in PXGL, the GSK3 inhibitor is replaced by a small molecule XAV939, which, oddly, has an opposite function. XAV939 activates GSK3 via inhibiting its upstream repressor, tankyrase, followed by degradation of the Wingless/Int-1 (WNT) downstream target β-catenin (Huang et al., 2009). This is curious, because in human naïve pluripotency, WNT/β-catenin is positively associated with cell self-renewal, although its link with naïve cell state maintenance is not fully understood (Xu et al., 2016). It is currently hypothesised that in naïve hPSCs both GSK3 activation and inhibition can lead to a stable naïve phenotype for at least two reasons: GSK3 signalling is not essential in regulation of human naïve cells and its expression levels could vary, and/or could be substituted by other pathways. Experimental evidence for this was provided by two independent studies. Theunissen and colleagues showed that inhibition of GSK3 is optional in 5i/L/A human naïve reprogramming culture (Theunissen et al., 2016), while Austin Smith's research group succeeded to maintain chemically reprogrammed cells in a double-titrated version of t2ilgo, tt2ilgo, containing 0.3 μM instead of 1 μM Chiron (Guo et al., 2017). Overall, even though PXGL medium is an effective tool of current stem cell culture, the exact role of GSK3 and WNT/β-catenin in maintenance of human naïve pluripotency requires further research.

Finally, a very recent study has identified specific signalling requirements that are important for the induction and maintenance of naïve human pluripotency (Khan et al., 2021). Briefly, the study demonstrated that for the naïve hPSC maintenance, MEK inhibition can be replaced with inhibition of its upstream and downstream targets, including the downstream kinase ERK. Simultaneously, MEK/ERK inhibition promotes primed-to-naïve reprogramming when applied together with ACTIVIN A, as well as PKC, tankyrase and ROCK inhibitors. This recent finding uncovered that the induction and maintenance of the naïve pluripotency require different signalling components; therefore, in future research, we might need to re-consider whether using the same medium recipes for establishing of the naïve state and its maintenance is optimal.

In summary, human naive culture derivation and maintenance can be achieved via different signalling routes and medium composition. Despite utilising similar 'base' medium components, human naïve culture is not equivalent to mouse naïve pluripotency, as it requires additional signalling

molecules, and human GSK3 pathway seemingly plays a more redundant role. Several studies have shown that there are differences in the response to signalling inhibitors between human and mouse blastocysts (Blakeley et al., 2015; Niakan and Eggan, 2013; Niakan et al., 2012; Roode et al., 2012); this line of human embryo research will therefore be important for establishing conditions that would enable the best possible naïve hPSCs culture. Meanwhile, investigating ways to improve primed-to-naïve reprogramming and naïve cell derivation is continuing.

1.2.3.3 Distinguishing primed and naïve pluripotent stem cells

Characterisation of a cell phenotype involves identification of its protein markers that are expressed in a particular cell type in a specific developmental context, time or culture system. Currently existing hPSC investigation methods are rarely 100% efficient, such as reprogramming, which does not normally produce a completely homogenous reprogrammed culture. Therefore, researchers have to utilise stem cell distinguishing methods in the attempt to detect the correct combination of transcription factors and/or cell surface proteins in order to distinguish a certain cell type.

Transcriptional characterisation of human embryo development only recently became available (Blakeley et al., 2015; Guo et al., 2014; Okamoto et al., 2011; Petropoulos et al., 2016; Vallot et al., 2017; Yan et al., 2013). *In vivo* characteristics provided by these and similar publications have been used in transcriptional analysis and flow cytometry, in order to identify markers of specific cell types.

Core pluripotency factors *OCT4*, *SOX2* and *NANOG* are expressed in both naïve and primed hPSCs. While the levels of *OCT4* and *SOX2* do not differ significantly between the two states, *NANOG* transcription is slightly elevated in the naïve hPSCs (Messmer et al., 2019; Rostovskaya et al., 2019). In order to distinguish naïve and primed hPSCs from each other, as well as from somatic lineages and various assay-specific intermediate cell types, an array of markers was found in multiple independent studies. Commonly used naïve cell markers include: *DNMT3L, DPPA3, DPPA5, GATA6, IL6ST, KLF4, KLF5, KLF17, TBX3* and *TFC2PL1* (Blakeley et al., 2015; Collier et al., 2017; Dunn et al., 2014; Guo et al., 2017; Shahbazi et al., 2016; Theunissen et al., 2016; Weinberger et al., 2016; Yan et al., 2013), as well as *ALPP, ALPPL2, FAM151A, HORMAD1, HYAL4, KHDC1L, KHDC3L, LYZ, MEG8, OLAH, TRIM60* and *ZNF729*, found in a recent study conducted by Messmer and colleagues (Messmer et al., 2019). Primed hPSCs are characterised by expression of *OTX2, SFRP2, TFT* and *ZIC2* (Buecker et al., 2014; Guo et al., 2016) and additional *CYTL1, DUSP6, FAT3, HMX2, KLHL4, NEFM, PLA2G3, PTPRZ1, SOX11, STC1, THY1* and *ZDHHC22*, also described in (Messmer et al., 2019). Along with these intracellular markers, Collier and colleagues tested and validated cell surface proteins, specific to naïve and primed hPSCs (Collier

et al., 2017). In addition to the newly-identified markers, this validation study also included the analysis of surface markers discovered in previous publications (Chen et al., 2015b; Gafni et al., 2013; Pastor et al., 2016; Qin et al., 2016; Shakiba et al., 2015; Ware et al., 2014). As a result, Collier et al., 2017 described the optimised method allowing to distinguish primed and naive reprogramming populations from intermediate products during reprogramming experiments, as well as enabled fluorescence-activated cell sorting (FACS) sorting of desired populations. These cell surface markers include CD75, CD7, CD130, CD77, CD320 (naïve hPSC); CD24, CD57, CD90, HLA-A/B/C (primed hPSCs) and CD90.2, used for eliminating mouse feeder cells. Another cell surface marker frequently used to label the naïve population and distinguish it from other hPSCs is SUSD2 (Bredenkamp et al., 2019a)

One more alternative method for distinguishing the reprogramming population from by-products utilises reporter systems, labelling regulatory regions that are active in a specific pluripotent state. Examples of constructs developed were 1) *OCT4*-ΔPE-*GFP*, which contained a truncated version of a proximal primed-specific *OCT4* enhancer, labelled with *GFP*; 2) *OCT4*-ΔDE-*GFP*, which included a truncated version of a distal naive-specific *OCT4* enhancer, labelled with *GFP* (Gafni et al., 2013); and 3) EOS-(*OCT4*-DE)-*GFP*, which contained an early transposon, active in the undifferentiated pluripotent cells, as well as *Oct4* and *Sox2* ESC binding motifs (Hotta et al., 2009), where the *Oct4* component was modified to contain the naïve-specific enhancer only (Takashima et al., 2014).

In summary, distinguishing primed and naïve hPSCs from other populations requires, separately or in combination, such approaches as identification of transcriptional markers by Real-time quantitative PCR (RT-qPCR), identification of cell surface markers by flow cytometry, and the use of naïve/primed-specific reporter constructs.

1.2.4 Capacitation of human pluripotent stem cells for differentiation

Naïve hPSCs have an almost unlimited differentiation potential; by the definition of pluripotency, they have the potential to differentiate into all embryonic lineages *in vitro*. Remarkably, they were also shown to form expandable extraembryonic endoderm (Linneberg-Agerholm et al., 2019), trophectoderm cells (Cinkornpumin et al., 2020; Dong et al., 2020) and even form blastoids (Kagawa et al., 2021; Yanagida et al., 2021; Yu et al., 2021), including those that contain amnion-like cells, according to a study recently uploaded to BioRxiv (Zhao et al., 2021).

Notably, the naïve state is not particularly responsive to signalling that drives embryonic differentiation, and before following any particular differentiation route, naïve cells first have to achieve competence (Guo et al., 2017; Rostovskaya et al., 2019). Similar conclusion was made in attempts of differentiating mouse 2iLIF ESCs, demonstrating that they have to undergo formative capacitation (Hayashi et al., 2011; Kalkan et al., 2017; Mulas et al., 2017). Capacitation represents the

transition from pre-implantation through an early post-implantation stage, allowing cells to gain an ability to differentiate towards a somatic fate. In mouse ESCs, formative capacitation is achieved in 24-48 hours as described in (Hayashi et al., 2011; Mulas et al., 2017). In hPSC culture, formative capacitation can be induced, however, the initial design of a human capacitation protocol was not robust enough and was taking more than three weeks to achieve the desired competent cell state (Guo et al., 2017). In 2019, Rostovskaya and colleagues presented an improved capacitation protocol, enabling efficient transition of naïve cells towards the formative fate (Rostovskaya et al., 2019). They discovered that culturing naïve cells in N2B27 base medium supplemented with a tankyrase inhibitor XAV939 is sufficient to obtain a post-implantation-like expanding population in just a few days. Overall, a novel robust formative capacitation protocol opened up a plethora of possible applications for naïve hPSCs, both in fundamental research and medical application, however, this *in vitro* culture still requires further studies in order to understand its properties in more detail.

1.2.5 Uses, applications and limitations of human pluripotent stem cells

hPSCs have promising applications in regenerative medicine, cell therapy and drug discovery. The first hPSC clinical trials in cell therapy had been launched more than 10 years ago, as reviewed in (Liu et al., 2020). In such tests, conventional hPSCs have been used for differentiation towards oligodendrocyte progenitor cells (clinical trial number NCT01217008, NCT02302157), retinal pigment epithelial cells (NCT01345006, NCT01344993, NCT02286089), cardiomyocyte (NCT02057900) and pancreatic progenitor cells (NCT02239354). More recently, in 2019, Doss and Sachinidis proposed a detailed list of quality criteria that should be met by hPSCs in cell therapy: 1) sterility and absence of mycoplasma/endotoxins; 2) expression of pluripotency markers in the starting cell population; 3) differentiation marker profile must be unique for each particular therapeutic cell line; 4) normal karyotype; 5) absence of undifferentiated and malignant cells in the final product; 6) absence of any other contaminating cell types; 7) successful validation *in vivo* during pre-clinical trials; 8) transgene and vector-free; 9) ability to convey genotyping short tandem repeat assays in autologous induced pluripotent stem cell (iPSC) lines; 10) clinical-grade cell viability (Doss and Sachinidis, 2019).

While clinical trials attempt to create efficient cells lines, fundamental research is working towards advancing differentiation set-ups, such as creating 3D scaffolds, facilitating prolonged culture with optimal cell-cell interactions and signal transition (Kraehenbuehl et al., 2011; Lei and Schaffer, 2013; Sant et al., 2010). Faulkner-Jones and colleagues went even further and developed a bioprinting protocol, which enabled direct printing of hPSCs into a 3D alginate scaffold for further differentiation towards hepatocyte-like cells (Faulkner-Jones et al., 2015).

While conventional hPSCs are being widely used for developing novel cell models and therapies, direct application of naïve hPSCs in such differentiation studies is not currently possible due to the inability of 5i/L/A or t2iLGo cultures to efficiency differentiate into somatic tissues (Lee et al., 2017; Liu et al., 2017) and, therefore, related to this requirement of the formative capacitation for multilineage differentiation (Section 1.2.4). Additionally, it is possible that if naïve hPSCs were to be used as a starting point of the differentiation protocol in a clinical application, the final product could have issues with the regulation of imprinted genes due to the erasure of methylation marks and general 'lack of restriction' in the naïve hPSCs compared to the primed (reviewed in Collier and Rugg-Gunn, 2018). Nevertheless, naïve hPSCs are crucial for investigating fundamental human naïve-specific processes such as, for instance, X chromosome silencing (Sahakyan et al., 2017; Vallot et al., 2017) and transposable element regulation (Grow et al., 2015; Pontis et al., 2019; Theunissen et al., 2016). Naïve cells would also be advantageous in understanding such biological topics as: early human blastocyst plasticity (Section 1.1), development of extraembryonic lineages and amnion (Section 1.2.4) including, for instance, disease modelling in the cells of trophectoderm. Overall, naïve hPSCs have a very high potential of filling the gaps in understanding mechanisms of pre-implantation biology in human development.

Naive hPSCs are relatively new compared to the primed cultures and therefore less is known about their biology, which adds to other obstacles in their medical application. However, use of primed hPSCs also has its limitations and potential side-effects. A classic example is the risk of cancer mutations occurring in clinical applications such as, for example, mutations in p53 (Lin and Lin, 2017; Rivlin et al., 2011). Another example was demonstrated by two studies which showed that somatic lineages derived from hPSCs can become immunogenic (Swijnenburg et al., 2005; Swijnenburg et al., 2008). Another study showed that hPSCs obtained via reprogramming for neuronal differentiation were less efficient and more variable than embryo-derived hPSCs in the differentiation protocol (Hu et al., 2010). Two other studies demonstrated that different hPSC lines have distinct differentiation propensity, providing such examples as pancreatic differentiation, cardiomyocyte generation, hematopoietic differentiation and others (Bock et al., 2011; Osafune et al., 2008). Notably, these findings comprise only several examples of potential issues that could occur while using hPSCs in-clinic; they, as well as other numerous concerns, require thorough investigation and further trials.

In summary, hPSC culture is becoming a very promising tool in the medicinal field and in fundamental research. However, before its application becomes widespread, the scientific community has to eliminate potentially dangerous side-effects of using stem cells in humans. In my opinion, some ongoing clinical trials could be considered premature and potentially dangerous, since it is impossible

to claim that all potential side-effects and long-term outcomes of stem cell therapy techniques have already been investigated.

Overall, in order to successfully apply hPSC culture in-clinic and in fundamental research, we ought to deepen our general understanding of naïve and primed hPSC properties and general biology. One of the key questions to address here would be differences and similarities in expression of transcription factors between naïve and primed hPSCs. An example of an important transcription factor in human pluripotency and the description of its properties in the naïve and primed hPSCs is presented in the next sections.

1.3 Homeobox protein NANOG - pluripotency transcription factor

1.3.1 *NANOG* in naïve and primed pluripotency

Nanog is a homeobox gene, originally investigated and described by four independent groups as a component of mouse pluripotency (Chambers et al., 2003; Hart et al., 2004; Mitsui et al., 2003; Wang et al., 2003). Mouse *Nanog* is expressed in the developing pre- and post-implantation epiblast, developing germ cells, as well as in the pluripotent cell cultures (Chambers et al., 2003; Hart et al., 2004; Mitsui et al., 2003; Wang et al., 2003; Yamaguchi et al., 2005). Today, we have evidence that its human orthologue *NANOG* is involved in maintenance of pluripotency and cell self-renewal in human pluripotency as well, as described below.

In the human pluripotency network, high expression levels of *NANOG* are essential for maintaining the undifferentiated state and sustaining self-renewal of both primed and naïve states. High expression of *NANOG* changes significantly only between the primed and naïve states (Messmer et al., 2019; Rostovskaya et al., 2019) and could therefore be assumed to be equally important in both. Downregulation of *NANOG* expression in primed hPSCs destabilises the pluripotent state and facilitates acquisition of extraembryonic (Hyslop et al., 2005; Zaehres et al., 2005), neuroectodermal (Vallier et al., 2009) and definitive endoderm (Lie et al., 2012) fates. In contrast, its overexpression promotes increased cell proliferation; the cells remain pluripotent but appear to obtain a transcriptional signature of the epiblast and not ICM (Darr et al., 2006). Alternatively, cultures of *NANOG*-deficient naïve hPSCs upregulate several trophectoderm marker genes, while the overexpression of *NANOG* in naïve hPSCs supresses the induction of trophectoderm fate (Guo et al., 2021).

Nanog has an important reprogramming role in both human and mouse. Human and mouse *Nanog* orthologues are not included in the canonical OKSM (*Oct4*, *Klf4*, *Sox2*, *c-Myc*) reprogramming cocktail, which enables reprogramming fibroblasts into pluripotency (Takahashi and Yamanaka, 2006;

Takahashi et al., 2007). In mouse, somatic reprogramming appears to be *Nanog*-independent in the optimised full medium conditions, as shown in such studies as Carter et al., 2014 and Schwarz et al., 2014. However, exogenous *Nanog* was demonstrated to facilitate somatic stem cell reprogramming under minimal culture conditions (Theunissen et al., 2011a) and was found to be necessary for completing full reprogramming of pre-iPSCs into the 2iLIF mESCs (Silva et al., 2009). Similarly, in human, overexpressing *NANOG* together with *KLF2* in primed hPSCs in the naïve medium t2iL led to reprogramming and naïve fate induction (Takashima et al., 2014). Therefore, even though *NANOG* and *Nanog* are not specifically required in the canonical reprogramming, they can still facilitate it in certain conditions.

In primed hPSCs, *NANOG* expression is sustained by FGF and TGF-β signalling, with TGF-β/ACTIVIN-responsive SMADs binding directly to *NANOG* proximal promoter (Xu et al., 2008). Recently, SMAD2/3 signalling was also found to be important in naïve hPSCs (Osnato et al., 2021a). In this study, SMAD2/3 binding was detected upstream of *NANOG*, and, notably, inhibiting this signalling pathway led to the significant decrease of NANOG messenger RNA (mRNA) expression within 2 hours.

The protein complex SMAD2/3 can also directly interact with NANOG protein in primed hPSCs, which in turn assists recruiting histone methyltransferases to ACTIVIN/NODAL signalling targets (Bertero et al., 2015). NANOG protein also modulates SMAD2/3 promoter binding via direct interaction with the complex, this way promoting maintenance of pluripotent genes. Upon withdrawal of NANOG, SMAD2/3 complex becomes capable of binding endoderm-lineage promotors leading to exit from the pluripotency and cell differentiation (Vallier et al., 2009). The exact mechanism of this interaction is unknown; however, Hart and colleagues identified a SMAD4-homologous domain in the N-terminus of NANOG protein (Hart et al., 2004). , which suggests that it could potentially be involved in the complex formation

In addition to interacting with the SMAD proteins in primed hPSCs, NANOG cooperates with OCT4 and SOX2, binding numerous naïve and primed enhancers and super-enhancers, as well as promotors of pluripotency factors, including its own regulatory regions (Chovanec et al., 2021). NANOG directly interacts with other pluripotency factors as well, such as DPPA5 in primed hPSCs, which in its turn was shown to stabilise *NANOG* expression as well as promote self-renewal (Qian et al., 2016). NANOG protein regulation and upstream signalling in the naïve pluripotency is less studied and awaits further research. Molecular properties of NANOG, such as its domain activity and dimerisation ability, are discussed in Chapter 3 and 4 in the context of the findings of this thesis.

In summary, NANOG is an important regulator of pluripotency which promotes activation of pluripotent targets of SMAD2/3 proteins in the primed state, as well as binds multiple enhancer and promotor regions in both naïve and primed hPSCs. In some cases, NANOG acts as a transcriptional

repressor, like in the case of restricting the expression of *OTX2* and *CDX2* (Mendjan et al., 2014; Su et al., 2018). Maintaining its expression at a high level is crucial for the undifferentiated state. In human, *NANOG* function is relatively well studied in the context of primed pluripotency, while naïve roles await future research.

1.3.2 Role of *NANOG* in the pluripotency of non-human species

Orthologs of *NANOG* can be found in the majority of vertebrates and, while the gene structure varies widely between species, its homeodomain DNA sequence, crucial for the DNA binding, is highly conserved in all of them (Booth and Holland, 2004). Ectopic expression of rat, human, chick and zebrafish *Nanog* in mouse *Nanog* -/- somatic cells enabled the rescue of their reprogramming potential and produced induced PSCs in naïve defined medium conditions (Theunissen et al., 2011b) . Moreover, ectopic expression of mouse and zebrafish Nanog homeodomain alone was also sufficient to induce naïve pluripotency in *Nanog* -/- somatic cells (Theunissen et al., 2011b). These observations demonstrate that the conserved homeodomain is responsible for the reprogramming and pluripotency maintenance function of *Nanog* and its orthologs.

In mouse embryos, *Nanog* is essential for the ICM and epiblast formation (Mitsui et al., 2003; Silva et al., 2009), and necessary for survival and proliferation of developing mouse germ cells (Chambers et al., 2007; Yamaguchi et al., 2005; Yamaguchi et al., 2009). In rhesus macaque embryos, *Nanog* overexpression led to improved blastocyst formation with an increased number of ICM cells, suggesting the importance of *Nanog* in the macaque embryo formation as well (Tachibana et al., 2009). *NANOG* knock-down studies in the human embryo have not been performed yet, therefore whether these crucial *NANOG* functions are conserved in human, remains unknow.

An important insight into NANOG protein-protein interaction potential was provided by the assays conveyed in mouse PSCs. According to three independent research publications, NANOG functions as a dimer, and the dimerisation is mediated by a tryptophan-rich region within its C-terminal domain (Mullin et al., 2008; Torres and Watt, 2008; Wang et al., 2008a). In 2017, Mullin and colleagues additionally demonstrated that the interaction is likely to occur via stacking of tryptophan aromatic rings (Mullin et al., 2017). Loss of the tryptophan stretch eliminates Nanog interactions with other pluripotency-associated transcription factors, such as SALL4, NR0B1, ZFP198 and ZFP281 (Wang et al, 2008), while its dimerisation with SOX2 was found to be essential for proper functioning of the latter (Gagliardi et al., 2013)

In vivo regulation of human *NANOG* expression appears to be divergent from that of mouse. First, the genetic interaction between *OCT4* and *NANOG in vivo* in human differs from the interaction between *Oct4* and *Nanog* in the mouse embryo. *Oct4* -/- mouse blastocysts exhibit defects in the ICM

and cannot be maintained; nevertheless, it retains *Nanog* expression (Frum et al., 2013; le Bin et al., 2014), while human (and, interestingly, bovine) *OCT4* -/- embryos fail to form a blastocyst and exhibit overall downregulation of epiblast gene expression, including NANOG (Fogarty et al., 2017; Simmet et al., 2018). This could mean that loss of *OCT4* in human leads to the absence of *NANOG*, directly or indirectly, while in mouse *Nanog* expression is maintained independently of *Oct4*. Interestingly, if TGF/ACTIVIN/NODAL signalling, crucial for maintaining *NANOG* expression, is ablated in human blastocyst, the latter loses its epiblast (Blakeley et al., 2015). This is not the case for mouse and marmoset embryos, which preserve that compartment and *NANOG* expression (Boroviak et al., 2015), suggesting than *NANOG* regulation could utilise different signalling pathways in human early development compared to the other species. This has only been shown in one study, however, and the reason behind the phenotype awaits further investigation.

Interestingly, ectopic increase of *Nanog* expression level in mouse PSCs revealed that it is capable of regulating its own expression through a poorly understood autorepressive feedback mechanism by binding its own promoter (Navarro et al., 2012).

Nanog was also found in less well studied vertebrate species. In chicken ESCs, for example, it is involved in maintenance of pluripotency, along with the *Oct4* orthologue cPouV (Lavial et al., 2007). In felids *Nanog* was demonstrated to have preserved its reprogramming function (Verma et al., 2013). This experiment was aimed to obtain PSCs of endangered felids, such as Bengal tiger, serval and jaguar, to further tackle species preservation (Verma et al., 2013). However, not all vertebrates demonstrate similar conservation of function. In fish species, the *Nanog* orthologue participates in early development *in vivo,* although its function is not essential as in mammals. For instance, in medaka, *Nanog* is crucial for embryo proliferation, but its overexpression and inhibition do not cause any change in mRNA levels of pluripotency or differentiation markers, suggesting that it is not directly involved in embryonic development (Camp et al., 2009). In zebrafish, *Nanog* is required for normal extraembryonic yolk syncytial layer development, but is not required for progression of the embryo (Gagnon et al., 2017)

In conclusion, *Nanog* is a homeobox gene, present in multiple classes of vertebrates. Its function mostly lies in early embryogenesis, and the degree of involvement as well as the regulatory mechanism depend on the species. Despite extensive coding sequence (CDS) divergence, all tested species had *Nanog* reprogramming-to-pluripotency function, proving that conservation within the homeodomain is sufficient for preserving one of the key *Nanog* properties. Most of the knowledge on *Nanog* function came from murine studies, such as NANOG capability to form dimers and autorepress its own expression. Human *NANOG* is relatively less studied, and, therefore, existing knowledge of

mouse *Nanog* properties serve as a valuable source of insights and guidance for *NANOG* research in human pluripotency.

1.4 Gene duplications in evolution and embryo development

1.4.1 Evolutionary role of genomic duplications

Gene duplications play an important role in genome and species evolution. The majority of protein-coding genes in eukaryotes belong to multigene families that have evolved by gene duplication (Ohta, 2000). In addition to a paranome (all paralogous genes in a genome), numerous non-coding regulatory sequences have also multiplied and diverged over the course of evolution (Magadum et al., 2013). The majority of genomic duplications undergo decay due to gene silencing, loss-of-function mutations and/or lack of required regulatory regions (Magadum et al., 2013). However, some survive, with the new copy either acquiring a novel function (neofunctionalisation) or sharing the ancestor function with the parental gene (subfunctionalisation) (Fares, 2014; Force et al., 1999; Kondrashov and Kondrashov, 2006). Therefore, the presence of a new copy of a gene or a regulatory sequence could potentially allow the organism to adapt to a changing environment. Overall, it is clear from evolutionary studies that genomes undergo stochastic turnover, in some cases leading to the formation of novel adaptive functions.

A single gene duplication could have a substantial effect on a cell phenotype and physiology. Gene regulatory network analysis of hPSC and human embryo transcriptomes identified that transcription factors form extensively large interactomes involving numerous protein-protein interactions (Stevens et al., 2019). When a novel gene evolves by duplication, its network interactions could either be lost or conserved while the two copies diverge (Bhan et al., 2002; Vázquez et al., 2003). Additionally, a mathematical model describing proteome evolution showed that after a single duplication event, 40% of protein-protein interactions become independent from the original duplicated component of the protein-protein interaction network (Pastor-Satorras et al., 2003). Thus, a single gene duplication could affect almost one half of the gene regulatory network structure and potentially causing functional divergence (Pastor-Satorras et al., 2003).

1.4.2 Molecular mechanisms of gene duplication

Multiple copies of one ancestral genomic region could be formed by four different mechanisms: polyploidisation, retrotransposition, unequal crossing-over and duplicative transposition (Magadum et al., 2013). Polyploidisation involves multiplication of whole genomes due to

chromosomes failing to separate during DNA replication. Such mechanism is one of the major forces of species formation. At least 70% of flowering plant lineages have gone through this process (Masterson, 1994). Moreover, vertebrates had undergone whole genome duplication twice during their evolutionary progression (Ohno, 1970). Polyploidisation occurring in the human embryo is lethal, however, some adult tissues, such as liver and placenta, can contain polyploid cells (Schoenfelder and Fox, 2015)

The next duplication mechanism, unequal crossing-over, occurs between sister chromatids (mitosis) or homologous chromosomes (meiosis), and leads to the formation of continuous, or tandem repeats (Reams et al., 2012). The outcome of tandem duplication depends on the crossing-over positioning and, as a result, tandem repeats could contain either duplicated gene fragments or fully duplicated gene sequences, sometimes including their regulatory regions as well (Zhang, 2003).

Duplication by retrotransposition involves reverse transcription of mRNA molecules into cDNA followed by their stochastic insertions in the genome (Long et al., 2003). Retrogenes have different structures compared to tandem duplicates, for instance, they do not contain intronic sequences and lack regulatory regions. The latter was hypothesised to be one of the main reasons why such duplications are not maintained for a very long time during the course of evolution (Vinckenbosch et al., 2006). Vinckenbosch and colleagues had also shown that such duplicates have higher probability of expression and 'survival' if inserted near, or even inside 'host' genes, provided that CDS remains undisturbed.

Finally, duplicative transposition, is considered to be the most frequent mechanism of duplication in humans, which is initiated by a double-strand DNA break and is performed via either nonhomologous end joining or nonallelic homologous recombination. The latter requires presence of a homologous template to complete the DNA repair process, whereas duplication by nonhomologous end joining is often linked to the activity of transposable elements (or 'jumping genes'). Homologous recombination often occurs between segmental duplications (SDs; also known as low copy repeats), which contain at least two duplicated homologous regions, while transposons do not require homology and therefore could 'transport' novel copies, leading to more stochastic genome positioning of the duplicates.

1.4.3 Gene duplication in early embryo development

Human pre-implantation epiblast cells are pluripotent and contribute towards all adult somatic and germline lineages (Lawson et al., 1991). PGCs, arising from the pre-implantation epiblast, eventually go on to produce gametes (Leitch et al., 2013). Therefore, gene duplication events occurring in the pre-implantation epiblast would not only impact the developing organism, but could

be propagated onto subsequent generations. Therefore, gene duplications occurring in the early human development could be a major unexplored driver of divergence between mammalian developmental programmes, yet their contribution to early developmental programmes is poorly understood.

An example of gene duplications that could have influenced evolutionary development in the primate lineage are *SRGAP2C and ARHGAP11B.* These two copies are normally co-expressed in the developing human brain alongside their ancestral counterparts - *SRGAP2A* and *ARHGAP11* - and are truncated. Nevertheless, several studies have identified their potential functional roles and hypothesised that they could have had a role in the evolutionary expansion of human neocortex (Charrier et al., 2012; Dennis and Eichler, 2016; Florio et al., 2015). Overexpressing *SRGAP2C* in the developing mouse brain caused formation of human-like neocortex features, including juvenilisation during spine maturation and increased spine density. In addition to that, *SRGAP2C* is able to dimerise with its full-length ancestor *SRGAP2A,* preventing the latter from assembling into functional homodimers, and therefore, potentially blocking the ancestral function. Interestingly, the *SRGAP2*-containing locus duplicated around 2-3 mya immediately before the paleontological estimate of neocortex expansion. Hence, *SRGAP2C* may have contributed to the evolution of the hominin-specific neocortex (Charrier et al., 2012; Dennis and Eichler, 2016). The second gene in this example, *ARHGAP11B,* promotes basal neural progenitor proliferation and human-specific neocortex folding when ectopically expressed in mouse brain, whereas its predecessor, *ARHGAP11,* does not encode human-specific functions (Florio et al., 2015). *ARHGAP11B* appeared in the primate lineage after the divergence from chimpanzee and prior to the split from Neanderthals, which also points to its potential importance in evolutionary expansion of human neocortex (Florio et al., 2015). Overall, these experimental results provide evidence that gene duplicates, even truncated and missing ancestral functional domains, could lead to functional development and evolutional advancement. However, molecular mechanisms of such contributions remain undefined.

Pre-implantation epiblast could also be perceived as a permissive developmental state for duplication mutations to occur, although there is not yet much experimental evidence to support this hypothesis. In this early stage, cells express high levels of chromatin remodelling proteins, making the genome transcriptionally hyperactive and maintaining chromatin in the ‘open’ state, as shown in the mouse ESC system (Efroni et al., 2008). Genome-wide loss of methylation and mostly ‘open’ chromatin state allow for increased transposon-mediated mutagenesis by letting ‘jumping’ genes evade host genome (Hajkova et al., 2002). Additionally, Grow and colleagues have demonstrated that endogenous retroviruses are transcriptionally active in early human development (Grow et al., 2015).

Collectively, these data allow to hypothesise that preimplantation epiblast could provide an unusual period of genome vulnerability, allowing for DNA amplification to occur at a higher rate.

In summary, duplications occurring in the pre-implantation epiblast are not only likely to influence all embryonic lineages, but could also be passed on to new generations. In several particular cases they were shown to have influenced evolution of human-specific brain features. The pre-implantation epiblast state could also be particularly permissive for the duplications to occur, although this hypothesis would require additional investigation.

1.4.4 *NANOG*, a highly duplicated human pluripotency gene

The pluripotency transcription factor *NANOG* has an unusually high number of duplicated copies in the human genome, eleven in total (Booth and Holland, 2004). Therefore, it serves as a paradigm for studying the impact of gene duplication events on early embryonic development. In human pluripotency, *NANOG* serves as a critical transcriptional regulator, expressed in human epiblast, and its *in vitro* counterpart hPSCs, as well as primordial germ cells (PGCs). Elevated levels of *NANOG* expression are crucial for maintaining the pluripotent state in human (Section 1.3) and thus it is possible that *NANOG* pseudogenes, if expressed, could have overlooked roles in the regulation of pluripotency.

Ten of the eleven duplicates (*NANOGP2-NANOGP11*) were formed by retrotransposition and are therefore processed pseudogenes that lack regulatory sequences and possess various mutations that led to their functional decay (Booth and Holland, 2004). Only one member of the human *NANOG* pseudogene family, *NANOGP1*, has not been processed (Booth and Holland, 2004)(Figure 1.4).

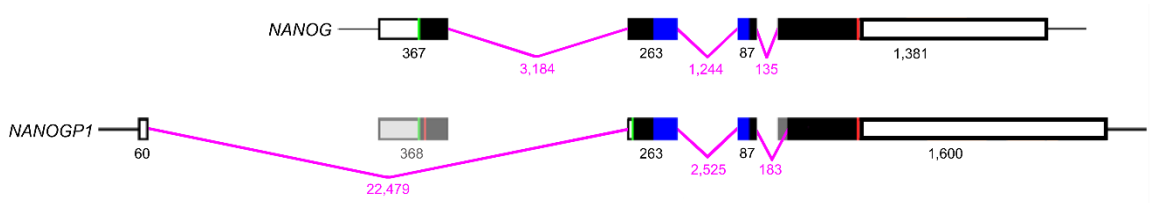


Figure 1.4 Diagram showing comparison of the two *NANOGP1* mRNA variants, predicted by Booth and Holland, 2004, with *NANOG* (adapted from Booth and Holland, 2004). Exons are shown as black and blue blocks, and 5' and 3' UTR as white blocks, respectively. Homeodomain is indicated as a blue block. Introns are shown as pink lines. Start (green) and stop (red) labels mark the predicted location of start and stop codons. mRNA component length is shown as a number of nucleobase pairs.

Human *NANOG* and *NANOGP1* share 97% CDS homology and have a similar predicted exon-intron structure. Importantly, the predicted *NANOGP1* CDS, if expressed, could potentially generate a protein with an intact homeodomain, presumably with the same DNA recognition and binding

capabilities as NANOG (Booth and Holland, 2004). However, while the *NANOGP1* sequence was shown to be highly conserved with *NANOG*, none of the previous studies have identified what open reading frame(s) *NANOGP1* uses in hPSCs and/or how its mRNA might be processed. Therefore, the two *NANOGP1* ORF variants shown in Figure 1.4 are hypothetical and require further investigating.

In summary, the existence of a highly conserved tandem duplicate of a key pluripotency factor indicates its potential functionality and emphasises the need for its detailed investigation.

1.4.5 History and limitations in studying *NANOGP1* pseudogene

According to the published literature, *NANOGP1* pseudogene was not investigated in detail, and therefore its potential role in pluripotency, and whether it encodes a protein in hPSCs, remains poorly understood. An evolutionary study that compared human and chimpanzee genomes suggested that *NANOGP1* could be expressed, and its ORF has probably undergone selection-driven conservation (Fairbanks and Maughan, 2006). Indeed, according to Booth and Holland, 2004, the predicted *NANOGP1* CDS is largely homologous to that of *NANOG* (Figure 1.4). However, as the two authors concluded, if *NANOGP1* ORF were to use the same ATG as *NANOG*, the protein would have been highly shortened containing only the first 8 amino acids. This is because *NANOGP1* has a cytosine to thymine substitution mutation at codon 25, which introduced a premature stop codon. Alternatively, *NANOGP1* could use a completely different start codon, 58 positions downstream of the *NANOG* ATG (Booth and Holland, 2004). This predicted NANOGP1 protein would include a conserved homeodomain and transactivation domain, which are involved in protein dimerisation, DNA binding and are responsible for pluripotency maintenance in mouse *Nanog* and its orthologs (Chambers et al., 2003; Chang et al., 2009; Hart et al., 2004; Mullin et al., 2020; Oh et al., 2005; Theunissen et al., 2011a). Therefore, it is possible that *NANOGP1* could encode a functional and nearly full-length protein which would lack the first *NANOG* exon and the very beginning of exon 2.

NANOG and *NANOGP1* transcripts have been detected in leukaemia cells (Eberle et al., 2010). This study also demonstrated that when these *NANOGP1* transcripts are ectopically expressed, they could be translated into three, nearly full-length protein variants. Another study showed that *NANOGP1* mRNA is also detectable by RT-qPCR in primed hPSCs, adult testes and epididymal tumours (Hart et al., 2004).

Whether *NANOGP1* can be translated into a protein in hPSCs is a controversial topic. The divergence between the predicted *NANOGP1* ORF and *NANOG* ORF, proposed by Booth and Holland, led to the belief that *NANOGP1* does not encode a protein (Booth and Holland, 2004), and *NANOGP1* is still classified as a non-protein encoding pseudogene in repositories like Ensembl. The study, however, did not analyse a full spectrum of synonymous and non-synonymous mutations between

the two genes due to high polymorphism of complementary DNA (cDNA) and ESTs available at that time. Additionally, the study by Booth and Holland did not fully explore the possibility that their second predicted variant, which encodes the homeodomain and the C-terminal domain, could have a functional role.

In summary, *NANOGP1* is an unprocessed tandem duplicate of *NANOG* with a highly conserved CDS. *NANOGP1* mRNA was detected in several human cell lines and tissues, including primed hPSCs. *NANOGP1* transcripts, detected in leukaemia cells, were also able to produce nearly full-length protein variants when ectopically expressed from a plasmid. However, based on the study conducted in 2004, *NANOGP1* was annotated as non-protein coding and therefore, it is not known whether it encodes a protein in hPSCs and if it is functional in human pluripotency.

1.5 Hypothesis

In this study, I hypothesise that *NANOG*/*NANOGP1* tandem duplication is a tractable model of gene duplication influencing early development.

NANOG is one of the core pluripotency transcription factors, expressed in human, mouse, NHPs and other species, where it has properties affecting regulation of early development (Section 1.3). In human, *NANOG* had been studied less well compared to the similar investigations in mouse, and therefore not all of its functions are currently understood there. It is clear, however, that *NANOG* expression is highly important for the maintenance of the undifferentiated state in both primed and naïve hPSCs. Therefore, an unprocessed tandem duplicate, highly conserved at a sequence level (Section 1.4.5), could also have an overlooked role, with significant relevance to human development.

Several independent *NANOGP1* studies, described in Section 1.4.6, demonstrated that *NANOGP1* mRNA can be detected *in vivo* and in various cell models, including primed hPSCs. *NANOGP1* mRNA, detected in leukaemia cells, also had the ability to translate into a protein when it was expressed from a plasmid. Moreover, the predicted CDS of *NANOGP1* also has conserved homeodomain and transactivation domain encoding regions. Finally, the past conclusion that *NANOGP1* cannot encode a protein was based on bioinformatic analysis and has not tested experimentally in cells that express endogenous *NANOGP1*.

Here I hypothesise that, based on the *NANOGP1* sequence, mRNA expression, and the overlooked potential ORF conservation, human *NANOGP1* could encode a protein with a potential functional role in human pluripotency. I also hypothesise that *NANOGP1* could have retained its ancestral function and/or has developed a novel role.

Collectively, this points to the necessity of studying *NANOGP1* in human pluripotency, which can be achieved now with the advanced experimental and bioinformatics tools presently available. In

this study, the efficiency of distinguishing between *NANOG* and *NANOGP1* expression ought to be higher than in 2004, with more datasets and bioinformatic tools available. Additionally, gene editing technologies that became available in the past few years have opened up new resources for *NANOGP1* epitope tagging and *NANOGP1* functional analysis.

Overall, filling the knowledge gap regarding the potential functionality of *NANOG* tandem duplicate could help understand the specifics of human early development, why and how it is different from the in-depth studied murine development, and discuss potential implications of unprocessed pseudogenes in early embryonic processes.

1.6 Aims of the project

Based on the Hypothesis, I created the following aims, containing fundamental questions about *NANOG*/*NANOGP1* duplication evolution and potential roles in human pluripotency.

1. Investigate evolutionary history and conservation of *NANOG*/*NANOGP1* duplication site
2. Investigate *NANOGP1* expression *in vivo* (human embryo) and *in vitro* (naïve and primed hPSCs)
3. Investigate whether NANOGP1 protein is expressed in hPSCs, and if it is, explore its chromatin binding profile using ChiP-seq
4. Investigate whether recombinant NANOGP1 protein is capable of homodimerising and/or forming heterodimers with recombinant NANOG
5. Investigate whether *NANOGP1* has any conserved and/or novel functions in hPSCs

2 Materials and Methods

2.1 Human pluripotent stem cell culture

2.1.1 Human pluripotent stem cell lines

hPSC lines used in this study are:

- CR-H9 naïve female hPSCs, derived from primed WA09/H9 hPSCs by chemical reprogramming (Guo et al., 2017)
- CRISPRi Gen1B primed and naïve female hPSCs, reprogrammed (Mandegar et al., 2016)
- HNES1 naïve male hPSCs, embryo-derived (Guo et al., 2016)
- WA09/H9 NK2 naïve female hPSCs, created by reprogramming primed H9 hPSCs transfected with *NANOG* and *KLF2* Doxycycline-inducible transgenes (Takashima et al., 2014)
- WA09/H9 primed female hPSCs, embryo-derived (Thomson, 1998)

All hPSC lines used in this study (except the reprogrammed lines) are listed in the UK Stem Cell Line Registry and all of the work was carried out with the appropriate approval by the UK Stem Cell Bank Steering Committee and BI committees.

2.1.2 Human pluripotent stem cell culture maintenance

All hPSC lines were maintained at 5% O_2, 5% CO_2 at 37°C in a humidified incubator. All maintenance culture medium was changed every day. The Countess™ Automated Cell Counter (ThermoFisher Scientific) or a glass haemocytometer were used for cell counting. See Table 2.1 for the reagent details.

Table 2.1 hPSC culture reagents

Reagent name	Reference	Company
Accutase	423201	Biolegend
Advanced DMEM	12491023	ThermoFisher Scientific
B27	17504-044	ThermoFisher Scientific
bFGF		WT-MRC Cambridge Stem Cell Institute
Blasticidin	A1113902	ThermoFisher Scientific
b-mercaptoethanol	31350-010	ThermoFisher Scientific
Bovine Serum Albumin	15260037	ThermoFisher Scientific
CHIR99021		WT-MRC Cambridge Stem Cell Institute
Collagenase, Type IV	17104019	ThermoFisher Scientific
Corning Matrigel Matrix (GFR)	354230	Scientific Laboratory Supplies
DMEM, high glucose	41965-062	ThermoFisher Scientific
DMEM/F-12	31330-095	ThermoFisher Scientific
Doxycycline	4090	Tocris
Foetal Bovine Serum	F7524	Sigma-Aldrich

Geltrex™	A1413302	ThermoFisher Scientific
Go6983	2285/10	Tocris
IM12	SM04-100	Cell Guidance Systems
Irradiated MF1 MEFs		WT-MRC Cambridge Stem Cell Institute
KnockOut™ Serum Replacement	10828028	ThermoFisher Scientific
L-glutamine	25030-024	ThermoFisher Scientific
MEM Non-Essential Amino Acids Solution	11140050	ThermoFisher Scientific
MTeSR-Plus	5825	Stem Cell Technologies
N2	17502-048	ThermoFisher Scientific
Neurobasal medium	21103-049	ThermoFisher Scientific
PD0325901		WT-MRC Cambridge Stem Cell Institute
Penicillin/Streptomycin	15140122	ThermoFisher Scientific
Phosphate-buffered saline buffer	D8537	Sigma-Aldrich
Puromycin	73342	StemCell Technologies
Recombinant human LIF		WT-MRC Cambridge Stem Cell Institute
SB431542	SM33-10	Cell Guidance Systems
SB590885	SM48-1	Cell Guidance Systems
Sodium Pyruvate	11360039	ThermoFisher Scientific
TeSR™-E8™	5990	StemCell Technologies
Valproic acid sodium salt	P4543	Sigma-Aldrich
Vitronectin	A14700	ThermoFisher Scientific
WH-4-023	2827-10	Cell Guidance Systems
XAV939	X3004	Sigma-Aldrich
Y-27632	SM02-10	Cell Guidance Systems

Irradiated MF1 mouse embryonic fibroblasts (MEFs) for primed and naïve hPSC feeder-dependent maintenance were plated in Dulbecco's modified Eagle medium (DMEM) high glucose, 10% foetal bovine serum (FBS), 2 mM L-glutamine, 1 mM sodium pyruvate, 0.1 mM MEM non-essential amino acids, 50 U/ml and 50 µg/ml penicillin-streptomycin, 0.1 mM β-mercaptoethanol on sterile plastic flasks pre-treated with 0.1% gelatine in phosphate-buffered saline buffer (PBS). MEFs were seeded at a density of $3.5x10^4/cm^2$ for most experiments or $5x10^4/cm^2$ for primed-to-naïve reprogramming experiments.

Feeder-dependent primed hPSCs were cultured on a layer of irradiated MF1 MEFs in Advanced DMEM, 20% KnockOut Serum Replacement (KSR), 2 mM L-glutamine, 50 U/ml and 50 µg/ml penicillin-streptomycin, 0.1 mM b-mercaptoethanol supplemented with 4 ng/ml bFGF (**'KSR/FGF medium'**, (Thomson, 1998). For passaging, cell medium was aspirated and replaced with 200 U/ml collagenase (Advanced DMEM, 20% KSR, 2mM L-glutamine, Collagenase IV), and cells were incubated for 5 min. After the incubation, collagenase was replaced with KSR/FGF medium, and cells were detached from the plate by gentle scraping with a plastic serological pipette. KSR/FGF medium with the detached cells was aspirated using the same pipette, ensuring that the cells were in small

aggregates and not single cells to support optimal survival. hPSCs were plated at a cell density of ~2.5-5x10^4/cm^2.

Feeder-free primed hPSCs was maintained in TeSR™-E8™ (routine maintenance) or mTeSR™-Plus (nucleofection and cell sorting experiments) medium on sterile plastic flasks pre-treated with 5 µg/ml Vitronectin in PBS. Feeder-free cultures were passaged by incubating for 5 min in Gentle Cell Dissociation Buffer (GCDB; 0.5 mM EDTA in PBS), aspirating the GCDB from the wells, re-suspending the cells in a fresh culture medium and passaging at a 1:4 – 1:20 split ratio for a routine near-confluent culture. Of note, in general single cell dissociation during passaging was avoided and the cells were passaged as small aggregates. For transfection and cell sorting experiments that required single cell dissociation, 10 µM Y-27632 (ROCK inhibitor) was added to the culture medium for 24 h after the passage, in order to increase cell viability (Watanabe et al., 2007).

Feeder-dependent and feeder-free naïve hPSCs were cultured in N2B27 medium (1:1 DMEM/F12 and Neurobasal, 1X B27-supplement, 1X N2-supplement, 2 mM L-Glutamine, 50 U/ml and 50 µg/ml penicillin-streptomycin, 0.1 mM β-mercaptoethanol) supplemented with 1 µM PD0325901, 1 µM CHIR99021, 2 µM Go6983 and 20 ng/ml human LIF (‘**t2iLGo medium’**, (Takashima et al., 2014)) or 1 µM PD0325901, 2 µM Go6983, 20 ng/ml human LIF and 2 µM XAV939 (**‘PXGL medium’**, (Bredenkamp et al., 2019b; Rostovskaya et al., 2019)).

Feeder-dependent naïve hPSCs were cultured on a layer of irradiated MF1 MEFs. t2iLGo naïve feeder-free culture was maintained on Matrigel® Matrix instead of MF1 MEFs. Cell culture plates were pre-coated with Matrigel in DMEM, 1:100 dilution, and incubated overnight at 4°C. Prior to cell passaging, Matrigel was aspirated and replaced with cell culture medium. In PXGL naïve hPSC culture, Geltrex™ Matrix was added to both feeder and feeder-free culture medium during cell replating at a 1:300 dilution.

For passaging, feeder-dependent and feeder-free naïve hPSCs were dissociated to single cells by incubating with Accutase for 5 minutes, collected in 15 ml Falcon tubes with wash buffer (0.01% BSA in N2B27, 2x of the Accutase volume total) and collected by centrifugation at 300xg for 3 min. After aspirating the wash buffer, cell pellets were re-suspended in the culture medium and passaged to fresh feeder or feeder-free plates at a density of 2x10^4/cm.

2.1.3 Cryopreservation of human pluripotent stem cells

Cryopreservation was used for long-term storage of naïve and primed hPSCs. In order to prepare cultures for cryopreservation, 1.5-2x10^6 cells were resuspended in 1 ml of freezing medium, transferred to a cryovial and stored at -80°C for at least 24 h. Then, cryovials were placed into liquid nitrogen tanks, where they could be stored indefinitely. In order to return cryopreserved cells into

culture, they were taken out of the liquid nitrogen tanks, thawed and passaged as normal. Freezing medium was based on hPSC maintenance medium (specific to the cell type, i.e., t2iLGo, PXGL) and also additionally contained 10% dimethyl sulfoxide (DMSO).

2.1.4 Cell transfection methods

2.1.4.1 Nucleofection

The Amaxa™ Nucleofector™ 4D (Lonza) and Neon™ Transfection system (ThermoFisher Scientific) were used for transfecting primed and naïve hPSCs, according to their respective manuals.

Nucleofector™ 4D was used for creating CRISPRi and TetOn hPSC lines. Briefly, plasmids were added to a $1x10^5$ cell pellet of primed CRISPRi Gen1B WT (wild type) hPSCs or primed H9 WT hPSCs which was re-suspended in 20 µl Amaxa™ P3 Primary Cell Solution + Supplement 1. Alternatively, the reaction was scaled-up by transfecting a $1x10^6$ cell pellet in 200 µl P3 Primary Cell Solution + Supplement 1. The nucleofection cell solution was then carefully added to an Amaxa™ Nucleo-microcuvette. The microcuvette was then placed into the Amaxa™ Nucleofector™ 4D and the cells were nucleofected using the settings CB150. Immediately after nucleofection, 150 µl of appropriate culture medium was added to the nucleofection cell solution, then, the whole volume was collected and carefully added to a well on a 12-well plate with pre-warmed culture medium, supplemented with 10 µM Y-27632. After 18 h, the culture medium was replaced with fresh media without Y-27632.

The Neon™ Transfection system was used for NANOGP1 epitope tagging. A $1x10^6$ cell pellet of CR-H9 hPSCs (per one reaction) was resuspended in 100 µl Buffer R (mixed with the CRISPR reagents (see Section 2.1.5)), aspirated using a Neon™ Pipette with a Neon™ 100 µl Pipette tip. The Neon™ Pipette tip, filled with the transfection mix, was then inserted into a Neon™ Tube in the Neon™ Pipette station, containing 3 ml Buffer E2. Transfection was performed using the 1300 V, 30 ms, 1 pulse settings. After the transfection, cells were seeded into PXGL naïve medium, supplemented with 10 µM Y-27632 for increased cell survival.

2.1.5 NANOGP1 epitope tagging

The experiment was performed using CRISPR/Cas12a gene editing method, described in Zetsche et al., 2015. Reagents were synthesised by IDT. During the reagent validation screen, crRNA_5' Cas12a was chosen out of six crRNA reagents as the most efficient one (see Chapter 4) to be used in *NANOGP1* epitope tagging. A *NANOGP1* Cas12a crRNA sequence targeting 10 bp upstream of the *NANOGP1* ATG site, and a repair template containing an epitope tag (V5 or 3xFLAG) + homology arms, were designed using CRISPOR (http://crispor.tefor.net/) and nucleotide sequence editing software

SnapGene. crRNA sequences can be found in Table 2.2. See Table 2.3, for the single-stranded oligodeoxynucleotide (ssODN) homology directed repair (HDR) sequences. For cell nucleofection, 5.6 µg NANOGP1 Alt-R® A.s. Cas12a crRNA and 40 µg Alt-R® A.s. Cas12a Ultra protein were pre-assembled for 15 min at RT, and then combined with 2 µl 200 pmol/µl repair template. The CRISPR reagents were then used for transfecting CR-H9 naïve hPSCs with the Neon™ Transfection System. Each transfection reaction was performed using $1x10^6$ cells per pellet. After the transfection, hPSCs were seeded into PXGL naïve medium, and, before being return to a 32°C incubator, were incubated in 'cold shock' conditions (Guo et al., 2018) at 32°C for 24 hr with 5% O_2, 5% CO_2 in a humidified incubator. Additionally, 2 µM M3814 (DNA-dependent protein kinase inhibitor) was added to the culture medium to repress the non-homologous end joining (NHEJ) DNA repair mechanism and aiming to increase HDR efficiency (Riesenberg et al., 2019). To improve survival, Y-27632 was added to the cell culture medium 2 h prior to the transfection and to the post-transfection medium. PXGL+M3814+Y-27632 culture medium was not replaced for 72 h after the transfection.

Table 2.2 crRNA molecules tested in *NANOGP1* and *NANOG* epitope tagging screening assay.

crRNA name	**crRNA sequence, 5'-3'**
crRNA_3' Cas9 *NANOG*	GATATTACTCAATTTCAGTC
crRNA_3' Cas9 *NANOGP1*	GATGTTACTCAGTTTCAGTC
crRNA_5' Cas12a *NANOGP1*	TGGGCCTGAAGAAAACCATCC
crRNA_5' Cas9 *NANOG*	CCTCTATACTAACATGAGTG
crRNA_5' Cas9 *NANOGP1_1*	GGCCCACAAATCACAGGTAT
crRNA_5' Cas9 *NANOGP1_2*	GAGATGCCTCACACAGAGAC

Table 2.3 ssODN templates used in the *NANOGP1* epitope tagging experiment. AS – antisense strand. S – sense strand. Tag sequence is in bold. Homology arms are in capital letters.

ssODN name	**ssODN sequence, 5'-3'**
NANOGP1_3xFLAG_AS	TTACCAGTCTCTGTGTGAGGCATCTCAGCAGAAGACATTTGCAAGGATGG**cttgtcatcgtcatcctgtaatcgatgtcatgatctttataatcaccgtcatggtctttgtagtc**CATATGGTTTTCTTCAGGCCCACAAATCACAGGTATAGGTGACCAGTCTTTAC
NANOGP1_V5_S	GTAAAGACTGGTCACCTATACCTGTGATTTGTGGGCCTGAAGAAAACCATATG**ggtaagcctatccctaaccctctcctcggtctcgattctacg**CCATCCTTGCAAATGTCTTCTGCTGAGATGCCTCACACAGAGACTGGTAA

2.1.6 Inducible gene expression knock-down (CRISPRi) cell line generation:

dCas9-iKRAB Gen1B CRISPRi *NANOGP1* and dCas9-iKRAB Gen1B CRISPRi *NANOG* hPSCs lines were generated as described in Mandegar et al., 2016. Gene-specific gRNA oligonucleotides were designed using the IDT gRNA design tool or are from published literature (Mandegar et al., 2016) and were synthesised by Sigma. The oligonucleotides were phsopho-annealed and cloned into pgRNA-CKB vector (CAG-mKate2-T2A-bsd^R, Mandegar et al., 2016), pre-digested with *BsmBI* (NEB) and pre-treated with FastAP (Life Technologies). Linearised vector and phospho-annealed gRNA oligonucleotides were ligated at RT overnight with T4 DNA Ligase (Invitrogen). Ligated products were validated by subcloning into competent *E. coli* with 100 µg/ml Ampicillin selection and by Sanger sequencing (Genewiz). Primed CRISPRi Gen1B WT hPSCs, kindly provided by Prof. Bruce Conklin (Gladstone Institutes, USA), were nucleofected with the *NANOGP1* or *NANOG* gRNA plasmids using Amaxa™ Nucleofector™ 4D, selected by blasticidin treatment (8 µg/ml for five days) and flow sorted for mKate2 expression. Primed CRISPRi Gen1B *NANOGP1* and *NANOG* cell lines were reprogrammed into the naïve state using 5i/L/A reprogramming method, adapted from Theunissen et al, 2014. On Day 0, primed feeder-free cultures were passaged onto feeders in mTeSR™ Plus medium supplemented with 10 µM Y-27632 at a density of 2x10^4 per cm^2. On Day 1, mTeSR™ Plus was replaced with 5i/L/A reprogramming medium, composed of N2B27 base supplemented with 1 µM PD0325901, 20 ng/ml human LIF, 20 ng/ml Activin A, 1 µM IM12, 0.5 µM SB590885, 10 µM Y-27632 and 1 µM WH-4-023. See reagent details in Section 2.1. Cultures were passaged every 5 days and transferred to t2iLGo medium on Day 18.

Sequencing primers can be found in Table 2.7; Section 2.3.4.

Forward and reversed primers used for gRNA cloning can be found in Table 2.4.

Table 2.4 Primers designed for the pgRNA-CKB gRNA cloning. +/- values, distance from the gRNA PAM (Protospacer adjacent motif) site to the target gene transcription start site (TSS) in bp; '+' indicates upstream location and '-' indicates downstream location. 'T' and 'NT' indicate whether the gRNA targets the template or non-template strand, respectively. TTGG and AAAC in bold – overhangs added to clone phospho-annealed oligonucleotides to pgRNA-CKB using *BsmBI* restriction.

Primer pair name	**Forward primer sequence, 5'-3'**	**Reverse primer sequence, 5'-3'**
gRNA-71_N (T)	**TTGG**ATTCACAAGGGTGGGTCAGT	**AAAC**ACTGACCCACCCTTGTGAAT
gRNA+18_N (NT)	**TTGG**CCAGCAGAACGTTAAAATCC	**AAAC**GGATTTTAACGTTCTGCTGG
gRNA+252_N (NT)	**TTGG**CAGTCGGATGCTTCAAAGCA	**AAAC**TGCTTTGAAGCATCCGACTG

gRNA+358_S (T)	**TTGG**TTCTGCTGAGATGCCTCACA	**AAAC**TGTGAGGCATCTCAGCAGAA
gRNA-120_S (NT)	**TTGG**CCCGTCTACCAGTCTCACCA	**AAAC**TGGTGAGACTGGTAGACGGG
gRNA-148_S (NT)	**TTGG**CAGAGTAACCCAGACTAGGT	**AAAC**ACCTAGTCTGGGTTACTCTG
gRNA+119_P1 (NT)	**TTGG**TGAGTCGCCTCCACAATAAC	**AAAC**GTTATTGTGGAGGCGACTCA
gRNA+116_P1 (NT)	**TTGG**GTCGCCTCCACAATAACAGG	**AAAC**CCTGTTATTGTGGAGGCGAC
gRNA+348_P1 (NT)	**TTGG**GGCCCACAAATCACAGGTAT	**AAAC**ATACCTGTGATTTGTGGGCC
gRNA+268_P1 (NT)	**TTGG**ATCCACACTCATGTCATTAT	**AAAC**ATAATGACATGAGTGTGGAT
gRNA+57_P1 (NT)	**TTGG**AGTCTTTAGATTTATAATGA	**AAAC**TCATTATAAATCTAAAGACT

2.1.7 Inducible gene overexpression (TetON) cell line generation:

TetON gene overexpression H9 hPSC lines generated and used in this thesis are:

1. TetON-*NANOG-GFP* + CAG-rtTa-Puro^R
2. TetON-*NANOGP1-1-GFP* + CAG-rtTa-Puro^R
3. TetON-*NANOG-GFP* + TetON-*KLF2-RFP* + CAG-rtTa-Puro^R
4. TetON-*NANOGP1-1-GFP* + TetON-*KLF2-RFP* + CAG-rtTa-Puro^R
5. TetON-*NANOGP1-2-GFP* + TetON-*KLF2-RFP +* CAG-rtTa-Puro^R
6. TetON-*NANOGP1-3-GFP* + TetON-*KLF2-RFP* + CAG-rtTa-Puro^R
7. TetON-*KLF2-RFP* + CAG-rtTa-Puro^R

This section describes the generation of inducible overexpression primed hPSC lines. Primed-to-naïve reprogramming of these seven lines is addressed in Sections 2.1.8. For simplicity, full hPSC line names are used in this section only; in the rest of the text, only the first part of the name (TetON-*gene-reporter*) is used.

To create TetON inducible gene overexpression vectors, gene cDNA was synthesised as a gBlock Gene Fragment (IDT), cloned into a pCAG-IRES-Puro^R backbone vector (Niwa et al., 1991) and amplified with primers containing an *attB* sequence at their 5' ends. All three *NANOGP1* isoforms shared the same forward attB primer. All *NANOGP1* isoforms and *NANOG* shared the same reverse attB primer.

attB primer sequences are shown in Table 2.5 attB primer sequences used for generating TetON hPSC lines. The amplification product (attB-gene_cDNA-attB) was cloned into a TetON-*GFP/RFP* plasmid (kindly provided by Andras Nagy (Lunenfeld-Tanenbaum Research Institute, Canada); (Woltjen et al., 2009) using the Gateway strategy (Hartley, 2000; Hartley, 2002) and was validated by Sanger sequencing (Genewiz). A cell pellet of $1x10^6$ primed H9 hPSCs were then transfected with 2 µg of a vector encoding constitutively expressed reverse tetracycline-regulated transactivator gene (pCAG-rtTa-Puro^R), followed by 48 hr selection with 1 µg/ml Puromycin. Then, CAG-rtTA-Puro^R H9 hPSCs were transfected with 100 ng of a vector encoding a piggyBac transposase (pCyL43) (Wang et al., 2008b), as well as with the four individual TetON plasmids: *NANOG-GFP*, *NANOGP1-1-GFP*, *NANOGP1-2-GFP* and *NANOGP1-3-GFP*, 2 µg each. This resulted in the generation of four separate hPSCs lines, which, after expansion, were treated with 1 µM Doxycycline for 48 h and then were flow sorted for GFP positive expression. The stable cell lines were expanded and stored.

In a separate experiment, the four TetOn cell lines described above, as well as the CAG-rtTA-Puro^R only hPSC line, were transfected with 100 ng of pCyL43 and 2 µg TetON-*KLF2*-RFP vector. After expansion, the cell lines were treated with 1 µM Doxycycline for 48 h and then were flow sorted for RFP positive expression. The stable cell lines were expanded and stored.

Table 2.5 attB primer sequences used for generating TetON hPSC lines. attB sequences are in bold.

Primer name	**Primer sequence (5'-3')**
attB-*NANOGP1*-F	**GGGGACAAGTTTGTACAAAAAAGCAGGCTCT**ATGTCTTCTGCTGAGATGCC
attB-*NANOG*-F	**GGGGACAAGTTTGTACAAAAAAGCAGGCTCT**ATGAGTGTGGATCCAGCTTG
attB-*NANOGP1*/*NANOG*-R	**GGGGACCACTTTGTACAAGAAAGCTGGGTC**TCACACGTCTTCAGGTTGC
attB-*KLF2*-F	**GGGGACAAGTTTGTACAAAAAAGCAGGCTCT**ATGGCGCTGAGTGAACCC
attB-*KLF2*-R	**GGGGACCACTTTGTACAAGAAAGCTGGGTC**CTACATGTGCCGTTTCATGTGC

2.1.8 Primed-to-naive human pluripotent stem cell reprogramming

2.1.8.1 Chemical reprogramming method

Primed TetON-*iNANOGP1-1-GFP* and TetON-*iNANOG-GFP* H9 hPSC lines were reprogrammed into the naïve state using the chemical reprogramming method, as described in Guo et al*.,* 2017. On Day 0, primed feeder-free hPSC cultures were passaged onto feeders in mTeSR™ Plus medium supplemented with 10 µM Y-27632 at a density of $1x10^4$ per cm^2. On Day 1, the culture medium was replaced with mTeSR™ Plus without Y-27632. On Day 2, the medium was changed to chemical reprogramming medium 1 (cRM-1) based on N2B27 supplemented with 1 µM PD0325901, 10 ng/ml

human LIF and 1 mM valproic acid (VPA) sodium salt. No medium change took place on Day 3, but, starting from Day 4, medium was changed daily. On Day 5, cRM-1 medium was aspirated from the plates and replaced with chemical reprogramming medium 2 (cRM-2) based on N2B27 medium supplemented with 1 µM PD0325901, 10 ng/ml human LIF, 2 µM Go6983 and 2 µM XAV939. Dome-shaped colonies became apparent by Day 9. Cells were passaged onto feeders in cRM-2 medium at a 1:1 or 1:2 split ratio. After several passages the culture would become homogeneous and could be passaged at a normal, 1:3 or 1:4 split ratio and maintained in PXGL or t2iLGo medium.

2.1.8.2 *iNANOGP1/iNANOG+iKLF2* reprogramming method

Primed *TetON-iNANOG-GFP-iKLF2-RFP*, *TetON-iNANOGP1-1-GFP-iKLF2-RFP, TetON-iNANOGP1-2-GFP-iKLF2-RFP*, *TetON-iNANOGP1-3-GFP-iKLF2-RFP* H9 hPSC lines were reprogrammed as described in Takashima et al., 2014. Briefly, prior to the reprogramming, on Day -2, feeder-free primed hPSCs were treated with 1 µM Doxycycline for 48 h and sorted by GFP/RFP double-positive fluorescence to establish transgene lines with similar levels of reporter expression between lines. On Day 0, flow sorted transgene lines were plated on feeders in KSR/FGF on Day 0, and on the following day, culture medium was changed to KSR/FGF supplemented with 1 µM Doxycycline. On Day 2, medium was changed to t2iLIF medium, composed of N2B27 base medium with 1 µM PD0325901, 10 ng/ml human LIF, 1 µM CHIR99021 and supplemented with 1 µM Doxycycline. t2iLIF medium was changed daily and cells were split every 5 days. On Day 12, Doxycycline was withdrawn and 5 µM Go6983 was added. Reprogrammed cells were further propagated in t2iLIF + 5 µM Go6983 medium on feeders.

2.1.9 Formative capacitation of naïve human pluripotent stem cells

Naïve hPSCs were capacitated to a formative state as described in Rostovskaya et al., 2019. On Day 0, naïve TetON-*iNANOGP1-1-GFP* CR-H9 hPSCs were seeded in PXGL medium supplemented with 10 µM Y-27632 in feeder-free conditions on plates pre-coated with Geltrex at a seeding density of $1.6x10^4$ per cm^2. The following day, culture medium was replaced with PXGL without Y-27632. On Day 2, medium was replaced with N2B27 supplemented with 2 µM XAV939, either with or without 1 µM Doxycycline. Medium was replaced every day and cells were passaged at the 1:2 ratio when confluent. In total, hPSCs were cultured in N2B27 with XAV939 and +/-Doxycycline for 14 days.

2.1.10 Colony formation assay and alkaline phosphatase staining

Alkaline phosphatase (AP) is a cell-surface protein expressed by PSCs; in this thesis formation of hPSC colonies was assessed by AP activity (Štefková et al., 2015). The AP activity assay contained pluripotent colonies that were characterised by high levels of AP activity and appeared purple (AP^{pos}), as well as non-pluripotent/differentiated colonies that were colourless (AP^{neg}).

hPSCs were dissociated into single cells and plated into the experiment-specific medium in 6-well plates. On Day 12 (Figure 5.53), the cells were assayed for AP activity and imaged using a Zeiss PALM MicroBeam. The AP assay involved fixing cells with 4% paraformaldehyde (PFA; Agar scientific, R1026) in PBS, followed by incubation in alkaline phosphatase staining solution (Millipore, SCR004) for 15 min and washing with PBS twice to stop the reaction. The colonies were counted and categorised as described above.

2.2 Flow cytometry assays

PSCs were dissociated with Accutase, washed with 2% FBS/PBS and filtered through 50 µm sterile strainers (Sysmex, 1050553). If the signal was conferred by a fluorescent reporter, no antibodies were added to the cell suspension. If additional fluorescent antibody staining was required (i.e., for a naïve or primed cell surface markers), the following protocol was used: hPSCs were incubated in Brilliant Stain Buffer (BD Biosciences, 563794) with antibodies for 30 min at 4°C in the dark, followed by a wash with 2% FBS/PBS, pelleting at 300xg for 3 min and re-suspension in 300 µl 2% FBS/PBS. Antibody details are summarised in Table 2.6. The strategy for the antibody choice can be found in the experimental set-up (described in the individual Results chapters). If hPSCs were grown on feeders, anti-mouse Cd90.2 antibody that recognises mouse fibroblasts was added to the staining mix as well. To distinguish live and dead cells, Fixable Viability Dye eFluor 780 antibody was used; alternatively, 0.1 µg/mL DAPI (Tocris, 5748) was added to the cell suspension after the immunostaining, at least 5 min prior to the analysis.

Flow cytometry analysis was performed on BD LSR-Fortessa™. Cell sorting experiments were performed on BD Influx™ or BD FACSAria™ Fusion. Before the fluorescence analysis, cells were gated using **a.** forward-scatter (FSC) and side-scatter (SSC) to exclude debris, **b.** SSC/FSC Area and Height to exclude cell doublets, and **c.** anti-mouse Cd90.2 and/or live-dead stain signal to exclude mouse and/or dead cells, respectively. Data processing was performed using FlowJo™ V10.1.

Table 2.6 Flow cytometry antibodies. Dilution ratios per 100 µl buffer per 500,000 cells**.** FVD* - Fixable Viability Dye (not an antibody). Na – not applicable.

Target	Conjugation	Reactivity	Dilution	Clone	Company	Reference

CD24	BUV395	Human	1:80	ML5 RUO	BD Biosciences	563818
CD75	eF660	Human	1:40	LN-1	eBioscience	50-0759-42
CD77	PE-CF594	Human	1:40	5B5	BD Biosciences	563631
Cd90.2	APC-Cy7	Mouse	1:40	30-H12	BioLegend	105328
*FVD**	eF780	na	1:33	na	eBioscience	65-0865-18
SSEA4	APC	Human/mouse	1:50	MC-813-70	R&D	FAB1435A
SUSD2	PE	Human	1:200	REA795	Miltenyi Biotec	130-111-641
SUSD2	FITC	Human	1:20	W5C5	Miltenyi Biotec	130-127-93
SUSD2	BV421	Human	1:200	W5C5	BD Biosciences	749533

An example of a full gating panel is given below (Figure 2.1).

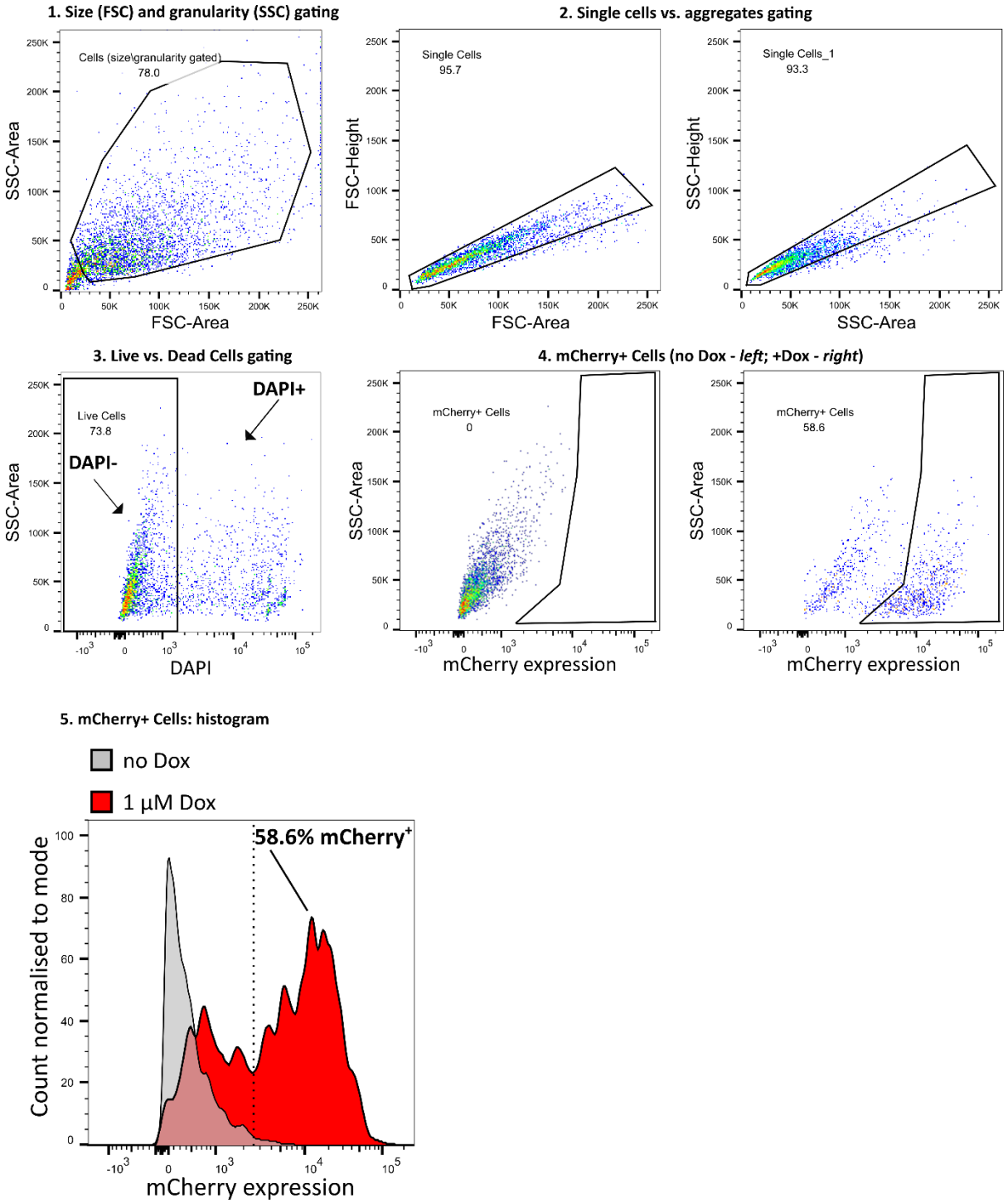


Figure 2.1 Flow cytometry gating strategy example. Experiment shown here is naive CRISPRi Gen1B +/- 1µM Doxycycline, 48 h. SSC – side scatter. FSC – forward scatter. Dox – doxycycline.

2.3 Molecular biology

2.3.1 Routine microbial culture

Escherichia coli DH5α (ThermoFisher Scientific, 18265-017) was used in this study for routine transformation in DNA cloning experiments. First, 1 µl of the DNA transformation mix was added to a 50 µl aliquot of competent bacterial cells, followed by incubation for 30 min on ice. The cells were transformed using the heat-shock method: 30 sec at 42°C, followed by 2 min on ice. 200 µl of RT SOC medium (Formedium, SOC0201) was carefully added to the transformed bacterial aliquot. The latter was then incubated at 37°C in a shacking incubator for 1 h. After that, the bacterial mix was spread onto a Luria-Bertani (LB) agar plate and left for 17 h at 37°C. For Blue-White screening, the agar would contain 40 µg/ml X-gal (Melford, MB1001). For antibiotic selection, the agar would contain 100 µg/ml ampicillin or 50 µg/ml kanamycin B (see details below). After the overnight incubation, individual colonies were picked and incubated in LB broth containing an appropriate antibiotic for 17 h at 37°C. Bacterial culture was then used for plasmid extraction.

2.3.2 Molecular cloning: TOPO-TA method

TOPO® TA Cloning® kit (ThermoFisher Scientific, 450641) was used to clone Taq polymerase-amplified PCR products into a plasmid vector for subcloning and Sanger sequencing, following manufacturer's instructions.

2.3.3 Molecular cloning: Gateway™ method

2.3.3.1 Gateway™ cloning method summary

Gateway™ cloning allows the transfer of a DNA fragment between cloning vectors while maintaining its reading frame. The method is based on recombination between *attB+attP* and *attL+attR* sequences, which flank the DNA fragment and are present in donor, entry and destination vectors. Two major steps are therefore called BP and LR reactions. In this study, Gateway cloning was used to generate TetON vectors, carrying *NANOGP1*, *NANOG* and *KLF2* cDNA.

2.3.3.2 BP reaction

A BP reaction was used to create a Gateway™ entry clone carrying an *attB*-flanked PCR product with BP Clonase™ II kit (ThermoFisher Scientific, 11789-100). The process of generating *attB*-

flanked cDNA sequences of *NANOGP1*, *NANOG* and *KLF2* is described in Section 2.1. In brief, 150 ng of the *attB*-flanked PCR product was mixed with 150 ng of the donor vector pDONR221 (ThermoFisher Scientific, 12536017) and TE buffer (8 µl total). 2 µl of BP Clonase™ II enzyme was then added to the mix, followed by vortexing and microcentrifuging. The reaction was then incubated for 1 h at 25°C and terminated by adding 1 µl of the Proteinase K and incubation 10 min at 37°C. 1 µl of the BP reaction was transformed into DH5α *E.coli.* The bacteria were then plated onto 50 µg/ml kanamycin B selective plates. Next day, individual bacterial colonies were placed into LB broth with added 50 µg/ml kanamycin B and incubated overnight in a shaking incubator. The following day, bacterial culture was pelleted, the PCR fragment-carrying entry vector was extracted and validated by Sanger sequencing.

2.3.3.3 LR reaction

150 ng of the sequence-verified entry plasmid was mixed with 150 ng of the destination vector TetON-*GFP/RFP* (kindly provided by Andras Nagy; Woltjen et al., 2009) and TE, pH 8.0 up to 8 µl. 2 µl of the LR Clonase™ II enzyme was added to the reaction mix, followed by vortexing and microcentrifuging. The reaction was then incubated for 1 h at 25°C and terminated by adding 1 µl of the Proteinase K and incubation 10 min at 37°C. 1 µl of the LR reaction was transformed into DH5α *E.coli.* The bacteria were then plated onto 100 µg/ml ampicillin selective plates. Next day, the bacterial colonies were placed into LB broth with added 100 µg/ml ampicillin and incubated overnight in a shaking incubator. The following day, the bacterial culture was pelleted, the PCR fragment-carrying destination vector was extracted and validated by Sanger sequencing.

2.3.4 Polymerase chain reaction (PCR) and genotyping primers

Polymerase chain reaction (PCR) was used to amplify various genomic and plasmid DNA fragments. PCR reactions were run in a BioRad Thermal Cycler T100™. Polymerases Q5® HiFi (NEB, M0491), LongAmp® Taq (NEB, M0323) and HotStarTaq® (Qiagen, 203203) were used according to the manufacturer's instructions.

Primer sequences used in PCR reactions, genotyping and DNA Sanger sequencing can be found in Table 2.7.

Table 2.7 Primers used for genotyping, cloning validation and Sanger sequencing. F, R – forward and reverse primer orientation.

Primer name	Assay	Primer sequence (5'-3')
M13-20-F	Sanger-Seq Genotyping	GTAAAACGACGGCCAGT
M13-R	Sanger-Seq Genotyping	CATGGTCATAGCTGTTTCC

T7-F	Sanger-Seq Genotyping	GTAATACGACTACTATAGGG
T3-R	Sanger-Seq Genotyping	ATTAACCCTCACTAAAG
P^{PH}-F	Sanger-Seq	AAATGATAACCATCTCGC
attL1-F	Sanger-Seq Genotyping	CTACAAACTCTTCCTGTTAGTTAG
attL2-R	Sanger-Seq Genotyping	ATGGCTCATAACACCCCTTG
pgRNA-CKB-F	Sanger-Seq	GAGATCCAGTTTGGTTAGTACCGGG
pgRNA-CKB-R	Sanger-Seq	ATGCATGGCGGTAATACGGTTAT
*NANOG*_2/5'-F	Sanger-Seq Genotyping	CAATGGCCTTGGTGAGACTG
*NANOG*_2/5'-R	Genotyping	GGTCCATCATTGCTCAAGAGG
*NANOG*_3/3'-F	Sanger-Seq Genotyping	ACCCCAGCCTTTACTCTTCC
*NANOG*_3/3'-R	Genotyping	AGGCTCCAACCATACTCCAC
*NANOGP1*_2/3'-F	Sanger-Seq Genotyping	CCCTGAAACACACAACTCCAG
*NANOGP1*_2/3'-R	Genotyping	GTTGTCTTTAGCAGCCAAGGT
*NANOGP1*_7/5'-F Alias: HA-F	Sanger-Seq Genotyping	TCCTGTTATTGTGGAGGCGA
*NANOGP1*_7/5'-R	Genotyping	GTGAAATGATCCCTTAACCCTCT
FLAG-R	Genotyping	TGGCTTGTCATCGTCATCCT
V5-R	Genotyping	GGAGAGGGTTAGGGATAGGC
P1-tag-seq-F	Sanger-Seq	GATCCAGCTTGTCCATAAAGCC

2.3.5 Plasmid and genomic DNA extraction

Genomic DNA extraction from hPSC pellets was performed using Monarch® Genomic DNA Purification Kit (NEB, T3010). Plasmid DNA extraction from bacterial pellets was performed using Qiagen Miniprep (27106) and Maxiprep (12162) kits. Miniprep kit was used for routine DNA cloning and Sanger sequencing. Maxiprep kit has an improved purification efficiency and lower endotoxin levels, therefore, it was used to prepare plasmids for hPSC transfection.

2.3.6 RNA extraction and cDNA synthesis

RNA extraction was performed using the RNeasy® Mini Kit (Qiagen, 74104) and QiaShredder Homogenizers (Qiagen, 79654) following the manufacturer's instructions. RNA concentration was measured using NanoDrop™ 2000 spectrophotometer. 500 ng of an RNA sample was then converted into cDNA using QuantiTect® reverse transcription kit (Qiagen, 205311) which involved **1.** a DNase-mediated genomic DNA removal step (2 min at 42°C) and **2.** cDNA synthesis using deoxyribonucleotide

triphosphates (dNTPs)-containing buffer, oligo-dT and random primer mix as well as a reverse transcriptase (1 h at 42°C/cDNA synthesis followed by 3 min at 95°C/reverse transcriptase inhibition).

2.3.7 Quantitative reverse transcription PCR (RT-qPCR) and RNA expression primers

cDNA was diluted to 50 ng/µl and used in RT-qPCR using SYBR® Green Jump Start™ Taq (Sigma-Aldrich, S4438) with 200 nM Forward and Reverse primers (Sigma-Aldrich; designed using Primer3 software (Untergasser et al., 2012)). Samples were run in technical triplicates (or duplicates) on 96-well plates on Bio-Rad CFX96 or 384-well plates on Bio-Rad CFX384. The results were analysed using the delta-delta cycle threshold method (relative quantity = $2^{-\Delta\Delta Ct}$) for which technical replicates were averaged and normalised to the expression of housekeeping genes *HMBS* and *GAPDH*. Data values represent Mean ± Standard Dation of three biological replicates, unless stated otherwise.

Forward and Reverse RNA expression primer sequences can be found in Table 2.8.

Table 2.8 RT-qPCR primer sequences.

Gene	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
DNMT3L	CTGCTCCATCTGCTGCTCC	ATCCACACACTCGAAGCAGT
DPPA3	AGACCAACAAACAAGGAG CCT	CCCATCCATTAGACACGCAGA
DUSP6	TTCCCTGAGGCCATTTCTTT	AGTGACTGAGCGGCTAATG
GAPDH	CGCTGAGTACGTCGTGGAGT	GGGCAGAGATGATGACCCTTT
GATA2	AAGGCTCGTTCCTGTTCAGAAG	CCCATTCATCTTGTGGTAGAGG
GATA3	TCACAAAATGAACGGACAGAAC	TTGTGAAGCTTGTAGTAGAGC C
GATA6	GCTCTACAGCAAGATGAACG	GACAGTTGGCACAGGACAATC
GFP	CTTCAAGATCCGCCACAACATC	GGGTGCTCAGGTAGTGGTTGT C
HMBS	AGGAGTTCAGTGCCATCATCCT	CACAGCATACATGCATTCCTCA
KLF17	CTGCAACTACGAGAACTGCG	GCAAGAATATGGCCTCTACC
KLF4	GCTGCCGAGGACCTTCTG	GCGAACGTGGAGAAAGATGG
NANOG endogenous	CCACTTTCTTGCACAGACCA	CTGGAGTTGCTGGCAGAAAG
NANOG_1	CTTGTCCCCAAAGCTTGCCT	AGGCCCACAAATCACAGGCA
NANOG_2	AAGCATCCGACTGTAAAGAATCT	ACATTTGCAAGGATGGATAGT
NANOGP1_ 1	CTTGTCCATAAAGCCTGCCT	AGGCCCACAAATCACAGGTA
NANOGP1_ 2	AAGCATCTGACTGTAAAGACTGG	ACATTTGCAAGGATGGATGGT
OCT4	GGATATACACAGGCCGATGTGG	ATGGTCGTTTGGCTGAATACCT
OLIG3	TGAGGCTGAAGATCAACGGACG	AGTTTCTGGCGAGCAGGAGTG T
OTX2	CACTTCGGGTATGGACTTGC	GGTACCGGGTCTTGGCAAA

SNAI1	AATCCAGAGTTTACCTTCCAGC	GAAGTAGAGGAGAAGGACGAAG
SOX1	AAGATGCACAACTCGGAGATCA	GCCAGCGAGTACTTGTCCTTCT
SOX17	CAAGGGCGAGTCCCGTATC	CGACTTGCCCAGCATCTTGC
SOX2	AACCAGCGCATGGACAGTTAC	GTTCATGTAGGTCTGCGAGCTG
T-BRACHYURY	TGCTGCAATCCCATGACA	CGTTGCTCACAGACCACA
TFCP2L1	TTTGTGGGACCCTGCGAAG	TGCTTAAACGTGTCAATCTGGA
ZIC2	GATGTGCGACAAGTCCTACAC	TGGACGACTCATAGCCGGA

2.3.8 Gel electrophoresis

Horizontal gel electrophoresis was used to separate nucleic acid molecules according to their size. Samples (PCR products, plasmids, DNA restriction digest products and RNA molecules) were first mixed with 5x Loading Dye containing bromophenol blue as a marker of sample migration. Samples and a DNA ladder (ThermoFisher Scientific, SM1334) were then loaded onto a 1% agarose gel in Tris/Borate/EDTA (TBE) buffer with SYBR® Safe cyanine DNA dye (ThermoFisher Scientific, S33102) or ethidium bromide (BioRad, 1610433). If the nucleic acid fragments differed in size by <10%, a 2% agarose gel was used for better fragment separation. The electric current was applied using BioRad Power Pac at 100-160V for 30-75min. Gel imaging was then performed on BioRad Gel Doc XR System. If separated fragments were to be extracted from the gel, bands of the correct size were first excised using a scalpel and then purified using Monarch® DNA Gel extraction kit (NEB, T1020).

2.3.9 Western blotting

Cell pellets were frozen on dry ice after harvesting and stored at -80°C; each pellet contained at least $1x10^6$ cells. Prior to the protein extraction, pellets were re-suspended in ice-cold Radioimmunoprecipitation assay (RIPA) buffer (25 mM Tris/HCl, 140 mM NaCl, 1% Triton X-100, 0.5% SDS, 1 mM EDTA, 1 mM PMSF, 1 mM Na_3VO_4, 1 mM NaF) supplemented with cOmplete™ protease inhibitor (Roche, 1836170). Cells were then lysed by incubating on ice for 30 minutes with vortexing every 5 min. After the incubation, lysates were centrifuged at 16,000xg for 30 min at 4°C. The supernatants were transferred to new tubes, and protein concentration was quantified using Bradford assay.

To prepare samples for gel loading, an appropriate volume of each lysate (containing 20-50 µg of the protein) was mixed with a 5x protein loading dye (5% β-mercaptoethanol, 0.02% bromophenol blue, 30% glycerol, 10% SDS, 250 mM Tris-Cl, pH 6.8) and incubated at 90^oC for 5 min. Then the samples were vortexed, briefly microcentrifuged and placed on ice. One polyacrylamide

vertical gel was composed of a stacking (top, used for the sample loading; REAGENTS) and a resolving (bottom, used for protein separation; REAGENTS) gel layers. Protein samples and a protein ladder (BioRad 1610374) were then loaded onto the gel, immersed into Tris-Glycine-SDS (TGS) buffer (BioRad, 1610772). The electric current was applied using BioRad Power Pac at 100V for 60 min. After the gel run finished, proteins were transferred onto a polyvinylidene fluoride (PVDF) membrane using iBlot2 gel transfer system. After the protein transfer, the membrane was blocked with 5% skim milk (Sigma-Aldrich, 70166) in the wash buffer TBST (Tris-buffered saline + 1% Tween20 (Sigma-Aldrich, P9416)) for 1 h at RT. A primary antibody was then added to a fresh aliquot of TBST + 5% milk, in which the blocked membrane was incubated overnight on a shaker at 4°C. Next day, the membrane was washed with TBST three times for 10 min on a shaker at RT. Then, a horseradish peroxidase (HRP) conjugated secondary antibody was added to a fresh aliquot of TBST + 5% milk, in which the washed membrane was incubated for 1 h on a shaker at RT. The membrane was then washed three times for 10 min on a shaker at RT, incubated in the ECL substrate (Cytiva, RPN2232) and imaged using X-ray films (Cytiva, 28906837). Alternatively, IRDye conjugated secondary antibodies were used for imaging membranes on LI-COR Odyssey.

Antibody details can be found in Table 2.9.

Table 2.9 Western Blotting antibodies. TFS – ThermoFisher Scientific. Na – not applicable.

Target	Conjugation	Reactivity	Host	Dilution	Clone	Company	Reference
IgG	HRP	Mouse	Goat	1:10000	Polyclonal	BioRad	1706516
IgG	HRP	Rabbit	Goat	1:10000	Polyclonal	BioRad	1706515
IgG	HRP	Goat	Rabbit	1:10000	Polyclonal	BioRad	1721034
IgG	Dylight 680	Mouse	Donkey	1:10000	Polyclonal	Cell signalling	5470
IgG	Dylight 800	Rabbit	D0nkey	1:10000	Polyclonal	Cell signalling	5151
FLAG	na		Mouse	1:10000	M-2	Sigma-Aldrich	F3165
HA	na		Mouse	1:5000	5B1D10	TFS	32-6700
NANOG	na	Human	Rabbit	1:1000	Polyclonal	Abcam	AB21624
NANOG	na	Human	Goat	1:1000	Polyclonal	R&D	AF1997
V5	na		Rabbit	1:1000	DBH8Q	Cell signalling	13202

2.3.10 RNA-sequencing: library preparation

RNA was extracted using RNeasy Mini Kit (Qiagen). Indexed libraries were made using 0.5 µg bulk RNA per sample with NEBNext® Ultra™ RNA Library Prep Kit for Illumina® with the Poly(A) mRNA Magnetic Isolation Module (NEB) and NEBNext® Multiplex Oligos for Illumina® (NEB). Agilent Bioanalyzer 2100 and KAPA Library Quantification Kit (KAPA Biosystems, KK4824) were used to identify library fragment size and concentration. Samples were then sequenced as 75 bp single-end libraries

on Illumina NextSeq 500 at the Babraham Institute Sequencing Facility, which generated 14-35 million uniquely mapped reads per library.

2.3.11 RNA-sequencing: data processing

2.3.11.1 Processing and analysing RNA-sequencing datasets generated in this thesis

Raw fastq sequencing files were quality control analysed by FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). RNA-sequencing (RNA-seq) reads were trimmed using trim galore v0.4.2 software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove the adaptor sequences. Then, using HISAT v2.0.5 (Kim et al., 2015) and guided by the gene models from Ensembl v70, trimmed reads were mapped to the human GRCh38 genome (Aken et al., 2017). Sequencing data was converted from. bam to .sam file type using Samtools (Li et al., 2009), and .sam files were imported to Seqmonk software (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). 'Raw read counts per transcript' was calculated using directional counts with help of an RNA sequencing quantitation pipeline on the Ensembl v70. DESeq2 software package (Love et al., 2014)was used to identify genes expressed differentially (cut-off of $p < 0.05$ without independent filtering and after testing correction) and to merge biological replicates. To correct for the library size and variance among counts, regularised log transformation was applied prior to data visualization. Principle component analysis (PCA) was performed using the top thousand most variable genes across the experiment. Finally, Enrichr Gene Ontology analysis was then used to identify specific biological patterns/cell types and simplify the data structure.

2.3.11.2 Analysing published RNA-sequencing libraries

The RNA-seq data of 1481 human embryo single cells from Petropoulos et al. 2016 were downloaded and categorised into the following groups: 8c, MOR, eICM, eTE, EPI, TE, PE, eUndef, Inter. Cell annotations were taken from Stirparo et al., 2018. The data was then mapped to the human GRCh38 genome using HISAT2 (v2.1.0; options --dta --sp 1000,1000), guided by known splice sites from Ensembl release 94 (Homo_sapiens.GRCh38.94.gtf) to which a custom NANOGP1 mRNA annotation had been added manually. Reads were then filtered for unique alignments (MAPQ >= 20), and log2 RPM counts for genes were calculated with SeqMonk (v1.43.1; assuming non-strand specific libraries and merging transcript isoforms). Beanplots of expression values for genes of interest were then calculated for different developmental stages using the beanplot library in R (in RStudio).

The RNA-seq data of 557 human embryo single cells from Xiang et al., 2020 were downloaded and categorised into the following groups: ICM, EPI, PrE, TrB. The data was then mapped to the human GRCh38 genome using HiSat2 (v2.1.0; options --dta --sp 1000,1000), guided by known splice sites from Ensembl release 94 (Homo_sapiens.GRCh38.94.gtf) to which a custom NANOGP1 mRNA annotation had been added manually. Reads were then filtered for unique alignments (MAPQ > 20), and log2 RPM counts for genes were calculated with SeqMonk (v1.43.1; assuming non-strand specific libraries and merging transcript isoforms). Violin plots of expression values for genes of interest were then calculated for different epiblast developmental stages using the ggplot2 package in R (in RStudio).

In Figure 5.26, Xiang et al., 2020 data was compared to the RNA-seq data produced in this study, using an R script, kindly shared by Maria Rostovskaya (publication is under review).

The RNA-seq data from Rostovskaya et al., 2019 and Collier et al., 2017 was available in the form of Annotated Probe Reports (Log2 RPM) and therefore did not require additional annotation and processing.

2.4 Bioinformatics

2.4.1 Identification of *NANOGP1* transcript variants

To identify putative *NANOGP1* transcripts, a combination of in-house generated datasets of naïve hPSCs as well as publicly available data from Theunissen et al. (2016; GEO accession GSE84382), Pastor et al. (2016; GEO accession GSE76970) and Takashima et al. (2014, ENA accession PRJEB7132) was used. All raw data was processed with Trim Galore (adapter and quality trimming, v0.6.5) and mapped to the human GRCh38 genome using HISAT2 (v2.1.0; options --dta --sp 1000,1000), guided by known splice sites from Ensembl release 94 (Homo_sapiens.GRCh38.94.gtf).

To find evidence for splicing, aligned reads were first imported into SeqMonk (v1.43.1) as introns rather than exons, which effectively uses the CIGAR operation 'N' as the start and end coordinates of putative introns. Multi-mapping reads were filtered out (MAPQ >= 20).

To identify likely exons, reads were then imported into SeqMonk as standard i.e., spliced, RNA-seq reads (MAPQ >=20). Using read counts of exonic reads and introns identified as described above, the data was inspected and manually curated further to identify potential *NANOGP1* transcript variants. Transcript candidates appearing well supported by both exonic and intronic reads were termed *NANOGP1* isoform 1-3 and taken forward for further analyses. GTF/GFF files were generated for *NANOGP1* isoforms 1-3 and included as additional annotations for both HISAT2 mapping and further analyses in SeqMonk.

To identify potential open reading frames of *NANOGP1* isoforms 1-3 their hypothetical cDNA sequences were then screened for open reading frames (ORF) using the NCBI Open Reading Frame Finder tool (https://www.ncbi.nlm.nih.gov/orffinder/). The longest ORFs, resulting in predicted proteins between 255 and 266 amino acids in length, were taken forward for multiple sequence alignments (ClustalW) and additional analyses.

2.4.2 Disambiguation of *NANOG* and *NANOGP1*

To investigate the cross-mapping of reads from the *NANOG* to the *NANOGP1* locus, and vice versa, cDNAs sequences for *NANOG* (NANOG-201, Ensembl) and *NANOGP1* (isoform 1) were used and converted them to simulated FastQ files (as 43bp (like in Petropulous et al., 2016) or 100bp single-end reads, in steps of 1bp from start to end). These *NANOG* and *NANOGP1* FastQ files were then aligned to the human GRCh38 genome (using HISAT2, v2.1.0); the amount of cross-mapping was either negligible or non-existent for unfiltered or multi-mapping filtered (MAPQ >=20) reads, respectively.

2.5 Protein Immunoprecipitation

All buffers (Table 2.10) were pre-chilled to 4°C. All centrifugation steps were performed at 4°C. NANOGP1-V5 and NANOGP1-3xFLAG hPSCs were harvested and centrifuged for 5 min at 300xg, with $5x10^6$ cells per sample. To fractionate nuclei, pellets were resuspended in ice cold Buffer A, incubated for 10 min on ice and centrifuged for 10 minutes at 2,000xg. Cell pellets were resuspended in 376 µl Buffer B, followed by 24 µl of 5 M NaCl. The resulting mix was homogenised using a Dounce on ice. Cell suspensions were kept on ice for 30 min followed by centrifugation for 20 min at 17,000xg. The supernatant was analysed by Bradford assay and stored on ice. Using a magnetic rack, Protein A and Protein G Dynabeads (Thermofisher Scientific) were washed twice with Dilution buffer. Then, 5 µg of anti-V5 and anti-FLAG antibodies (Table 2.11) were added to the Protein G and Protein A magnetic beads, respectively, which were diluted in 500 µl Dilution buffer. Tubes were kept on a rotating wheel at 4°C overnight. Next day, the beads were washed three times in the Dilution buffer. Then, 475 µg (95%) of the nuclear protein obtained in the lysis step was added to the beads. 25 µg (5%) of each protein sample were set aside as input. Immunoprecipitation samples were rotated at 4°C overnight. Next day, beads were resuspended in the Dilution buffer and washed for a total of three washes. To elute the immunoprecipitated complexes, beads were resuspended in 5x protein loading dye and boiled at 75° for 10 min. The eluate was stored at -80°C and used in Western blot assays.

Table 2.10 Protein Immunoprecipitation buffers

Buffer A	Buffer B	Dilution buffer
10 mM HEPES	5 mM HEPES	150 mM Tris-HCl pH 7.5
1.5 mM MgCl2	1.5 mM MgCl2	150 mM NaCl
10 mM KCl	0.2 mM EDTA	0.5 mM EDTA
0.5 mM DTT	0.5 mM DTT	cOmplete® EDTA-free protease inhibitor
0.05% NP40	26% Glycerol	Distilled water
250 u/ml Benzonase Nuclease (Sigma-Aldrich)	250 u/ml Benzonase Nuclease	
cOmplete® EDTA-free protease inhibitor (Roche, R1836170)	cOmplete® EDTA-free protease inhibitor	
Distilled water	Distilled water	

Table 2.11 Protein immunoprecipitation and ChiP-seq antibody details

Target	Conjugation	Reactivity	Host	Clone	Company	Reference
FLAG	na		Mouse	M-2	Sigma-Aldrich	F3165
V5	na		Rabbit	DBH8Q	Cell signalling	13202
NANOG	na	Human	Goat	Polyclonal	R&D	AF1997

2.6 Chromatin Immunoprecipitation (ChIP)

2.6.1 ChIP-sequencing: library preparation

All buffers (Table 2.12) were pre-chilled to 4°C. NANOGP1-V5 and NANOGP1-3xFLAG hPSCs were harvested and centrifuged for 5 min at 300xg at 4°C, $2x10^6$ cells per ChIP reaction. Pellets were re-suspended in PBS and fixed by incubating with 2 µM Di(N-succinimidyl) glutarate (Sigma-Aldrich, 80424) for 45 min at RT and then with 1% PFA (Agar Scientific, R1026) at a cell density of $1x10^8$ cells in 45 ml culture medium for 12.5 min at RT. 12.5 mM glycine was then added to quench the fixation, followed by a 5 min incubation at RT. the cells were then washed twice with PBS, re-suspended in 10 ml Wash Buffer 1 and incubated for 10 min at 4°C to lyse the cells. Nuclei were then separated from the other lysate components by centrifuging at 3,200xg for 5 min at 4°C, re-suspended in 10 ml Wash Buffer 2, incubated for 10 min at 4°C and centrifuged again at 3200xg for 5 min at 4°C. Then, each pellet was re-suspended in 1 ml of Nuclei Lysis Buffer per $1.2x10^7$ cells. Nuclear lysis was performed at 4°C for 30 min and was followed by sonication (30 sec on -> 45 sec off) x55 cycles per one ChIP

reaction on Bioruptor. To check fragment size distribution, a small amount of the sample was analysed by gel electrophoresis. Chromatin fragments obtained were ~400-500 bp in size. The chromatin-containing supernatant was separated from insoluble material by centrifugation at 10,000xg for 15min at 4°C, transferred to a new tube and diluted 1:10 with ChIP Dilution Buffer. 500 µl of the diluted chromatin was taken as input sample and frozen. The rest of the sample was incubated overnight with 5 µg of the antibody at 4°C. Next day, 120 µl Magnetic protein A Dynabeads (per IP) (Invitrogen, 10001D) were washed with Washing Buffer A and blocked with Bovine Serum Albumine (NEB, B9000) for 1 h at 4°C. The beads were then added to the antibody-bound chromatin, which was followed by 8 h incubation at 4°C. The beads-protein-antibody complex was washed twice with Wash Buffer A, once with Wash Buffer B, once with Wash Buffer C, and once with TE buffer (1 mM EDTA, 10 mM Tris pH 8).

Chromatin was eluted from the beads with 450 µl of Elution buffer, 11 µl of 20 mg/ml Proteinase K and 5µl of 10mg/ml RNaseA by incubating at 37°C for 2 hrs. Then, to reverse the crosslinks DNA and protein, the mix was incubated at 65°C overnight with intermittent shaking. DNA was then purified with 1x volume of AMPure XP beads (Beckman Coulter, A63880), eluted in 50 µl distilled water and quantified using the Qubit fluorometer dsDNA HS assay kit (ThermoFisher Scientific, Q32854). NEBNext Ultra II DNA library prep kit for Illumina (NEB) and NEBNext® Multiplex Oligos for Illumina® (NEB) were then used to prepare libraries according to the manufacturer's instructions). Agilent Bioanalyzer 2100 and KAPA Library Quantification Kit (KAPA Biosystems, KK4824) were used to identify library fragment size and concentration. Samples were then sequenced as 75 bp single-end libraries on Illumina NextSeq 500 at the Babraham Institute Sequencing Facility, which generated 20-43 million uniquely mapped reads per library.

ChIP-seq antibody details are provided in Table 2.11.

Table 2.12 ChIP-seq buffers

Wash Buffer 1	Wash Buffer 2	Nuclei Lysis Buffer	Dilution Buffer	Washing Buffer A	Washing Buffer B	Washing Buffer C	Elution Buffer
10 mM EDTA	1 mM EDTA	5 mM EDTA	5 mM EDTA	1 mM EDTA	1 mM EDTA	1 mM EDTA	1% SDS
0.5 mM EGTA	0.5 mM EGTA	150 mM NaCl	150 mM NaCl	150 mM NaCl	500 mM NaCl	1% Igepal CA-630	0.1 M $NaHCO_3$
10 mM Hepes pH 7.5	10 mM Hepes pH 7.5	1% SDS	0.1% SDS	1% NP40	1% NP40	250 mM LiCl	cOmplete® EDTA-free protease inhibitor

0.75% Triton X-100	200 mM NaCl	0.5% Sodium deoxycholate	0.5% Sodium deoxycholate	0.1% SDS	0.1% SDS	0.5% Sodium deoxycholate	Distilled water
cOmplete® EDTA-free protease inhibitor (Roche, R1836170)	cOmplete® EDTA-free protease inhibitor	25 mM Tris pH 7.5	25 mM Tris pH 7.5	0.5% Sodium deoxycholate	0.5% Sodium deoxycholate	50 mM Tris pH 8	
Distilled water	Distilled water	0.1% Triton X-100	1% Triton X-100	50 mM Tris pH 8	50 mM Tris pH 8.0	cOmplete® EDTA-free protease inhibitor	
		cOmplete® EDTA-free protease inhibitor	cOmplete® EDTA-free protease inhibitor	cOmplete® EDTA-free protease inhibitor	cOmplete® EDTA-free protease inhibitor	Distilled water	
		Distilled water	Distilled water	Distilled water	Distilled water		

2.6.2 ChIP-seq data processing and analysis

ChIP-seq reads (NANOGP1, NANOG from this study and NRSF/REST from ENCODE (GSM803365) were trimmed using Trim Galore v0.6.6 (Cutadapt v2.3) software and mapped to human genome GRCh38 with Bowtie2 (v2.4.2) (Langmead and Salzberg, 2012). Subsequent analysis was performed using SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/) software and R studio/R (v2021.09.0/v4.1.1). Quantitation values are represented as log2 RPM. NANOGP1 peaks were called on individual replicates using a SeqMonk implementation of MACS (Zhang et al., 2008) using parameters sonicated fragment size =300 and $p<10^{-9}$, coverage outliers were excluded. The intersection of individual replicate peaks was used as NANOGP1 peak annotation. NANOGP1 peaks overlapping NANOG peaks (Chovanec et al., 2021) were designated 'shared'.

ChromHMM states for naïve hESCs were taken from Chovanec et al., 2021. In order to assign only one ChromHMM state per peak, the peak centre location was used. Control regions were 1000 randomly selected 900 bp windows where 900 bp is an approximate average peak width.

De novo motif discovery and motif enrichment for *NANOGP1*, *NANOG* and shared peak regions (+/-100 bp around the peak centre) were performed using Homer (Duttke et al., findMotifsGenome.pl) and the human genome version hg38. Homer's scan MotifGenomeWide.pl was used to find the REST motif (REST/MA0138.2) in instances across the whole human genome.

2.7 Recombinant protein synthesis

2.7.1 Recombinant protein synthesis: cloning

NANOGP1-HA, NANOGP1-FLAG-His, NANOG-HA and NANOG-FLAG-His recombinant proteins were made using a Baculovirus expression system (ThermoFisher Scientific, 11827-011/11804-010) in *Sf9* insect cells. *NANOGP1* (isoform 1) and *NANOG* CDSs were generated using gBlocks, inserted into pCAG-IRES-Puro^R backbone vector (Section 2.1.7; Niwa et al., 1991) and then PCR-amplified from plasmid DNA. Here, PCR primers were different from those used for generating TetON vectors (Section 2.1.7). The primers here were designed to attach *SalI* and *XbaI* restriction sites to the 5' and 3' ends of the PCR products, respectively. Additionally, they amplified CDSs without their respective stop codons.

Table 2.13 Primer sequences used in the molecular cloning for recombinant protein synthesis. Enzyme restriction sites are in bold. 2-nt and 4-nt overhangs were added for the increased cutting efficiency. F – forward primer. R – reverse primer.

Primer name	Primer sequence (5'-3')
NANOG-SalI-F	ACGC**gtcgac**ATGAGTGTGGATCCAGCTTG
NANOG/*NANOGP1-XbaI*-R	GC**tctaga**CACGTCTTCAGGTTGCATGT
NANOGP1-SalI-F	ACGC**gtcgac**ATGTCTTCTGCTGAGATGCC

The PCR products were digested with *SalI* and *XbaI* enzymes (ThermoFisher Scientific, FD0684, FD0644) to be ligated into two different pBluescript vectors, carrying in-frame HA and FLAG-His tags. pBluescript vectors were also digested with *XbaI* and *SalI* enzymes and resolved on an electrophoresis gel. Bands of the correct size were excised and used in CDS+pBluescript ligation reactions. A 20 µl ligation reaction contained a digested pBluescript vector fragment, a digested *NANOG* or *NANOGP1* CDS, ATP-containing T4 ligase buffer (NEB, B0202) and T4 ligase (NEB, M0202). The reaction was incubated for 10 min at RT and terminated by 10 min incubation at 65°C. 1 µl of the reaction was transformed into DH5α *E.coli* which was selected on 40 µg/ml X-gal LB plates. Next day, white bacterial colonies were placed into LB culture with added 100 µg/ml ampicillin. The following day, bacteria were pelleted, plasmid DNA was extracted and validated by Sanger sequencing with T7-F primer. Plasmids produced in here were pBluescript-*NANOG*-HA, pBluescript-*NANOGP1*-HA, pBluescript-*NANOG*-FLAG-

His, pBluescript-*NANOGP1*-FLAG-His. Next, tagged CDS were cloned into pENTR3c vector containing *attL* sequences. First, pENTR3c, pBluescript-*NANOG*-HA, pBluescript-*NANOGP1*-HA, pBluescript-*NANOG*-FLAG-His and pBluescript-*NANOGP1*-FLAG-His were digested with *NotI* (ThermoFisher Scientific, FD0595) and *SalI* enzymes and run on the gel. The correct size bands were extracted and ligated (see the ligation reaction above). The reactions were transformed into DH5α *E.coli* which was then selected on 50 µg/ml kanamycin LB plates. Next day, bacterial colonies were placed in LB culture with 50 µg/ml kanamycin. Next day, bacterial cultures were pelleted, plasmids were extracted and validated by Sanger sequencing with attL1 and attL2 primers. Plasmids produced here were pENTR3C-attL1-*NANOG*-HA-attL2, pENTR3C-attL1-*NANOGP1*-HA-attL2, pENTR3C-attL1-*NANOG*-FLAG-His-attL2, pENTR3C-attL1-*NANOGP1*-FLAG-His-attL2. These plasmids were then recombined with a pDEST8 vector in an LR Gateway reaction and transformed in DH5α *E.coli* cells. Plasmids were extracted and validated by Sanger sequencing with the polyhedrin promotor-specific primer (P^{PH}-F). The verified plasmids were then transformed into DH10Bac *E.coli* cells (ThermoFisher Scientific, 10361012), which contain bacmid DNA with a mini-*att*Tn7 transposition target site and a helper plasmid, encoding transposition proteins. Bacterial colonies containing the recombinant sequences were white, as the transposition disrupts the *lacZ*α gene, and resistant to 50 µg/ml kanamycin (DH10Bac bacmid), 7 µg/ml gentamicin (pDEST8) and 10 µg/ml tetracycline (helper plasmid). Recombinant bacmid DNA mini-preps were prepared from the selected clones and were used to transfect insect cells.

Primer sequences can be found in Table 2.7.

2.7.2 Recombinant protein synthesis: insect cell culture

Sf9 insect cells were maintained in Hink's TNM-FH Insect Medium (Merck, 51942C) supplemented with 10% FBS and 1% Chemically Defined Lipid Concentrate (Gibco, 11905031; CDLC) as a suspension culture in vented 100 ml flasks at 27°C in a non-humidified shaking incubator (120 RPM) in the dark. When suspension culture reached $2x10^6$ cells/ml density, it was diluted to 0.5-1 $x10^6$ cells/ml. For transfection, log-phase *Sf9* cell culture was plated as a monolayer into a 6-well plate, $1x10^6$ cells/well in 2 ml TNM-FH + 10% FBS + 1% CDLC and incubated for 2 h. Before the transfection, medium was changed to TNM-FH without FBS/CDLC. For each transfection, 6 µl Cellfectin™ II (ThermoFisher Scientific, 10362100) and 5 µl recombinant bacmid DNA were pre-incubated at room temperature for 30 min in 200 µl TNM-FH and then added to a well dropwise. Ttransfection reactions were incubated for 5 h, followed by medium change to 2 ml TNM-FH + 10% FBS + 1% CDLC and a five-day incubation period without feeding. On day 5, when signs of viral infection were visible, 1 ml of the transfected cell culture (P1 stock) was transferred into 50 ml of fresh suspension culture at $1x10^6$ cells/ml and in TNM-FH + 10% FBS + 1% CDLC. Suspension culture was incubated for 5 days, then the cells were

centrifuged (5 min at 500xg). The supernatant was then filtered and 50 ml of fresh suspension culture at $1x10^6$ cells/ml and in TNM-FH+10% FBS + 1% CDLC was inoculated with 1 ml P2 stock. Suspension culture was incubated for 5 days, then, the cells were centrifuged (5 min at 500g), supernatant was filtered. Then, 2 L of fresh suspension culture at $1x10^6$ cells/ml and in TNM-FH+10% FBS + 1% CDLC was inoculated with 40 ml of P3. At this stage, viruses were combined and added to WT suspension culture in the following combinations (2 L WT culture per inoculation condition):

1. *NANOG*-HA + *NANOG*-FLAG-His
2. *NANOGP1*-HA + *NANOGP1*-FLAG-His
3. *NANOG*-HA + *NANOGP1*-FLAG-His
4. *NANOG*-FLAG-His + *NANOGP1*-HA

After 2.5 days of incubation, cells were centrifuged, pellets were snap-frozen in liquid nitrogen. Protein was extracted using ice-cold lysis buffer (50 mM HEPES (pH 7.5), 500 mM NaCl, 0.5 mM TCEP, 5% (v/v) Glycerol) and tissue pulveriser. Proteins were purified using an AKTA system.

2.8 Immunofluorescent staining

hPSCs were fixed in 12-well cell culture plates for 15 min at 4°C in 4% PFA (Agar Scientific, R1026) in PBS, washed once with PBS and permeabilised with 0.4% Triton X-100 (Sigma-Aldrich, T8787) in PBS for 10 min at RT. Nonspecific antibody binding was minimised by incubating cells with 3% BSA (Sigma-Aldrich, A7906) + 0.1% Triton X-100/PBS for 1 h at RT. Then, the cells were incubated with the appropriate primary antibody in 3% BSA + 0.1% Triton X-100/PBS overnight at 4°C, before being washed four times with 0.1% Triton X-100/PBS and incubated with the appropriate secondary antibodies in 3% BSA + 0.1% Triton X-100/PBS for 1 h at RT in the dark. Finally, the cells were washed three times in 0.1% Triton X-100/PBS (for nuclei staining 1 µg/mL DAPI (Tocris, 5748) was added to the first wash) and two times in PBS. Wells were then filled with PBS, plates were sealed and stored at 4°C.

Immunofluorescent staining antibody details are provided in Table 2.14.

Table 2.14 Immunofluorescent staining antibody details. CST - Cell Signalling Technology. SC – Santa Cruz. TFS - ThermoFisher Scientific. Na – not applicable.

Target	Conjugate	Reactivity	Host	Dilution	Clone	Company	Reference
IgG	AlexaFluor 555	Goat	Donkey	1:1000	Polyclonal	TFS	A21432
IgG	AlexaFluor 647	Mouse	Donkey	1:1000	Polyclonal	TFS	A31571
IgG	AlexaFluor 555	Rabbit	Donkey	1:1000	Polyclonal	TFS	A31572
NANOG	na	Human	Goat	1:200	Polyclonal	R&D	AF1997

OCT4	na	Human/mouse	Mouse	1:300	C-10	SC	SC5279
V5	na	na	Rabbit	1:150	DBH8Q	CST	13202

2.9 Microscopy

Imaging of live hPSC cultures was performed on ZOE fluorescent cell imager. Imaging of immunostaining experiments was performed on Nikon Live Cell Imager. Imaging of alkaline phosphatase imaging was performed on Zeiss Palm MicroBeam; image stitching was done using the microscope internal software. All images were processed and analysed using Fiji/ImageJ software (Schindelin et al., 2012).

2.10 Evolutionary genetics

This section has been taken and adapted with permission from the MSc thesis of Gökberk Alagöz; it described experiments performed by Gökberk Alagöz (see Section 3.2).

To investigate genomic structure of the *NANOG/NANOGP1* locus throughout evolution, most recent assemblies of nine primate species (Table 2.15 Primate genome assemblies used in the evolutionary genetics' assays.) were analysed. Approximate genomic coordinates of *NANOG* and *NANOGP1*, if present, were identified using BLAST ((Basic Local Alignment Search Tool (BLAST))) and Needle (Madeira et al., 2019) pair-wise sequence alignment tools. Within each assembly, a ~250 kilobase genomic region including *NANOG*, *NANOGP1* and their surrounding genes, was extracted using a Python script. DNA and its corresponding amino acid sequences of *NANOG* and *NANOGP1* were aligned using MEGA (Tamura et al., 2007)and ClustalW (Thompson et al., 1994) tools. Codeml and codonml PAML 4.8a programs were run for the phylogenetic analysis of amino acid sequences (Yang and Nielsen, 2000). Dotter (Barson and Griffiths, 2016) and Miropeats (Parsons, 1995) tools were used for visualising *NANOG*/*NANOGP1* duplication site, detecting its boundaries and conserved regions.

For GC content calculation, enhancer regions were first extracted from the human genome assembly. GC content ratios were then calculated by dividing the sum of G and C nucleotide counts (G+C) to the total nucleotide count (G+C+T+A) at a genomic region, using 30 base-pair sliding-window approach. GC percentages were plotted against genomic coordinates using matplotlib in Python.

Table 2.15 Primate genome assemblies used in the evolutionary genetics' assays.

Species	Assembly	First release date
Human	hg38	2013
Chimpanzee	panTro6	2018

Bonobo	panPan2	2015
Gorilla	gorGor5	2016
Orangutan	ponAbe3	2018
Gibbon[1]	*nomLeu3*	*2012*
Crab-eating macaque	macFas5	2013
Rhesus macaque	rheMac8	2015
Marmoset	calJac3	2009

1 - Gibbon nomLeu3 assembly was found to be not suitable for investigating *NANOG* region due to having large gaps. To resolve this, unpublished raw gibbon genome data, kindly provided by Prof. Evan Eichler research group (University of Washington), was analysed. To visualise *NANOG*-containing locus, human *NANOG* and *NANOGP1* sequences were mapped to gibbon contigs using Minimap2 (Li, 2018)**.**

3 Human pseudogene *NANOGP1*: characterisation of its evolutionary conservation and expression pattern in human pluripotency

3.1 Introduction

3.1.1 Background

Segmental and tandem duplications are thought to be one of the main drivers of species evolution (Section 1.4). Marques-Bonet and colleagues demonstrated a positive correlation between evolutionary development of mammalian species and the increasing percentage of segmental duplication in their genomes (Marques-Bonet et al., 2009a; Marques-Bonet et al., 2009b). Moreover, a sudden burst of gene duplication events occurred following the divergence of apes from Old World monkeys. According to their work, duplication rates peaked in the arguably most evolutionarily advanced primates, the Great Apes, and, more specifically, in humans.

Many of the duplicated genes eventually lose their function and either disappear completely, or become pseudogenes (Section 1.4). Interestingly, despite the commonly accepted assumption that pseudogenes are unfunctional, pseudogenisation has been linked with the evolutionary development of species-specific mechanisms (Niimura and Nei, 2007); discussed in more details in Chapter 6). Currently there is a lack of understanding, at the molecular level, how gene duplications, especially pseudogenes, contribute to development of species-specific mechanisms. The majority of duplications and pseudogenisation events therefore require focused attention from computational and experimental scientists to assess their potential significance and test their functional contribution.

Gene duplications that are active in early development could allow different species to evolve different reproductive and developmental strategies, therefore also requiring particular attention from researchers (Section 1.4). Moreover, as preimplantation epiblast is used for deriving hPSCs, investigating the role of gene duplicates in early development is also important for understanding how pluripotency (and hPSCs) could be regulated differently between species.

The highly duplicated human pluripotency factor *NANOG* serves as a tractable model for gene duplication in human embryo development, as described in Sections 1.3 and 1.4. Human *NANOG* has an unusually high number of copies: ten processed pseudogenes and one unprocessed and highly conserved tandem duplicate, *NANOGP1,* currently annotated as a pseudogene (Booth and Holland, 2004). Four studies have previously investigated this duplication, including the demonstration that *NANOGP1* mRNA is detectable in primed hPSCs (Booth and Holland, 2004; Eberle et al., 2010; Fairbanks and Maughan, 2006; Hart et al., 2004; Sections 1.4.4, 1.4.5). However, those prior studies were mostly computational and did not explore whether *NANOGP1* could have a function in human pluripotent cells or states. Additionally, those studies had some important discrepancies, such as an inability to agree on the structure of *NANOGP1* transcript or whether endogenous *NANOGP1* is

capable of making a protein. Because high expression levels of *NANOG* are crucial for maintaining naïve and primed pluripotency in hPSCs (Section 1.3), investigating a potential role of *NANOGP1* in these cells is important to understand how gene duplication could influence human embryo development, as is exploration of the role these processes might play in species-specific developmental and stem cell programmes.

Until this study, very little was known about the evolutionary history of the *NANOG/NANOGP1* tandem duplication and whether *NANOGP1* was consistently expressed in human pluripotent compartments. Moreover, it was not understood what structure *NANOGP1* mRNA had in the hPSC context and whether it had the potential to encode a full-length protein.

In this chapter, I test the hypothesis that *NANOGP1*, as an unprocessed duplicate of an important pluripotency regulator, is expressed in human pluripotency. I also address such key questions as the evolutionary conservation of *NANOGP1* sequence, the structure of *NANOGP1* mRNA, and whether it has functional upstream regulatory regions. Finally, I briefly discuss pseudogenes of other pluripotency factors and their expression in hPSC, highlighting the potential importance of pseudogenes in early development.

Sections of this chapter contain results that were obtained in collaboration with scientists from the Department of Genetics, University of Cambridge, and the Babraham Bioinformatics Group. All of these data have been essential for establishing the wet-lab project that I focused on during my Ph.D. project. To present the full scope of the project and demonstrate how it was initiated, I therefore describe here the results provided by our collaborators, with their permission.

Data presented in Figures 3.2-3.7, and 3.16*, were produced by Gökberk Alagöz under the supervision of Dr. Aylwyn Scally (Department of Genetics). Data generated by Gökberk Alagöz are described in his M.Sc. Thesis in Evolution, Ecology and Systematics, which was submitted to the University of Cambridge on 26.08.2019.

Data presented in Figures 3.9-3.10, and 3.13, were produced by Dr. Felix Krueger (Babraham Bioinformatics Group).

All figures in this chapter were adapted and/or made by myself, and Figures 3.1, 3.8, 3.11-12, 3.14-3.15, 3.16*, 3.17-3.20 are entirely my own work.

3.1.2 Aims

1. Investigate the *NANOG/NANOGP1* duplication locus using evolutionary genetics

1.1 Study the architectural structure and boundaries of the *NANOG/NANOGP1* locus in the human genome

1.2 Compare the degree of *NANOGP1* conservation between primate species

1.3 Convey protein conservation analysis between *NANOG* and *NANOGP1* primate orthologs

2. Investigate and compare *NANOG* and *NANOGP1* mRNA expression patterns in hPSCs and developing embryos
3. Examine *NANOGP1* mRNA and protein structure and conservation
4. Examine and compare predicted *NANOGP1* and *NANOG* regulatory regions in hPSCs
5. Analyse other highly expressed pseudogenes in naïve hPSCs

3.2 Results

3.2.1 *NANOG/NANOGP1* duplication locus: choice of nomenclature

In the human genome, *NANOG* and its tandem duplicate *NANOGP1* are located on the short arm of chromosome 12. In addition to them, this region contains another pair of duplicated genes, *SLC2A3* and its ancestor *SLC2A14* (Figure 3.1). In this thesis, I refer to the duplication locus as '*NANOG/NANOGP1'*, excluding the names *SCL2A14* and *SCL2A3* for brevity.

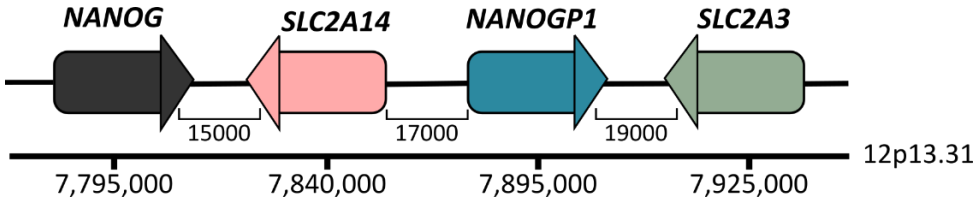


Figure 3.1 Diagram showing *NANOG/NANOGP1* tandem duplication locus. Orientation of *NANOG, SLC2A14, SLC2A3* genes and *NANOGP1* pseudogene is indicated by arrows. Distance between the genes/pseudogene is indicated with brackets, bp. Scale at the bottom of the diagram represents position of the locus within the chromosome, bp. 12p13.31 – region on human chromosome 12. Genome assembly GRCh38 used.

3.2.2 Characterising conservation of the *NANOG/NANOGP1* duplication locus

To study the architectural structure of human *NANOG/NANOGP1* duplication, as well as its evolutionary history among primates, a series of genomic analyses were performed.

First, to visualise the duplication locus and identify its boundaries, the sequence within a 250 kb region that included *NANOG, NANOGP1, SLC2A14, SLC2A3* and *NANOGNB*, as well as their flanking regions, was self-aligned. *NANOGNB* is located upstream of *NANOG* and is outside of the duplication

locus, although *NANOGNB* was recently hypothesised to be a highly diverged copy of *NANOG* (Dunwell and Holland, 2017). Here, *NANOGNB* was included in the analysis to see if we could detect conservation between the two genes.

The alignment was performed using Dotter software and visualised as a dot plot matrix. As a result, three clusters of duplication were identified: 1) *NANOG-NANOGP1*, 2) *SLC2A14-SCL2A3*, and 3) an intergenic region ~60 kb downstream of *SLC2A3*, which was composed of three smaller areas of duplication (Figure 3.2). No noticeable similarly was detected in other areas, including between *NANOGNB* and *NANOG*.

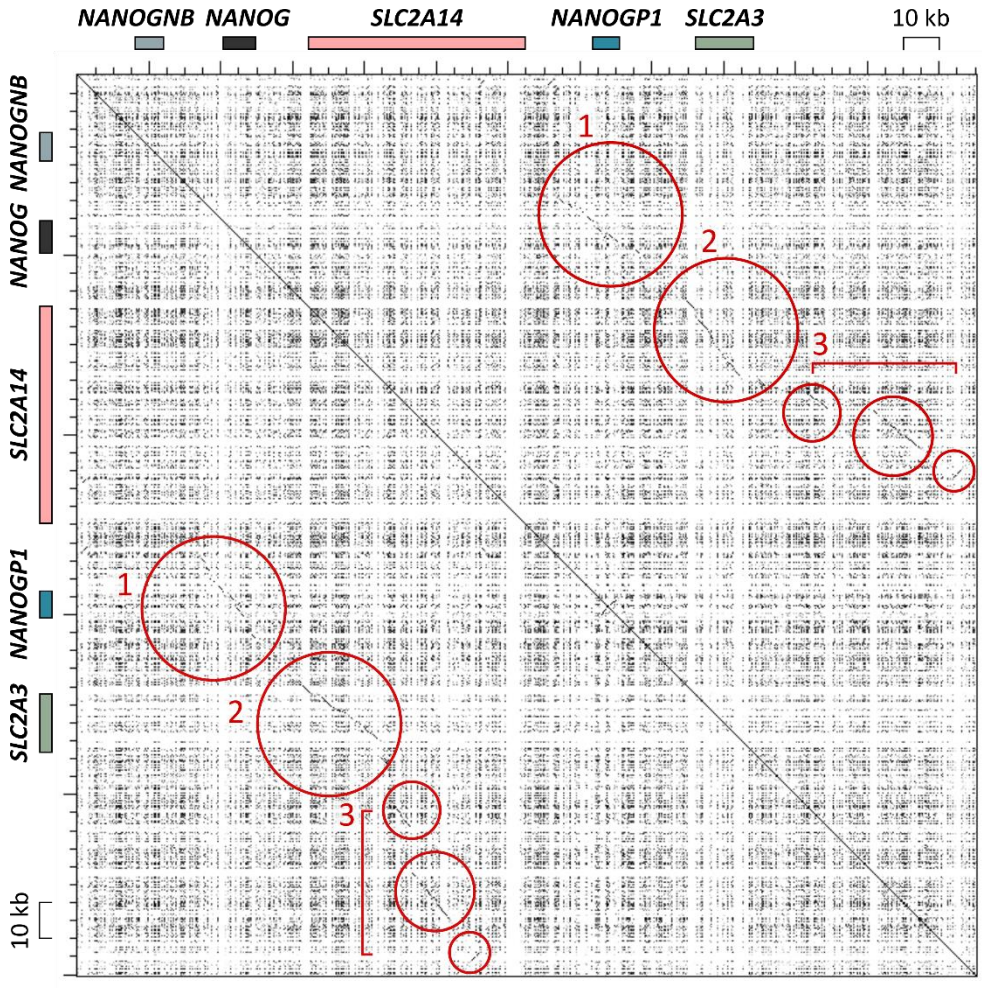


Figure 3.2 Dot plot showing self-alignment of a 250 kb region, containing *NANOG,* its tandem duplicate *NANOGP1* as well as another duplicated pair, *SLC2A14* and *SLC2A3*. *NANOG*, *NANOGP1, SLC2A14*, *SLC2A3* and *NANOGNB* are depicted as rectangles along the x and y axis. Individual dots represent matching base pairs between the two aligned sequences. Red circles indicate three areas or of conservation (1, 2, 3) between the ancestral and duplicated regions. All the other dots visible on the plot as a grid represent background. Top right and bottom left plots, separated by the diagonal line, are mirror images of each other. Scale, 5 kb.
Adapted with permission from MSc thesis of Gökberk Alagöz.

From the self-alignment analysis, the tandem duplication event at this locus was concluded to involve copying and inserting an ~80 kb region, which contained *NANOG* and *SLC2A14* genes,

downstream (3') of its original location. *NANOGP1* and *NANOG* were also observed to have high level of conservation not only within their exons, but upstream of their CDSs as well, as shown in the Miropeats plot (Figure 3.3).

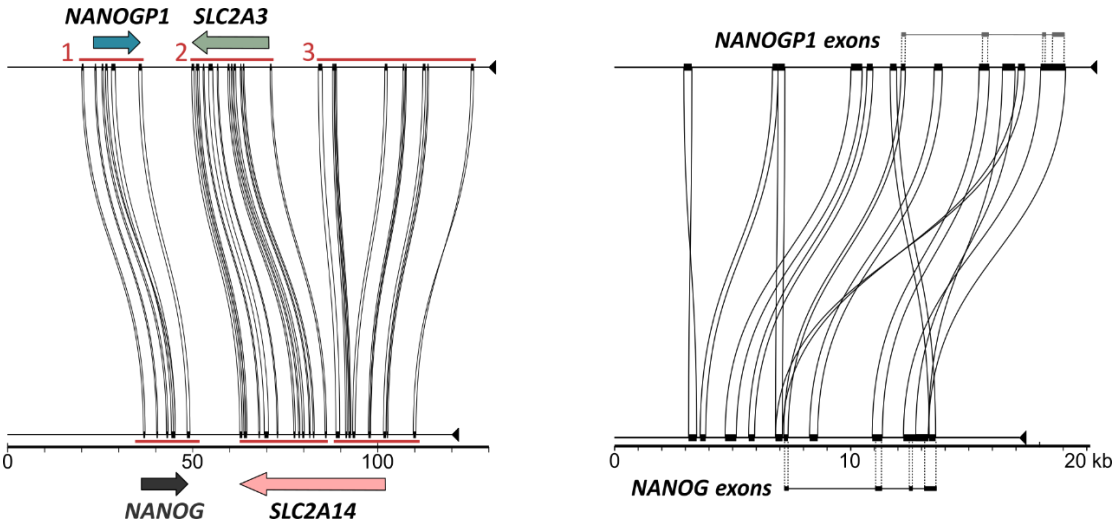


Figure 3.3 Miropeats plots showing sequence similarity between *NANOG-SLC2A14* and its tandem duplication *NANOGP1-SLC2A3* (left) and between *NANOG* and *NANOGP1* exons, as well as their upstream regions (right). Gene/pseudogene orientation is indicated by arrows. The three regions of conservation are indicated by red lines and numbers 1, 2, 3 Chromosome distance, kb.
Adapted with permission from MSc thesis of Gökberk Alagöz.

3.2.3 Characterising evolutionary origin of the *NANOG/NANOGP1* duplication locus

To study the evolutionary history of *NANOG/NANOGP1* duplication, gene localisation data was obtained from genome assemblies for the following species: Hominoids/Apes (human, chimpanzee, gorilla, orangutan, gibbon), Old World monkeys (green monkey, rhesus macaque, crab-eating monkey, baboon, golden snub-nosed monkey), New World monkeys (marmoset) and Prosimians (mouse lemur). *NANOGP1* was originally absent from most non-human genome assemblies and therefore had to be located and annotated manually using human *NANOGP1* as a reference.

As a result, *NANOGP1* sequences, full or truncated, as well as *SLC2A14*, were found in almost all Hominoid and Old World monkey genomes, but not in the mouse lemur genome (Figure 3.4)

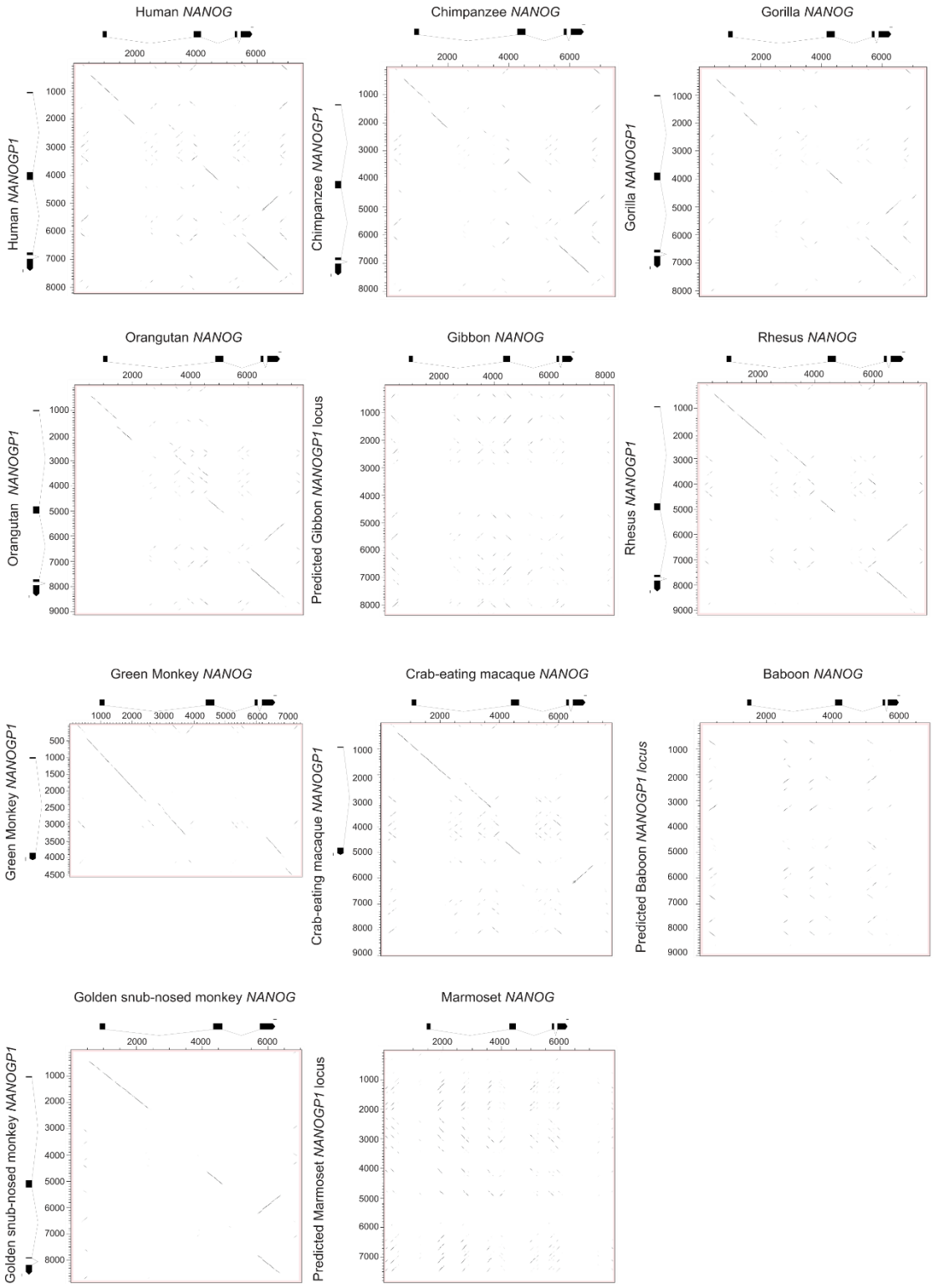


Figure 3.4 Dot plots showing alignment of primate *NANOG* orthologs to their corresponding *NANOGP1* duplicates. Individual dots represent matching base pairs between the two aligned sequences. In areas of sequence conservation individual dots form diagonal lines. Gene/pseudogene structure is shown as black rectangles (exons) and lines (introns). Scale, bp.
Adapted with permission from MSc thesis of Gökberk Alagöz.

This alignment analysis led to the conclusion that the duplication event had occurred at least ~40 million years ago (mya), before the split between the Hominoid-Old World monkey branches (25-32 mya) (Figure 3.5). It was not clear, however, whether it occurred after the Old World-New World monkey taxa had separated (50 mya), or prior to that event. Indeed, while *NANOGP1* sequence was not found in the marmoset genome, the *SLC2A14* duplicate *SLC2A3* was located there. This led to proposing two alternative duplication timing scenarios, namely, prior or after the Old and New World Monkey branch split.

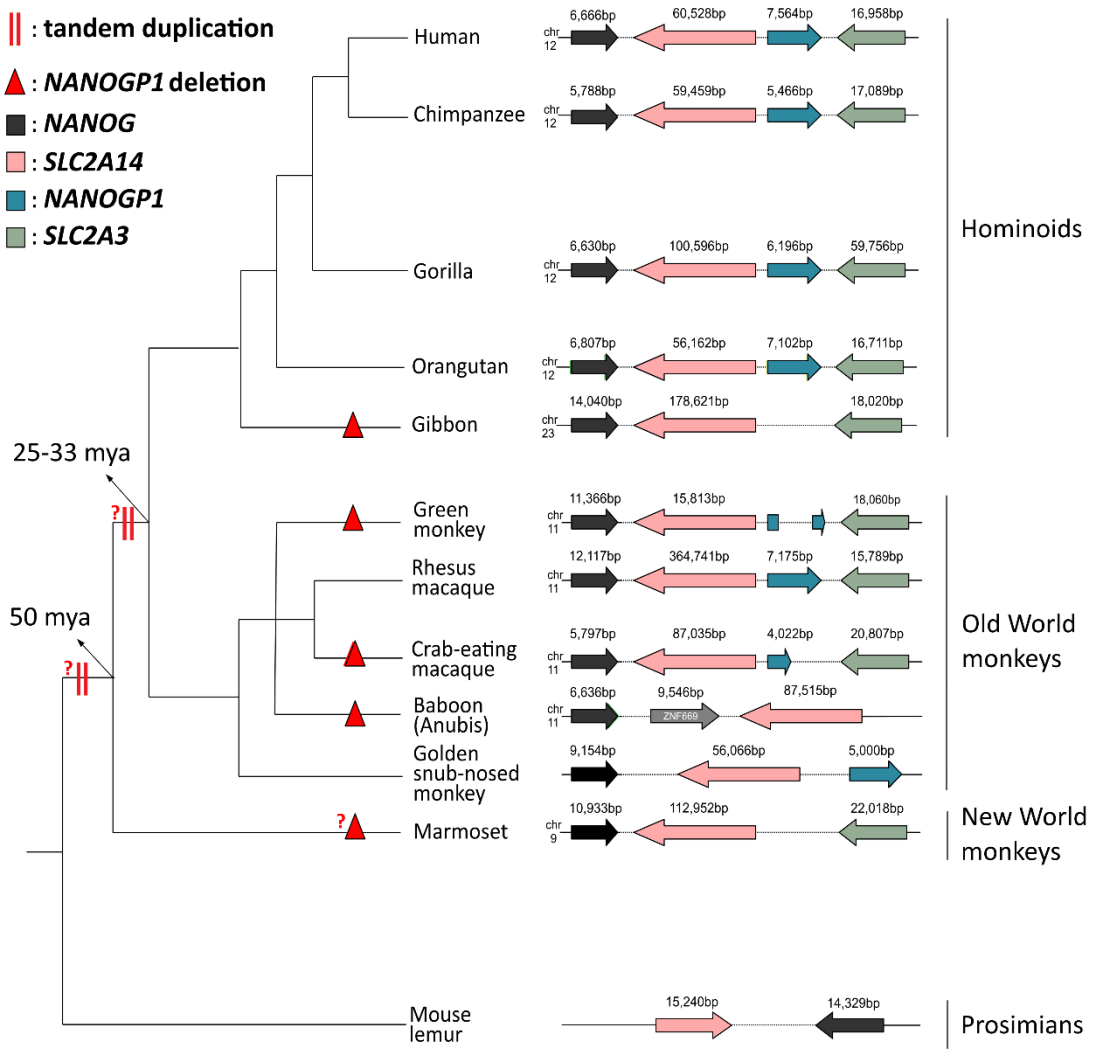


Figure 3.5 Summary diagram showing conservation of *NANOG/NANOGP1* tandem duplication locus among analysed species. Genes/pseudogenes are colour-coded - see legend in the top left corner. Predicted duplication dates are indicated with two red vertical lines. Predicted *NANOGP1* deletions are indicated with red triangles. Gene/pseudogene orientation is indicated by arrows. Chr – chromosome. Gene/pseudogene length, bp.
Adapted with permission from MSc thesis of Gökberk Alagöz.

In the hominoid branch, *NANOGP1* sequence was found in all analysed Great Ape species: human, chimpanzee, gorilla and orangutan. In that branch, *NANOGP1* had a noticeably high degree of sequence conservation with *NANOG*, whereas in gibbon, which belongs to Lesser Apes, the whole structure of the locus was altered and *NANOGP1* was absent. Interestingly, in contrast to Great Apes, the duplication in Old World monkeys had been noticeably affected by mutagenesis. For instance, *NANOGP1* sequence in green monkey and crab-eating macaque possessed large deletions, while in baboon *NANOGP1* was completely absent. In the two remaining Old World monkeys, rhesus macaque and golden snub-nosed monkey, *NANOGP1* seemed to be mostly intact (with the exception of several fine-scale mutations, discussed further in the text). Therefore, *NANOGP1* was concluded to be highly conserved and resemble *NANOG* in all analysed Great Ape genomes, while only two Old World Monkey species demonstrated similar conservation. Moreover, the Great Ape branch was unique not only in its level of sequence conservation within *NANOGP1* but also within the tandem duplication locus itself, which included gene order and orientation. Therefore, it was concluded that the high level of conservation of the duplicated locus in the Great Ape branch could suggest potential functional importance of *NANOGP1* in that group of species.

3.2.4 Characterising conservation of the protein-CDS within the *NANOG*/*NANOGP1* duplication locus

To study whether the protein-CDS was also conserved among the *NANOG* and *NANOGP1* primate orthologs, amino acid alignment plots were generated for the seven following species: human, chimpanzee, gorilla, orangutan, gibbon, rhesus macaque and crab-eating macaque (Figure 3.6). Crab-eating macaque was only included in the NANOG multiple sequence alignment since its NANOGP1 sequence is severely truncated. This method did not involve identification of potential ORFs/splicing patterns, but analysed the likelihood of conservation for each codon using Phylogenetic Analysis by Maximum Likelihood (PAML) software. This analysis showed that in all species except gorilla, amino acid sequences were largely conserved.

Interestingly, in gorilla NANOG and NANOGP1 predicted protein sequences, a single nucleotide deletion was identified in exon 4. This deletion would cause frameshifts towards the end of the amino acid sequence in both proteins. Due to these frameshifts, both NANOG and NANOGP1 obtained an early stop codon, V276STOP. The predicted NANOGP1 frameshift is located downstream of the predicted early stop codon, which would lead to losing just over 25 amino acids in the C-terminal end of NANOGP1. In NANOG, however, the frameshift occurs upstream of the early stop codon, leading to a change of over 80 amino acids of the C-terminal domain, in addition to the deletion of ~30 amino acids downstream of the early stop codon. Collectively, the frameshift and early stop codon

would cause the loss of one third of the conserved NANOG protein sequence in gorilla. Currently, this gorilla-specific NANOG frameshift, as well as its premature stop codon, are not annotated in publicly available protein repositories.

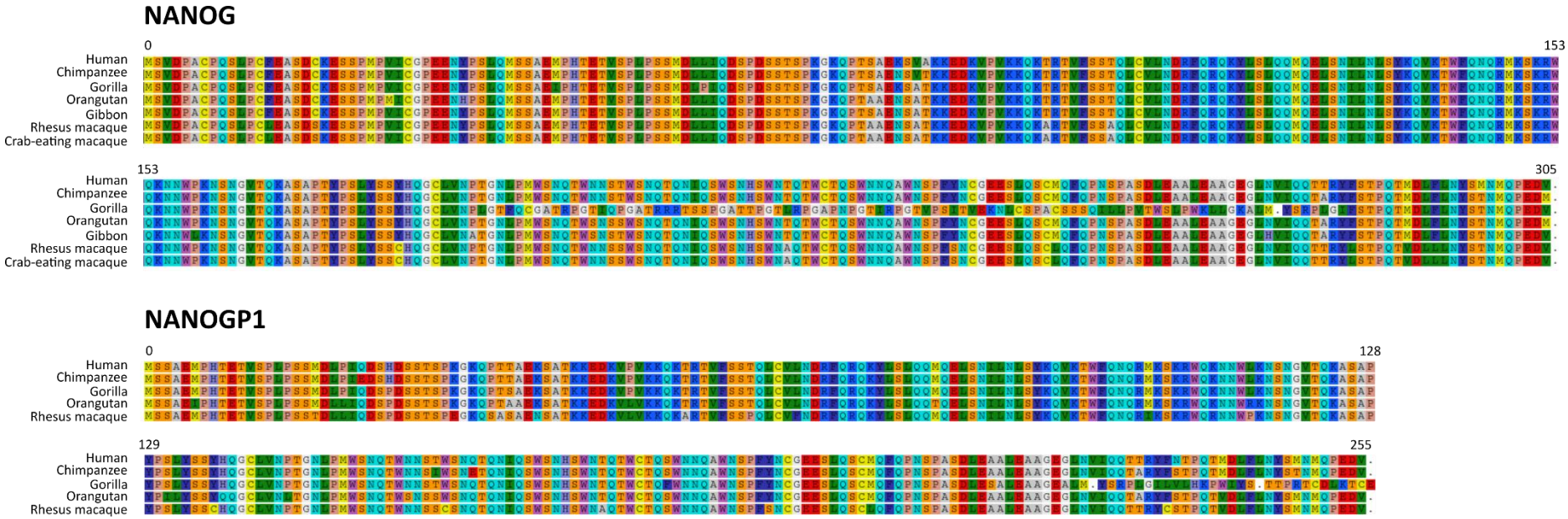


Figure 3.6 Amino acid sequence alignment of primate NANOG and NANOGP1 orthologues. Amino acid residues are indicated with the universally accepted single-letter code. Colour coding indicates different types of amino acids, according to their biochemical properties. Stop codon – dot. Names of the analysed primate genomes are listed on the left from each alignment matrix. The alignment figures for both NANOG and NANOGP1 are split into two parts (top and bottom) for better visual representation.
Adapted with permission from MSc thesis of Gökberk Alagöz.

The next step in the analysis of NANOGP1 protein sequence focused on the conservation of its key functional region, the homeobox DNA-binding homeodomain. Multiple sequence alignment was generated using the data shown in Figure 3.6. As a result, the homeodomain was found to be fully conserved between human, chimpanzee and gorilla genomes whereas in orangutan and rhesus macaque it possessed one and four amino acid substitutions, respectively (Figure 3.7). Based on the DNA recognition properties of conserved NANOG homeodomain (see Section 3.3.2), it is likely that rhesus macaque DNA recognition would be impaired, while the other species would have likely preserved the domain functionality.

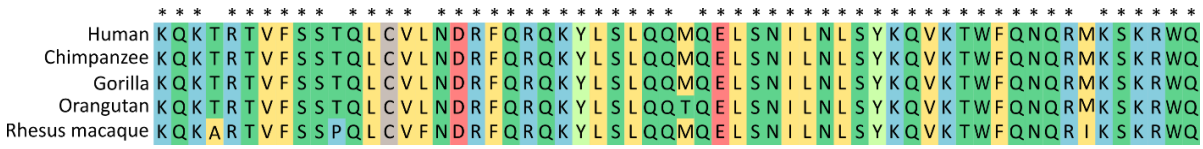


Figure 3.7 Amino acid sequence alignment showing aligned homeodomain sequences of NANOGP1 orthologs. Amino acid residues are indicated with the universally accepted single-letter code. Colour coding indicates different types of amino acids, according to their biochemical properties. ***** - amino acid residue is the same for all aligned sequences.
Adapted with permission from MSc thesis of Gökberk Alagöz.

In summary, *NANOGP1* sequence is retained and conserved only in Hominids. In contrast, in all other species, including Gibbon and Old/New World monkeys, *NANOGP1* has been disabled through a variety of different types of deletions and mutations. 'Predicted protein sequence' analysis demonstrated that the protein-CDS is mostly conserved in human, chimpanzee, orangutan and gorilla and the protein likely would be functional in those species. The consequences of mutations in rhesus macaque homeodomain would likely be more detrimental and await future research.

To conclude, *NANOGP1* pseudogene sequence remained highly conserved in Hominid species for at least 40 million years, suggesting that it could be functional there. This is particularly curious in relation to the human pluripotency.

3.2.5 Characterising *NANOGP1* mRNA expression in naïve hPSCs

The high degree of *NANOGP1* sequence conservation within the Great Ape branch implies that the gene might have a functional role in these species. To investigate this topic further, in this section, conservation and expression of human *NANOGP1* were analysed at the transcript level.

NANOGP1 transcripts were reported as detectable in primed hPSC cultures (Hart et al., 2004), providing an exciting prospect to investigate *NANOGP1* expression in hPSCs further in this thesis. The first question here was whether *NANOGP1* RNA could be detected in hPSCs, i.e., whether the result published in 2004 could be replicated. To investigate this, *NANOG* and *NANOGP1* RNA expression was analysed using published RNA-seq naïve and primed hPSC datasets (Collier et al., 2017). As a result,

NANOGP1 transcripts were detected both in primed and naïve hPSCs (Figure 3.8 *NANOGP1* and *NANOG* have different expression patterns in the naïve and primed hPSCs. RNA-seq reads are mapped against the published genome assembly, GRCh38_v9.0. Position of the locus within the chromosome, bp. RNA-seq datasets (n=3) are from Collier et al., 2017.. As seen in the wiggle plot and RNA-seq data tracks below, reads mapping to the *NANOGP1* sequence revealed a clear exon-intron structure, suggesting that *NANOGP1* mRNA undergoes splicing in hPSCs. A striking difference was discovered between *NANOGP1* and *NANOG* expression patterns in the hPSCs. *NANOG* was highly expressed both in the naïve and primed states, while *NANOGP1,* in contrast*,* exhibited high expression in naïve cells only, and its primed transcript level was noticeably downregulated.

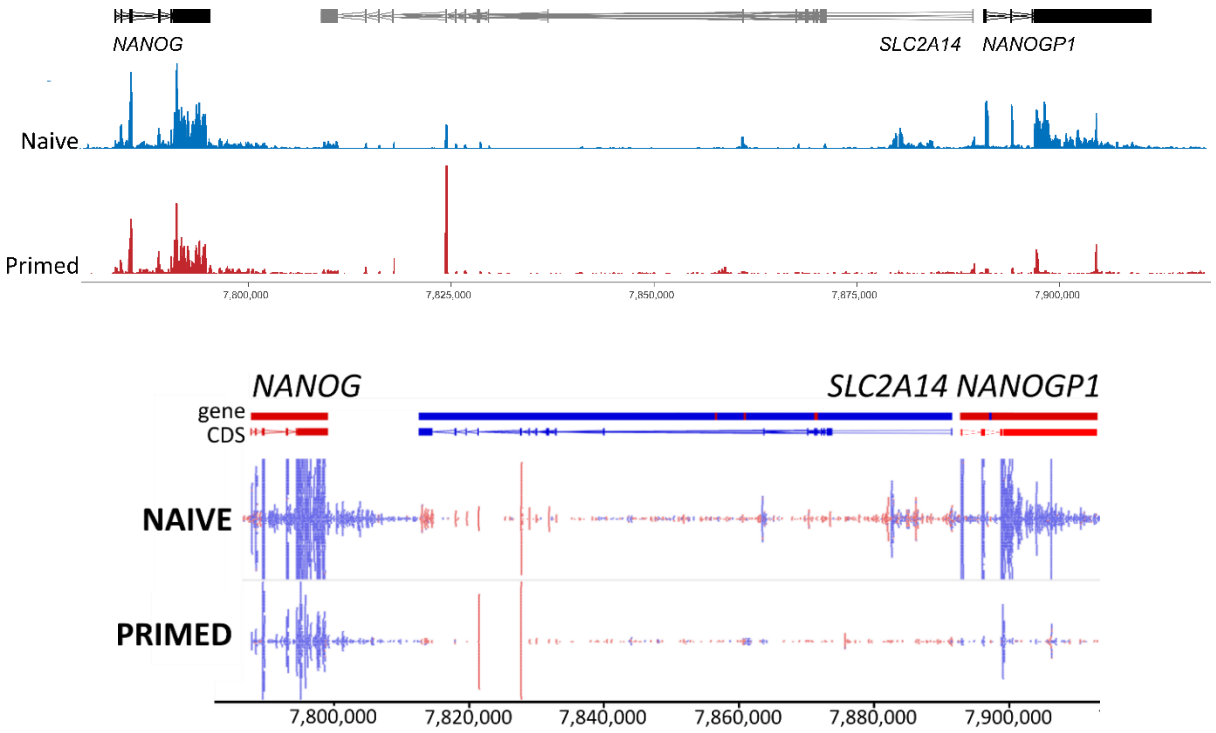


Figure 3.8 *NANOGP1* and *NANOG* have different expression patterns in the naïve and primed hPSCs. RNA-seq reads are mapped against the published genome assembly, GRCh38_v9.0. Position of the locus within the chromosome, bp. RNA-seq datasets (n=3) are from Collier et al., 2017.
Top: Wiggle plot. Naïve hPSC RNA expression is in blue, primed hPSC RNA expression is in red. CDS is shown as rectangles (exons) and lines (introns).
Bottom: RNA-seq data tracks. Numbers of mapped reads in the 'naive' and 'primed' tracks contribute to larger peaks, representing higher gene expression. Mapped reads, gene and CDS are in red and blue, corresponding to the two opposing DNA strands. Gene structure is shown as rectangles under gene names. CDS is shown as rectangles (exons) and lines (introns). Data tracks were generated using SeqMonk sequence analysis tool.

At the time of analysis, the precise structure of *NANOGP1* mRNA was not known. Since *NANOGP1* expression appeared particularly elevated in the naïve hPSCs, the exon-intron composition

of *NANOGP1* transcripts were investigated using three publicly available naïve hPSC RNA-seq datasets: WIBR3 (cultured in 5i/L/A medium), UCLA1 (cultured in 5i/L/A/F medium) and NK2 CR-H9 (cultured in t2iLGo medium), originally published in Theunissen et al, 2016, Pastor et al., 2016 and Takashima et al., 2014, respectively. The analysis involved finding splicing evidence among reads mapping against the *NANOGP1* genome sequence, as well as identifying its likely exons and introns. Collectively, this led to a conclusion that *NANOGP1* could be transcribed into three mRNA isoforms (Figure 3.9). All three isoforms were very similar in their exon-intron structure, bearing some differences in the last exons, which is addressed further. All three isoforms were also detected within each of the three analysed datasets, demonstrating that *NANOGP1* transcripts had the same mRNA structure in different cell lines and medium conditions, strengthening my confidence in the predicted structures. It was not possible, however, to identify whether either of the isoforms were expressed more than the others.

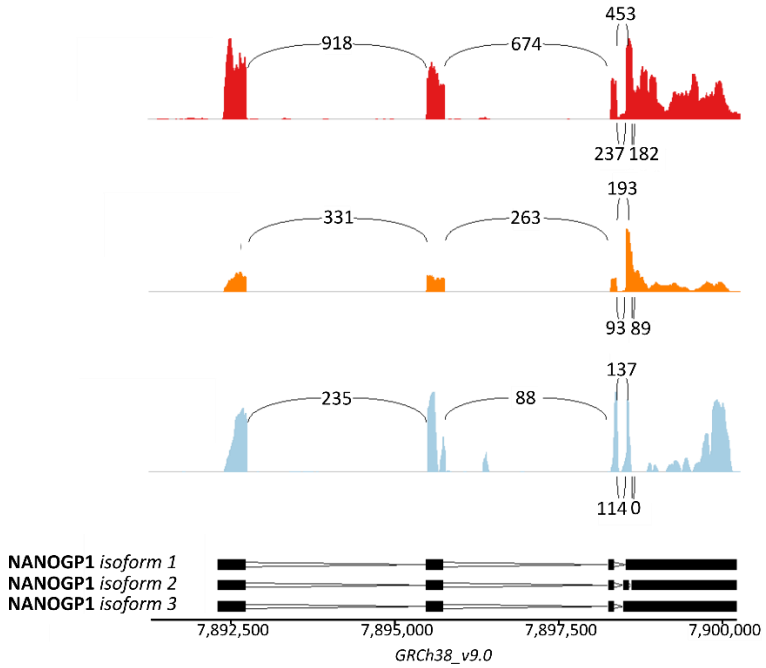


Figure 3.9 Sashimi plot showing splicing analysis summary (top) and three predicted mRNA isoforms for *NANOGP1* (bottom). RNA-seq peaks corresponding to exons are shown in red, orange and blue, each colour representing one individual dataset (Theunissen et al., 2016 – red; Pastor et al., 2016 – ornate; Takashima et al., 2014 – blue). Values in between the RNA-seq peaks represent the number of times a splicing event was recorded.
The scale at the bottom of the mRNA isoform diagram represents position of the locus within the chromosome, bp. GRCh38_v9.0 – human genome assembly version.
Adapted with permission of Dr. Felix Krueger.

To further characterise *NANOGP1* mRNA variants, The NCBI Open Reading Frame Finder tool was used to identify potential ORFs for the three *NANOGP1* transcript isoforms. Their structure and comparison to *NANOG* mRNA is summarised in (Figure 3.10).

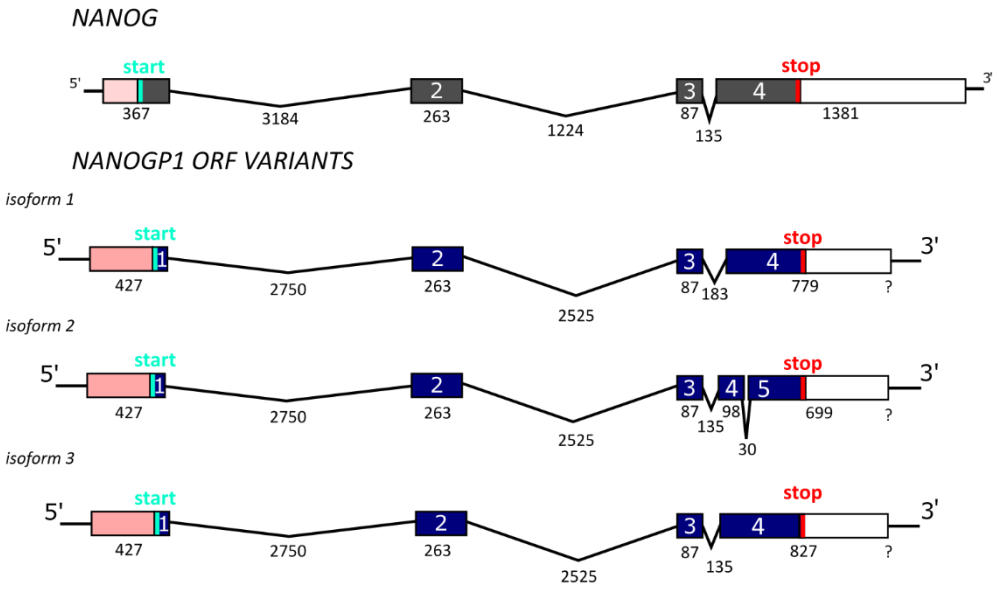


Figure 3.10 Diagram showing *NANOGP1* open reading frame (ORF) variants predicted based on exon-exon splicing analysis of naïve hPSC RNA-seq data and compared to *NANOG* ORF. ORF variants were predicted based on the exon-exon splicing analysis of naïve hPSC RNA-seq data. Exons are shown as blue blocks, introns – as black lines, 5' and 3' UTR as pink and white blocks respectively. Start (green) and stop (red) labels mark the predicted location of start and stop codons. An ORF component length is shown as a number of nucleobase pairs. Question marks indicate that the exact end point of *NANOGP1* mRNA tail was not determined.
Adapted with permission of Dr. Felix Krueger.

Interestingly, for all three isoforms, the predicted ORFs used an alternative start codon to the one that was previously put forward (Booth and Holland, 2004). Additionally, no splicing to an upstream putative exon was detected, as had been originally proposed by Booth and Holland. This demonstrated that the structure of *NANOGP1* mRNA differed from the previously hypothesised structure. According to the splicing analysis in this thesis, instead of the distant upstream exon 1, *NANOGP1* exon 1 was predicted to be the same as that of *NANOG*, but significantly shortened at its 5' end, as shown in the figure above. This was due to *NANOGP1* using an alternative start codon, located downstream of the start codon that *NANOG* would normally use. The reason for this was a stop codon present soon after the *NANOG*-specific ATG in the *NANOGP1* sequence, which would terminate translation only several amino acids after the start.

Usage of the newly-predicted start codon would result in *NANOGP1* missing 117 bp that is found at the start of *NANOG*, resulting in a decrease in the size of the first exon from 50 to 11 codons. Similar truncations were also suggested previously by Hart et al., Eberle et al., Fairbanks et al. and Booth and Holland. Despite the substantial difference in the size of exon 1, the rest of the *NANOGP1* ORF was very similar to that of *NANOG*, which included almost identical exons. Finally, it was not possible to determine the exact end point of the *NANOGP1* 3'-UTR (untranslated region). However,

this did not prevent establishment of either the exon-intron structure or the most likely start and stop codons.

In the next step, *NANOGP1* mRNA variants were used to predict protein sequence, which resulted in prediction of nearly full-length proteins in the case of all three isoforms (Figure 3.11). In agreement with previously published data, all three variants were lacking a 39 amino acid region at the N-terminus. The rest of the protein sequence was highly similar to the sequence of NANOG. All three protein variants had highly conserved homeodomains, identical to that of NANOG at the amino acid level. One substitution mutation, S285N, was found in the transactivation domain in all three NANOGP1 isoforms. Additionally, isoforms 1 and 2 had one short deletion each. The deletion in NANOGP1 isoform 1 (herein, NANOGP1-1) was located between the homeodomain and the tryptophan-rich region, while the NANOGP1-2 deletion included two tryptophan residues in the beginning of the tryptophan-rich region. All the other synonymous and non-synonymous amino acid mutations were identical between the three isoforms and were predominantly contained in the N-terminal region, outside of the key functional domains.

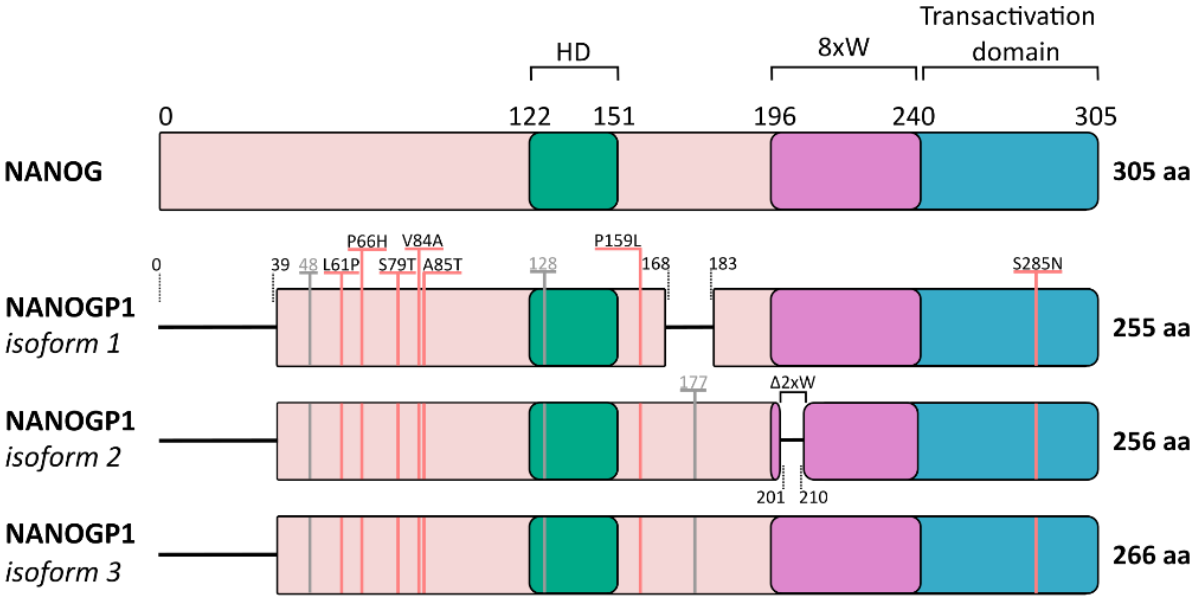


Figure 3.11 Diagram showing three predicted NANOGP1 protein isoforms compared to NANOG protein domain structure. Domain structure of the three NANOGP1 protein variants was estimated from the three predicted *NANOGP1* open reading frames. Peptide sequence deletions are labelled by vertical black dotted lines. NANOGP1 vs. NANOG amino acid substitutions caused by missense DNA mutations are labelled by red vertical lines. Silent mutations are labelled by grey vertical lines. Amino acid letter code: A – alanine, H – histidine, L – leucine, N – asparagine, P – proline, S – serine, T – threonine, V – valine. 8xW – tryptophan–rich subdomain/region containing 8 tryptophan (W) residues, Δ2xW – deletion of two tryptophan residues from the tryptophan-rich subdomain, HD – DNA-binding homeodomain, aa – amino acid.

Having identified that *NANOGP1* is highly expressed in naïve hPSCs and has three predicted mRNA isoforms which encode three nearly full-length protein variants, it was important to validate

that none of the reads mapping to *NANOGP1* belonged to *NANOG*. Hence, predicted *NANOGP1-1* was used to investigate potential cross-mapping of reads from the *NANOG* to the *NANOGP1* locus, and vice versa. To do this, cDNA sequences for *NANOG* (*NANOG-201*, Ensembl) and *NANOGP1-1* were converted into the two types of simulated FastQ files: 43bp (like in Petropoulos et al., 2016) and 100 bp single-end reads, in steps of 1bp. Collectively, these simulated reads represented *NANOG-201* and *NANOGP1-1* sequences from start to end, and were used to assess their potential cross-mapping between the two duplicates. To do this, the simulated reads were aligned to the human genome and, as a result, the degree of cross-mapping between *NANOG-201* and *NANOGP1-1* was either negligible or non-existent for unfiltered (MAPQ =0) or multi-mapping filtered (MAPQ >=20) reads, respectively. This confirmed that the identified putative transcripts were indeed based on *NANOGP1* reads and not *NANOG*.

After showing that the transcription described above indeed corresponded to *NANOGP1*, based on the three variants three new *NANOGP1* annotation tracks were created and added to SeqMonk sequence analysis software for all subsequent RNA analysis. Here and further in the thesis the annotation tracks were used as a tool for quantificating *NANOGP1* transcript abundance.

Previously in this chapter, *NANOGP1* was shown to be highly expressed in naïve hPSCs and noticeably downregulated in the primed cells, which was in contrast to the expression pattern of *NANOG* (Figure 3.8). Using the new annotation tracks, the difference between *NANOGP1* primed and naïve expression levels was additionally confirmed by analysing several other naïve and primed hPSC RNA-seq datasets, including embryo-derived and reprogrammed cell lines cultured in different media conditions (Figure 3.12).

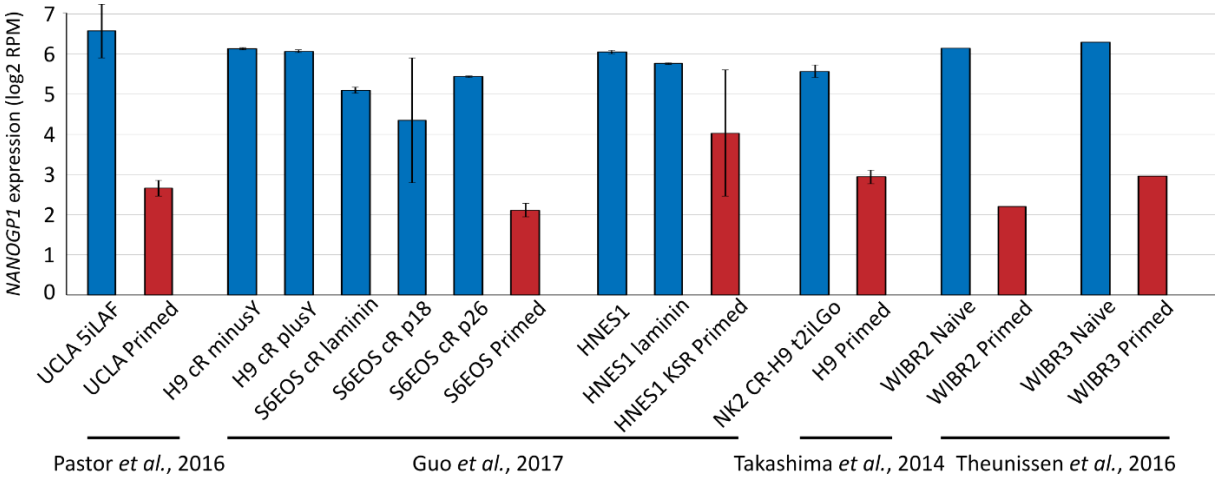


Figure 3.12. Bar chart showing *NANOGP1* RNA expression in naïve (blue) and primed (red) hPSC lines. Gene expression values are in log2 RPM (reads per million). Error bars represent mean ± standard deviation. Cell line names are used to label the x-axis. RNA-seq data was taken from the four published studies; citations are shown at the bottom of the graph.

In summary, three *NANOGP1* mRNA isoforms were identified in naïve hPSCs. All the three mRNA variants and their associated ORFs were highly similar to those of *NANOG*, with the major protein domains being present and seemingly intact in the predicted NANOGP1 protein variants. This suggests that NANOGP1 protein is likely functional, if it is expressed in hPSCs. The data presented here also resolved inconsistences in the literature rddiaegarding *NANOGP1* mRNA and ORF structure, providing splicing evidence for the *NANOGP1* exon-intron composition and identifying a new, previously unknown, *NANOGP1*-specific start codon. Finally, conservation of the *NANOGP1* CDS compared to its ancestral gene suggests presence of positive conservation acting on *NANOGP1,* implying potential functionality of *NANOGP1* in human early development.

3.2.6 Characterising *NANOGP1 RNA* expression in human embryo and hPSCs

In the previous section, *NANOGP1* transcripts were detected in naïve and primed hPSCs, and the expression pattern appeared to be similar to that of *NANOG* in the naïve state. In primed hPSCs, however, *NANOGP1* appeared to become downregulated, while NANOG did not exhibit such a noticeable change in its expression levels. To further investigate *NANOGP1* expression, its RNA expression was examined in cell types corresponding to where its ancestor *NANOG* is normally expressed, including the developing human embryo and the developing germ line.

A published embryo scRNA-seq dataset, Petropoulos et al., 2016, was used to investigate *NANOGP1* expression patterns in human embryos. The analysis demonstrated that *NANOGP1* is highly expressed in ICM and pre-implantation epiblast, similar to *NANOG*. Notably, in these two compartments *NANOGP1* was expressed at the same level as its ancestor, and also additionally had lowered transcript levels in the extraembryonic lineages, which also mirrored the *NANOG* expression pattern. One noticeable difference in the expression of the tandem duplicates was that *NANOG* mRNA was found in the developing embryo at the 8-cell stage, morula, and then gradually increasing in the developing ICM and epiblast; *NANOGP1* expression, however, was mostly limited to ICM and epiblast and was significantly lower in the 8-cell and morula stages, compared to *NANOG*. Surprisingly, *NANOGP1* was also detected in a subpopulation of extraembryonic primitive endoderm, unlike *NANOG* whose expression was absent in primitive endoderm and, instead, slightly elevated in some cells of the trophectoderm (Figure 3.13).

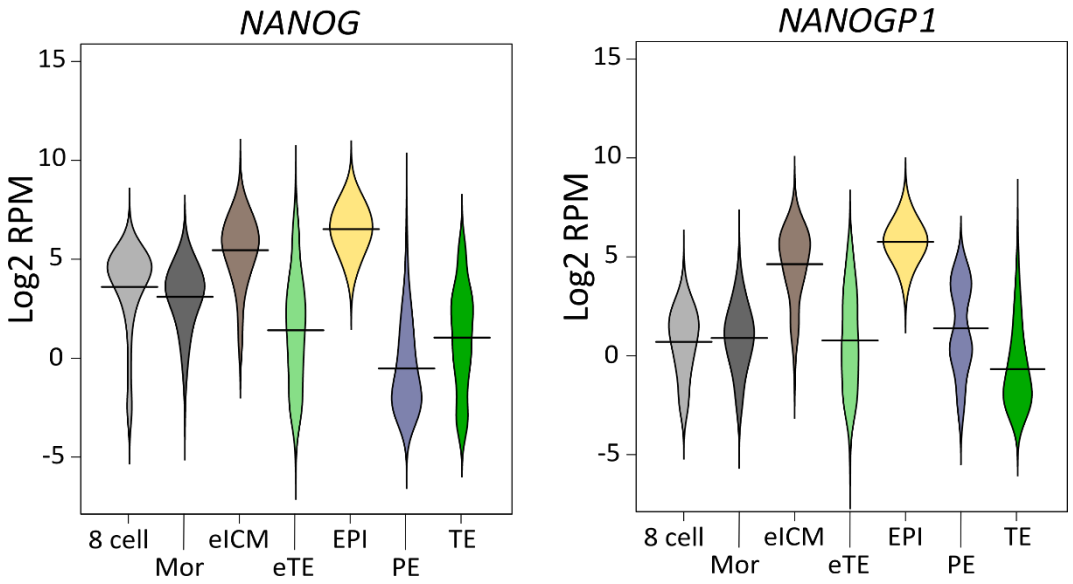


Figure 3.13. Violin plots show *NANOG* and *NANOGP1* RNA expression in the developing human embryo. Gene expression values are in log2 RPM (reads per million). Data is presented as violin plots and contains markers for the median (horizontal lines). 8 cell – 8-cell stage, Mor – morula, eICM – early inner cell mass, eTE – early trophectoderm, Epi - epiblast, PE – primitive endoderm, TE – trophectoderm. RNA-seq data was taken from Petropoulos et al., 2016.
Adapted with permission of Dr. Felix Krueger.

Since *NANOGP1* expression was different between naïve and primed hPSCs, I decided to examine whether it would change during pre-and post-implantation epiblast development as well. To do this, I used an in vitro-cultured embryo scRNA-seq dataset (Xiang et al., 2020). Analysing the expression on a per-cell-basis, I observed two different transcriptional profiles (Figure 3.14). In the developing epiblast, *NANOG* exhibited high expression starting from Day 6 until Day 14, while the *NANOGP1* expression window was narrower. It was also detected between Day 6 and Day 12, but then dropped in the majority of cells on Day 14.

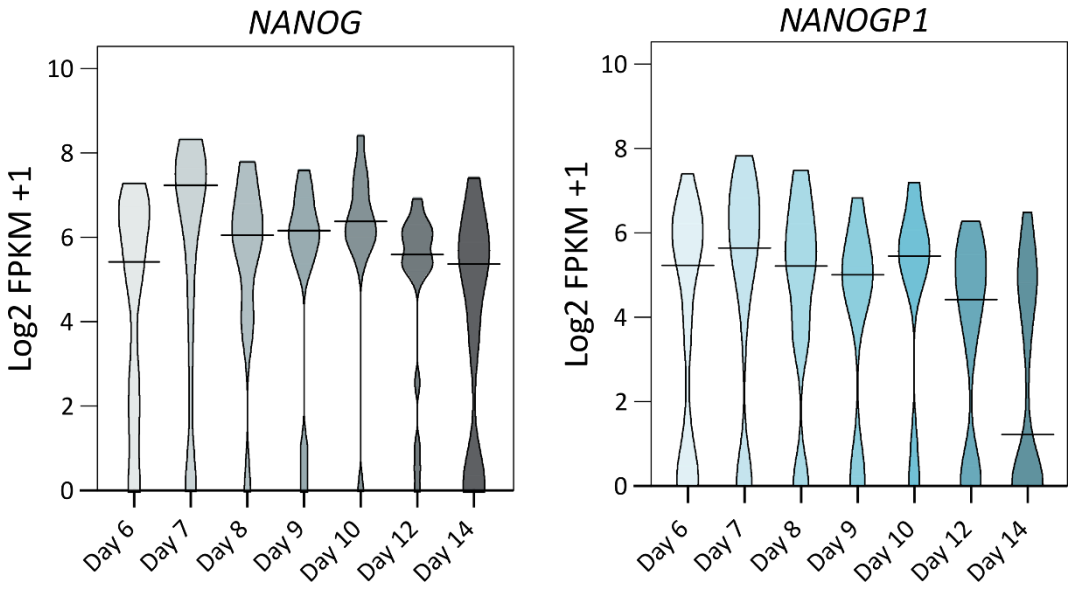


Figure 3.14 Violin plots showing *NANOG* and *NANOGP1* RNA expression in the developing human epiblast. Gene expression values are in log2 FPKM+1 (fragments per kilobase of exon per million mapped fragments). Data is presented as violin plots and contains markers for the median (horizontal lines). Day 6, Day 7, Day 8, Day 9, Day 10, Day 12, Day 14 – embryo developmental time points. RNA-seq data was taken from Xiang et al., 2020.

Finally, *NANOGP1* and *NANOG* expression was analysed in the developing human germ lineages of male and female embryos (Gkountela et al., 2015). Throughout the whole developmental timeline, starting from PGC stage and until advanced germ cell stage, *NANOGP1* was expressed in the same pattern and at the same level as *NANOG* (Figure 3.15).

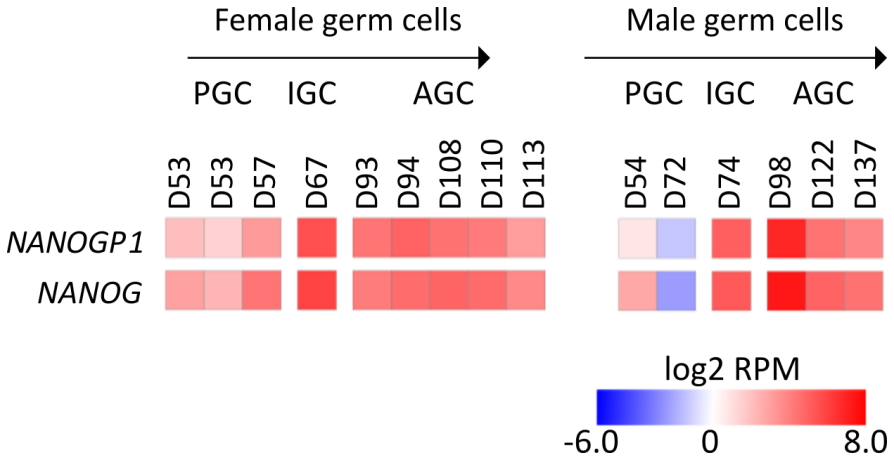


Figure 3.15 Heat maps showing *NANOG* and *NANOGP1* RNA expression in the developing male and female germ line. Progression of development is indicated with arrows. Gene expression values are in log2 RPM (reads per million) and are plotted as a colour gradient (blue – lower expression level, red – higher expression level). PGC – primordial germ cells, IGC – intermediate germ cells, AGC – advanced germ cells. RNA-seq data was taken from Gkountela et al., 2015.

In summary, in addition to the previously shown high expression of *NANOGP1* in naïve hPSCs, *NANOGP1* is also highly transcribed in human embryos, and its expression pattern and levels there are similar to those of *NANOG*. These results lead me to conclude that, since the two duplicates have closely overlapping expression patterns, *NANOGP1* could have a conserved function in the cells and embryo compartments where its expression is high. It is worth noting, though, that the expression patterns are very similar, but not identical. Indeed, *NANOGP1* transcription appears to be more temporally restricted. Compared to *NANOG*, its transcript levels are lower in the early stages of embryo development (8-cell, morula), by Day 14 in the developing epiblast, as well as in primed hPSCs. To conclude, this chapter has revealed previously unknown *in vivo* expression patterns of *NANOGP1*, strengthening the hypothesis that it could have conserved functionality in human early development with, potentially, a more specific role due to a shorter developmental window of expression.

3.2.7 Characterising putative regulatory regions of *NANOGP1*

NANOG and *NANOGP1* DNA sequences, predicted mRNA and protein structure appear to be mostly similar between each other, yet their RNA expression patterns in hPSCs and human embryo do not completely overlap. As shown in Sections 3.2.5 and 3.2.6, both duplicates are highly expressed in the ICM and epiblast, as well as naïve hPSCs. However, in human embryo, the expression of *NANOGP1* is more temporally restricted compared to *NANOG*, and it is also significantly lower in the primed hPSCs in comparison to its ancestral copy. The similar, but non-identical, expression patterns allowed me to hypothesise that the two duplicates could have divergent upstream regulation.

First, I aimed to investigate whether any *NANOG* regulatory regions can be found upstream of *NANOGP1*, suggesting that they had been created during the tandem duplication event. I also checked whether any other, non-duplicated, regulatory regions were present near *NANOGP1* that could be involved into its regulation. Here, understanding what mechanisms could be involved in establishing RNA expression patterns of the two duplicates was essential for learning potential functional conservation and/or divergence that *NANOG*/*NANOGP1* might exhibit.

In this section, I used Chovanec et al., 2021 to obtain coordinates of putative *NANOG* and *NANOGP1* regulatory regions (enhancers and super-enhancers), that had been originally annotated using the ChromHMM chromatin state identification tool (Ernst and Kellis, 2012), and the super-enhancer ROSE pipeline (Lovén et al., 2013; Whyte et al., 2013). I also used published datasets to characterise putative regulatory regions of *NANOGP1* and *NANOG* by their chromatin status (Assay for transposase-accessible chromatin sequencing (ATAC)-seq: 'active'/'inactive' chromatin) (Pastor et al., 2016), DNA methylation level (Theunissen et al., 2016), transcription factor binding profile (ChiP-seq: NANOG, OCT4 and SOX2) (Chovanec et al., 2021; Ji et al., 2016) and histone modifications (ChiP-

seq: H3K27ac, H3K4me1 and H3K4me3, marking ‘active’ chromatin; H3K27me3, marking ‘inactive’ chromatin) (Chovanec et al., 2021; Gifford et al., 2013; Ji et al., 2016; Theunissen et al., 2014) that occur in the *NANOG*/*NANOGP1* locus. Results of the analysis are depicted in Figure 3.16 and described in detail in the text below.

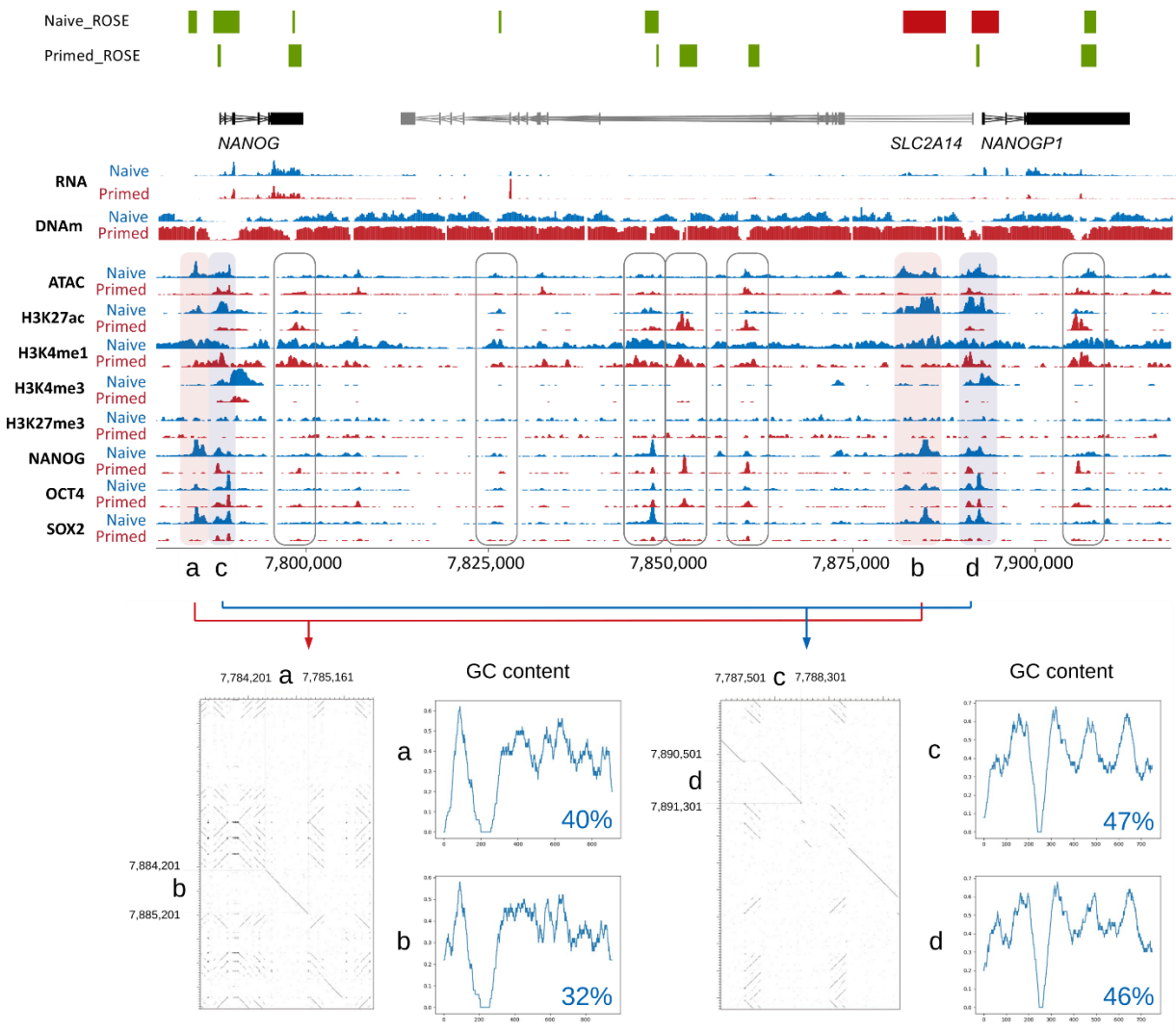


Figure 3.16 Summary figure of the RNA-seq, ATAC-seq, ChIP-seq (top) and GC content analysis (bottom) of the predicted *NANOG* and *NANOGP1* regulatory regions.
Top: Sequencing reads are mapped against the published genome sequence. GRCh38_v9.0 – human genome assembly version. Higher numbers of mapped reads in the 'Naïve’ and 'Primed' tracks contribute to larger peaks, representing higher gene expression (RNA-seq), more abundant protein binding and histone modifications (ChiP-seq), more ‘open’ chromatin state (ATAC-seq) and higher levels of DNA methylation (DNAm).
Four shaded boxes (a, b, c, d) represent two duplicated pairs of regulatory regions. Green blocks – enhancers, red blocks – super enhancers, according to the Chovanec et al. annotation (based on H3K27ac ChIP-seq data run through ROSE super-enhancer ranking pipeline).
Scale at the bottom of the track diagram represents position of the locus within the chromosome. Gene/pseudogene CDS structure is shown as rectangles (exons) and lines (introns).
See main text for dataset references.

Bottom: Dot plots and GC content ratio line graphs showing comparison of the 'naïve' ('a' and 'b') and 'shared' ('c' and 'd') putative *NANOG* and *NANOGP1* regulatory regions.
Dot plots: Individual dots represent matching base pairs between the two aligned sequences. In areas of sequence conservation individual dots form diagonal lines. Chromosome position, bp.
GC content ratio graphs: x-axis represents the length of a putative regulatory region, bp; y-axis shows (G+C)/(G+C+A+T) values. Average GC content ratio is shown in the right bottom corner of each graph, %.
Bottom: adapted with permission from MSc thesis of Gökberk Alagöz.

Collectively, the data revealed the location of several potential regulatory regions, annotated as enhancers and superenhancers (Chovanec et al., 2021) in the *NANOG/NANOGP1* locus. Four of those regions were positioned as two pairs directly upstream of *NANOG (***a, c***)* and *NANOGP1* (**b, d**). A pair-wise comparison was performed to identify whether they were formed as a result of the tandem duplication. Pair-wise alignments showed that the sequences within the two individual pairs, **a/b** and **c/d**, were very similar; additionally, each pair had matching GC content profiles, proving that they had| formed during a duplication event.

GC content ratio values for **a** and **b** were 32% and 40%, compared to 47% and 46%, respectively, for **c** and **d**. In the case of the **c**/**d** pair, GC content ratios were close to typical GC content ratio values, which normally average ~50% in the promoter areas (Villar et al., 2015), unlike the **a/b** pair which had lower GC content values. Additionally, in contrast to **a/b**, the **c/d** pair was enriched for the H3K4me3 modification, a known mark of active promoters (Barski et al., 2007). This allowed me to conclude that **c/d** are likely to serve as promoters while **a/b** serve as enhancers**.**

According to the ATAC-seq profile, sites **a, b, c** and **d** were characterised as having highly accessible, or 'active', chromatin. Additionally, all four regions had high acetylation and 'active' methylation histone marks and were bound by pluripotency factors in either one or both pluripotent states. These data supported the hypothesis that the four regions could have a regulatory role.

Interestingly, based on the protein binding signal, histone marks, chromatin status and, finally, ROSE pipeline, putative promoters **c** and **d** appeared active in both naive and primed states, and were hence referred to as 'shared', while the putative enhancers **a** and **b** were predominantly marked as active in the naïve hPSCs, and therefore, were called here 'naïve'.

Pluripotency factor binding, histone modification and chromatin status profiles were very similar between *NANOG* and *NANOGP1* putative promoter regions. The only prominent difference was the SOX2 protein binding and H3K4me3 profiles within the 'shared' putative promoters. More specifically, SOX2 and H3K4me3 peaks were detected near *NANOG* in both primed and naïve hPSCs, but at the *NANOGP1* locus they were only present in the naïve state.

Finally, in addition to putative enhancers and promoters **a, b, c** and **d**, six other enhancers were identified near the pseudogene (highlighted in the figure above). Their roles would require additional investigation in the future.

In summary, this section demonstrates that *NANOGP1* expression regulation is integrated within the regulatory circuitry of pluripotent cells, as its putative regulatory sites are targeted by OCT4, SOX2 and NANOG binding. Additionally, the two duplicated potential regulatory regions upstream of *NANOGP1* exhibit high sequence, GC content, histone modification and pluripotency factor binding similarity with the putative regulatory regions found upstream of *NANOG*. Collectively, these data explain the overall similarity of the *NANOGP1* and *NANOG* expression in human embryo and naïve hPSCs. Some differences were identified between the duplicated putative promoter profiles of *NANOG* and *NANOGP1* in the primed pluripotent state, which could be interesting to investigate in the future, asking if these factors could be responsible for the difference in the expression patterns of *NANOGP1* and *NANOG*.

3.2.8 Exploring the expression of other pseudogenes in human naïve pluripotency

Previously in this chapter, I described *NANOGP1* expression in the developing human embryo and PSCs, while comparing it to the expression of the ancestral copy, *NANOG*. To conclude the chapter, I decided to broaden the analysis and examine pseudogene expression in human naïve pluripotency more generally, as well as the pseudogene localisation, putative origin and function of their ancestral copies.

First, I analysed the expression of human pseudogenes using a naïve hPSC RNA-seq dataset, produced in this study (described in Chapter 5). By applying a name filter method to select all gene name entries ending with 'PN' (where N is a number from 1 to 9) pseudogenes were selected. As a result of this approach, 1880 protein-coding genes in the human genome were found to have copies, labelled as pseudogenes. A total duplicate copy number analysed here was 6922.

Here, I identified that 563 pseudogenes out of 6922 were expressed at a reasonable level (log2 RPM>0) in naïve hPSCs (Figure 3.17).

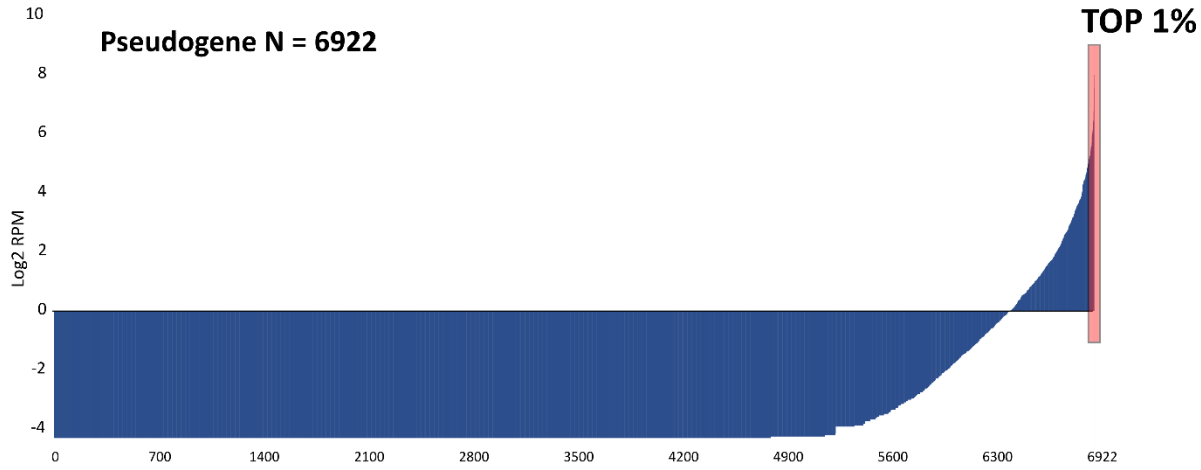


Figure 3.17 Bar chart showing pseudogene RNA expression in naïve hPSCs. RNA-seq dataset was produced in this study, see Chapter 5). Gene expression values are in log2 RPM (reads per million). Values show the mean of three biological replicates. Top 1% box highlights the top 1% highest expressed pseudogenes.

In addition to *NANOG*, another core pluripotency regulator, *POU5F1* (the gene encoding OCT4), as well as a naïve-specific factor, *DPPA3*, had highly expressed pseudogene copies in naïve hPSCs. Three of these duplicates, *NANOGP1, POU5F1P3* and *DPPA3P2*, were among the top 1% highest expressed pseudogenes in the whole genome, and their expression levels were comparable to the three core pluripotency regulators *NANOG, POU5F1* and *SOX2* (Figure 3.18).

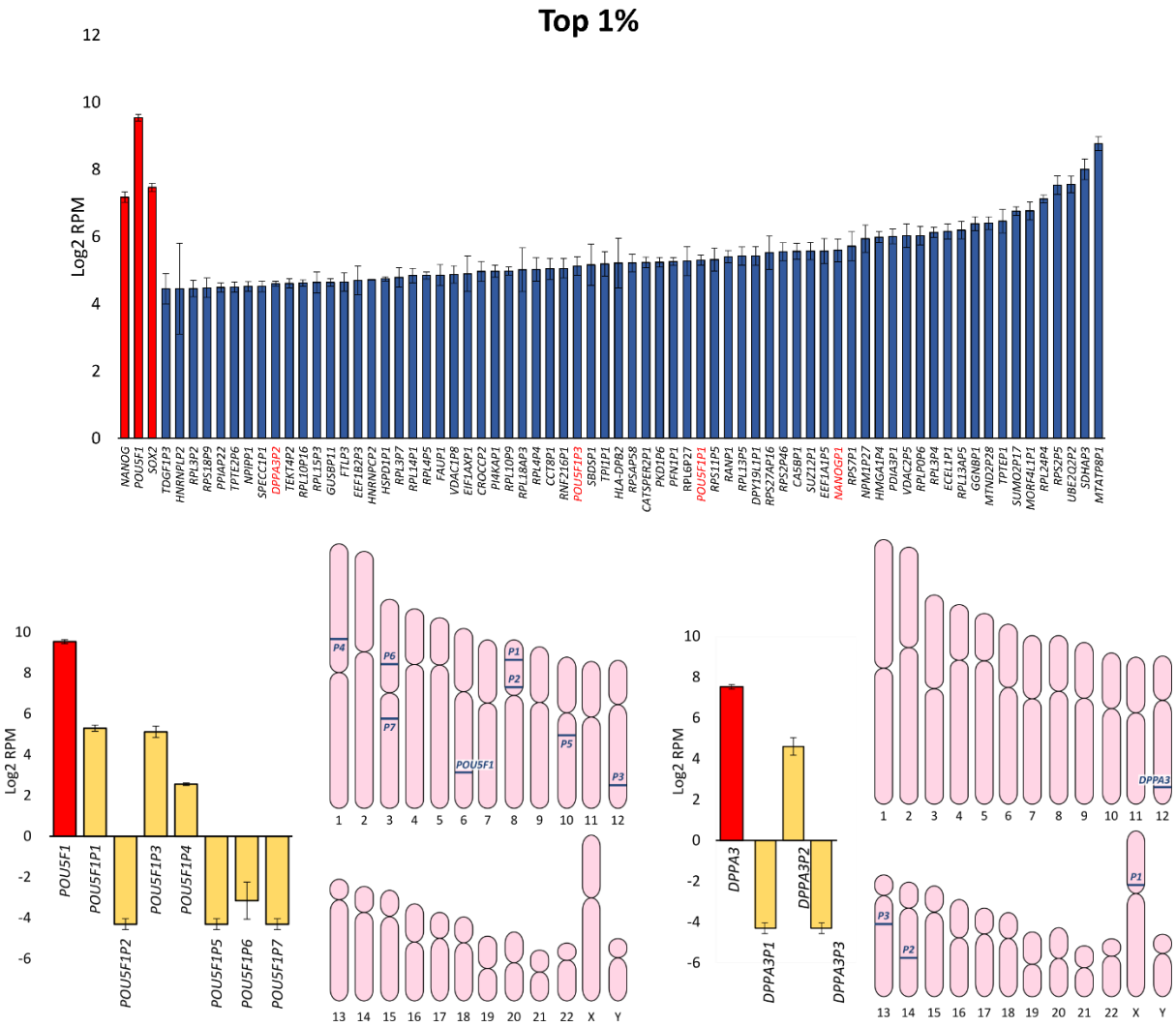


Figure 3.18 Bar chart showing top 1% highest expressed pseudogenes in naïve hPSCs (top) and diagram showing naïve *hPSC* expression and chromosomal location of *POU5F1* and *DPPA3* pseudogenes.

Top: Top 1% highest expressed pseudogenes in naïve hPSCs. For the expression level reference, 69 highest expressed pseudogenes (blue bars) are compared to the expression of three core pluripotency factors*, NANOG, SOX2, POU5F1* (red bars). Pseudogenes of pluripotency transcription factors are in red.

Bottom: Bar charts show *POU5F1* and *DPPA3* RNA expression levels in naïve hPSCs (red) compared to the expression of their pseudogenes (yellow). Idiograms show 22+XY human chromosome set (pink) with locations of the pluripotency genes and their corresponding pseudogenes (blue). Gene and pseudogene chromosome location information was obtained from UCSC Genome Browser. RNA-seq dataset was produced in this study - see Chapter 5. Gene expression values are in log2 RPM (reads per million). Values show the mean of three biological replicates. Error bars indicate ± standard deviation.

Interestingly, in addition to the structurally conserved *NANOGP1*, processed duplicates *NANOGP4* and *NANOGP8* were also expressed in naïve hPSCs (Figure 3.19).

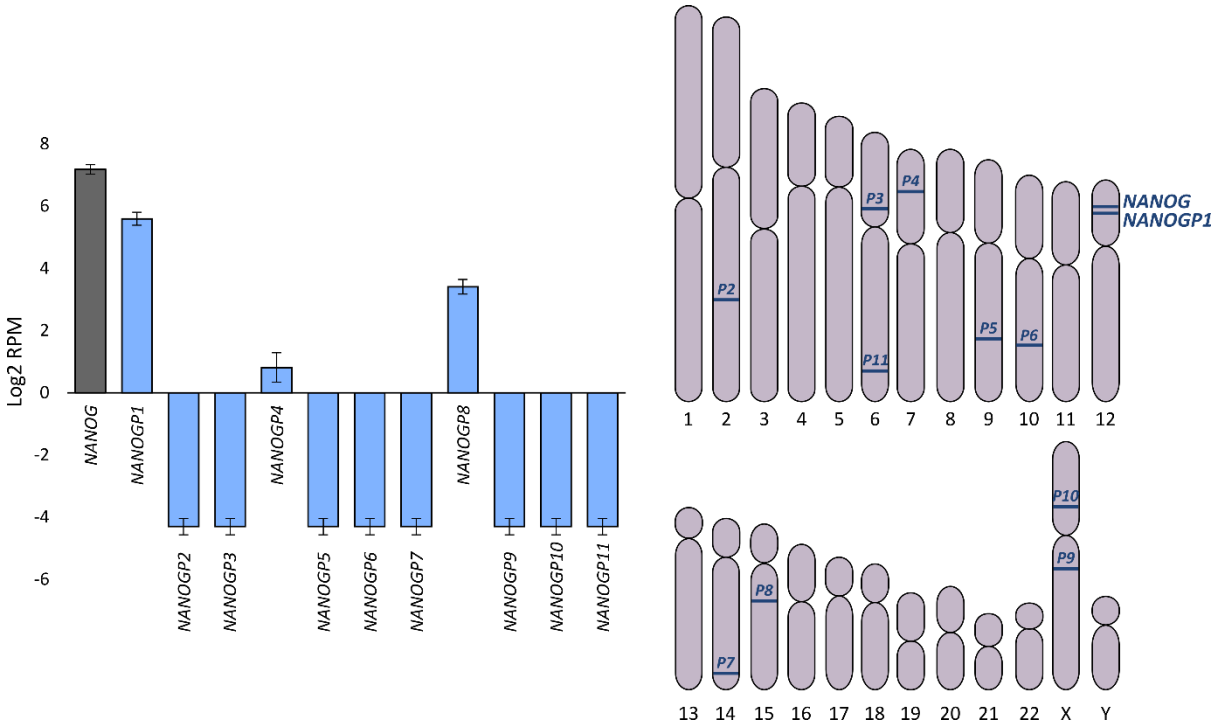


Figure 3.19 Bar chart showing RNA expression of *NANOG* and its pseudogenes in naïve hPSCs (left) and idiogram showing chromosomal location of *NANOG* and its duplicates (right).
Bar chart shows RNA expression levels of *NANOG* and its duplicates in naïve hPSCs (left). The RNA-seq dataset was produced in this study; see Chapter 2. Gene expression values are in log2 RPM (reads per million). Values show the mean of three biological replicates. Error bars indicate ± standard deviation. Idiogram shows 22+XY human chromosome set (grey) with locations of *NANOG* and its pseudogenes (blue). Gene and pseudogene chromosome location information was obtained from UCSC Genome Browser.

Detailed investigation of all other highly expressed pseudogenes was beyond the scope of this thesis; however, I anticipated that examining the functions that their ancestral copies perform could be insightful in studying the global role of highly expressed pseudogenes genes in naive hPSCs. To narrow down the list of pseudogenes to be examined here, I chose an arbitrary parameter: 'genes that exhibited higher expression than that of *NANOGP1'*. As described previously, *NANOGP1* was among the top 1% highest expressed pseudogenes, more specifically, the 20th from the top. Hence, here, I explored the origin of the remaining 19 highest expressed pseudogenes. I found that their ancestral copies were involved in processes such as ribosome functioning and protein synthesis/modification/ubiquitylation (*RPS7, PDIA3, RPLP0, RPL3, RPL13A, RPL24, RPS2, NPM1, SUMO2, UBE2Q2*), mitochondrial functioning (*MTATP8, MTND2, SDHA, VDAC2*), regulation of peptide hormone activity (*ECEL1*), chromatin organisation (*MORF4L2*), endocrine function (*TPTE*) and mouse spermatogenesis (*Ggnb)*. Notably, one of the highly expressed pseudogene ancestral copies was found to have a somatic-to-PSC reprogramming ability (*HMGA1*), (Kim et al., 2016).

Examining pseudogene locations within the genome, I observed that the majority of pseudogenes were positioned far from their ancestral copies, often on different chromosomes, which meant that they were likely created with help of transposons/retrotransposons (see *DPPA3* and *POU5F1* pseudogene chromosome location as examples, Figure 3.18). *NANOGP1,* formed by tandem duplication*,* was one of few pseudogenes found in the same locus as its ancestral copy, making it unique not only among the *NANOG* copies but other highly-expressed human pseudogenes as well.

In summary, I identified that *NANOGP1*, as well as copies of other pluripotency genes, *POU5F1* and *DPPA3*, were among the top 1% highest expressed genes in naïve hPSCs. In addition to the high expression level, *NANOGP1* was found to be unique in its genome location, being present in the same locus as *NANOG*, whereas the majority of other pseudogenes were located far away from their ancestral copies. Collectively, these results uncovered the large set of pseudogenes that are expressed in naïve hPSCs. In particular, the high expression of the duplicated and unprocessed pseudogene *NANOGP1* raises the possibility that this gene might have a functional role in hPSCs.

3.3 Discussion and future work

In this chapter, I described crucial topics related to the evolutionary conservation, mRNA expression pattern, regulation and predicted protein sequence of the pseudogene *NANOGP1*. The results strengthen the rationale for my hypothesis that *NANOGP1* might play a previously overlooked role in human pluripotency and development.

NANOGP1 was found to be highly conserved within the Great Ape branch, which included a highly similar structure between human *NANOG* and *NANOGP1* mRNA isoforms identified in naïve hPSCs. *NANOG* and *NANOGP1* were found to have overlapping but distinct expression profiles, both *in vivo* and *in vitro*. Putative regulatory regions of *NANOG* and *NANOGP1*, on the contrary, had very similar chromatin profiles. Finally, *NANOGP1* expression was compared to other pseudogenes expressed in naïve hPSCs, and *NANOGP1* was found to be among the highest expressed.

Taken together, Chapter 3 described previously unknown properties of *NANOGP1* pseudogene, highlighting its structural conservation and, at the same time, unique expression pattern in human pluripotency and embryo development. Thus, it emphasised the importance of functional assays that would help further characterise *NANOGP1* properties and test whether its potential function is also conserved and/or has unique properties, different from that of *NANOG*.

The functional assays are described further in the thesis, in Chapter 4 and Chapter 5.

3.3.1 Conservation of *NANOG*/*NANOGP1* duplication in primate evolution suggests potential functionality of *NANOGP1* in Great Apes

The first topic described in Chapter 3 addressed the structure of the human *NANOG*/*NANOGP1* duplication locus. Identification of the duplication boundaries had been challenging due to the high divergence in intergenic and intron sequences between the two duplicates. Moreover, this region contained not only *NANOG* and *NANOGP1*, but also another duplicated pair, *SLC2A14* and *SLC2A3*. The SLCA14/SLCA3 proteins are unrelated to NANOG function and, instead, are involved in glucose transport (Gould and Holman, 1993). Nevertheless, the approximate duplication boundaries were identified by analysing several distinct genomic features. The first of these is *NANOGNB*, which is located ~15 kb upstream of *NANOG*, but a copy of this gene is not present near to *NANOGP1*, indicating that the duplication area upstream of *NANOG* is likely less than 15 kb in size. Moreover, a 4 kb region upstream of *NANOG* that contains putative regulatory regions was found to be duplicated and thus also present upstream of *NANOGP1*. Collectively, these findings place the predicted start of the duplication in the ~ 11 kb intergenic region between *NANOG* and *NANOGNB*.

NANOGNB is worth discussing here separately due to its recently proposed connection to *NANOG*. Originally, *NANOGNB* (NANOG NeighBour) homeobox gene had not been considered as a *NANOG* copy due to the very high sequence divergence. Hence, it had been placed in a separate gene family (Zhong and Holland, 2011a; Zhong and Holland, 2011b). Therefore, initially, 'NANOG' in its name had only reflected its chromosomal location and not evolutionary origin. More recently, however, *NANOGNB* was suggested to be a cryptic duplicate of *NANOG* that had acquired such a high number of mutations as to be almost unrecognisable (Dunwell and Holland, 2017). Dunwell and Holland compared eutherian mammal NANOGNB domain structure to that of eutherian mammal, reptile and avian NANOG proteins. As a result, they discovered two small N-terminal motifs that were present in NANOGNB as well as in NANOG orthologs from multiple species. These new findings about *NANOGNB* gene indicate that *NANOG* could have more duplicates than initially assumed. According to Dunwell and Holland, this duplication occurred long before the *NANOG*/*NANOGP1* duplication event, specifically before mammals, birds and reptiles diverged. Additionally, the results presented in this chapter demonstrate that no indication of a *NANOGNB* pseudogene can be found between *NANOG* and *NANOGP1*, which means that if *NANOGNB* is a cryptic copy of *NANOG*, then *NANOG*/*NANOGNB* and *NANOG*/*NANOGP1* formed by two separate tandem duplication events, and the former has diverged at the sequence level significantly more than *NANOGP1*.

The 3' end of the *NANOG*/*NANOGP1* duplication was predicted based on the presence of *SLC2A3*, as well as its downstream conserved intergenic flanking region. Overall, the duplication of the original *NANOG-SLC2A14*-containing area, which was approximately 80 kb in size, resulted in the formation of a new *NANOGP1-SLC2A3* genomic locus. It is likely that in addition to this main duplication event, the area has undergone other genomic re-arrangements, since the *NANOGP1-SLC2A3* locus is noticeably longer than its ancestral copy, as seen in Figure 3.3. Based on the sequence conservation profile, shown in Figure 3.3, this is probably explained by insertional mutagenesis in *SLC2A3*. Together, the human *NANOG*/*NANOGP1* duplicated locus is estimated to cover ~ 250 kb in size and to contain two duplicated pairs of genes, as well as their corresponding intergenic regions.

Findings described in this chapter indicated that the duplication event occurred before the Hominoid and Old World monkey branches split. However, it is possible that this event took place even earlier, before the Old and New World monkey separated. The evidence for this is that *SLC2A14* and *SLC2A3*, but not *NANOGP1*, are present in the marmoset genome, which is a member of the New World monkey branch. This finding created two different potential duplication scenarios. It is possible that the duplication of *NANOG* and *SLC2A14*, described above, had indeed occurred in the marmoset genome, but *NANOGP1* is now missing due to its subsequent deletion. Alternatively, it could mean that the two *SLC2A* duplicates were formed in a separate duplication event in the common ancestor of New/Old World monkeys and Great Apes, and only later NANOGP1 was formed in a separate tandem duplication in the ancestor of the Old World monkeys and Great Apes. Unfortunately, the current quality of primate genome assemblies does not allow for distinguishing between the two scenarios, i.e., it is not possible to detect the 'scars' of *NANOGP1* duplication and subsequent deletion, or to compare putative New World monkey *SLC2A14*/*SCL2A3* duplication with the one detected in Great Apes and Old World monkeys. Therefore, at this point it is only possible to conclude that the duplication event took place at least ~ 40 mya; any further clarifying experiments will require better quality genome assemblies as they become available.

The overall conservation of the *NANOGP1* CDS in Great Ape genomes led to speculation of its potential functionality in the branch. Indeed, it appears that all members of the Old World monkey branch, which were investigated here, had disabled *NANOGP1* via different mechanisms, such as deletions and nucleotide substitutions. It looks clear that in Old World Monkey rhesus macaque, NANOGP1 has lost (or significantly altered) its DNA binding due to the substitution mutation in the key DNA recognition site within the homeobox domain, M109I. Indeed, this residue is conserved among NK homeodomain proteins and serves as the primary determinant of the DNA binding recognition sequence (Weiler et al., 1996; Weiler et al., 1998). Residue 109 often varies between species and mutating it usually leads to a significant decrease in protein binding affinity (Gehring et

al., 1994; Gruschus et al., 1997; Weiler et al., 1996; Weiler et al., 1998). In mouse, residue 109 was shown to be important for NANOG DNA recognition and binding (Jauch et al., 2008; Weiler et al., 1998), and a recent publication of human NANOG protein crystal structure demonstrated that mouse and human NANOG HDs are virtually identical in their structure (Hayashi et al., 2015). Therefore, replacing methionine 109 with isoleucine in rhesus macaque NANOGP1, which is highly similar to its human ortholog in the rest of the gene, probably altered its DNA binding site completely. The only homeobox protein that has isoleucine in position 109 is the transcription factor PBX1, and the binding consensus of PBX1 is different from that of NANOG: TGAT vs. TAAT, respectively (Chang et al., 1996; Piper et al., 1999). Therefore, it is possible to hypothesise that the M109I mutation in rhesus macaque NANOGP1 has led to the complete altering of the putative protein binding, potentially making it non-functional.

In summary, I conclude that the *NANOG*/*NANOGP1* duplication event occurred at least 40 mya and involved two genes, *NANOG* and *SLC2A14*, that were inserted downstream of their original location. While *NANOGNB* could be a highly diverged copy of *NANOG*, it did not appear to be involved in the NANOG/*NANOGP1* tandem duplication. *NANOGP1* sequence is highly conserved in Great Apes, which implies a potential function.

3.3.2 NANOGP1 mutations and their consequences on the putative protein functionality

In Section 3.2.5, *NANOGP1* was predicted to have three mRNA isoforms in the naïve hPSCs. Using an ORF-finding tool, the three isoforms were concluded to use the same start and stop codons and, as a result, likely to be translated into functional full-length protein variants with conserved NANOG-like domains, in addition to containing several small mutations. Here I discuss the potential effects of *NANOGP1* mutations on its functional properties.

According to the published literature (Chang et al., 2009; Do et al., 2009; Oh et al., 2005), human NANOG functionality is mediated by its central homeodomain (HD, tryptophan-rich region (8xW, or WR), and the C-terminal transactivation domain, which are involved in motif recognition and DNA binding, protein dimerisation, and transactivation respectively, as described further.

Three separate studies of *NANOG*, Chang et al., 2009; Do et al., 2009; Oh et al., 2005, conveyed detailed analysis of NANOG domain transactivation activity using variations of the luciferase reporter assay. Collectively, these studies tested the activity of the following NANOG domains: N-terminus (ND), HD, and C-terminus, which was separated into CD1, WR and CD2, as shown in Figure 3.20.

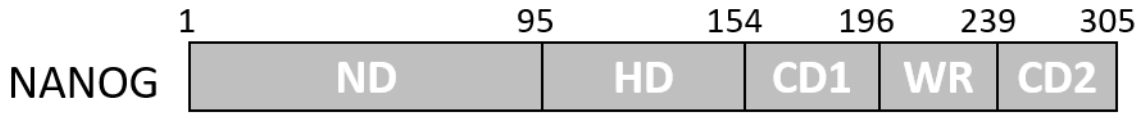


Figure 3.20 Diagram showing NANOG domain structure. Numbers indicate the number of amino acid residues. ND – N-terminus; HD – homeodomain; CD1, WR, CD2 – C-terminal subdomains. *Adapted from Do et al., 2009.*

The main conclusions of these three studies, relevant to the transactivation activity, were:

1) Only the C-terminus, CD1-WR-CD2, has transactivation activity, while ND and HD do not.
2) The CD1 domain within the CD1-WR-CD2 region cannot mediate transactivation, while WR and CD2 are highly potent.
3) Complete deletion of the WR region increases transactivation of the CD domains, while substituting W and Q residues in the WR stretch leads to the decrease of transactivation.
4) Only amino acid residues number 253 – 273 are crucial for the CD2 transactivation function.

Chang and colleagues also demonstrated that the WR region is required for human NANOG dimerisation (Chang et al., 2009). From their data, however, it is not clear what effect deleting the region containing two tryptophan residues, like in NANOGP1-2, would have on the dimer formation.

Therefore, I could hypothesise that the small deletion within CD1, the substitution mutation in CD2 in position 287, and the series of substitution mutations in ND detected in human NANOGP1 isoforms, are all unlikely to be detrimental to NANOGP1 transactivation. However, the 2ΔW deletion found in NANOGP1-2 may have altered its dimerisation activity (Figure 3.11).

Finally, the presence of a fully conserved homeodomain in NANOGP1 suggests identical DNA binding properties to those of NANOG.

Taken together, NANOGP1 substitution mutations and small deletions, described in this thesis, are unlikely to critically interfere with the conserved NANOG-like properties in NANOGP1. The ability of NANOGP1-2 to dimerise would need to be investigated in more detail. A possible future experiment could test protein dimerisation ability of recombinant NANOGP1-2 in insect cell culture, with further protein extraction and AKTA purification. A similar experiment was performed in this thesis using NANOGP1-1 (described in Chapter 4).

The last and the most prominent difference between NANOGP1 and NANOG proteins is a large, 39 amino acid deletion in the NANOGP1 N-terminus, conserved between all three variants, and which does not overlap with any currently known functional domains. However, it could still have an effect on NANOGP1. For instance, one possible scenario originates from the NANOG bivalency model (Chang et al., 2009). Chang and colleagues proposed that instead of being functionally redundant, the

NANOG N-terminus is in fact responsible for mediating transcriptional interference and could attract co-repressors of cell differentiation, opposing the transactivation role mediated by the C-terminus. Indeed, deleting the whole length of the ND, 122 amino acids, led to a significant increase in transactivation activity mediated by the remaining HD-CR1-WR-CR2 domains. This was demonstrated by two separate research groups, in Chang et al., 2009 and Oh et al., 2005, by testing ΔND-HD-CR1-WR-CR2 luciferase assay constructs in non-human primates and human cell lines. This, however, has not been tested directly in hPSCs.

According to our predicted isoform structure, NANOGP1 only has a partial deletion of the N-terminal domain, 39 amino acids, and not a full truncation of the N-terminal as described above. This means that, if the bivalent model applies to NANOG in hPSCs, then NANOGP1 could have partially maintained co-repressor binding domains in its remaining N-terminus, but with potentially altered protein interaction properties. Therefore, even though the N-terminal deletion is not likely to cause functional decay of NANOGP1, it could still affect its transactivation activity. The key question here is which part of the 122 aa is important for mediating the repressive effects and whether it overlaps with the first 39 aa, missing from NANOGP1.

Another potential function for the NANOG N-terminus was suggested by Oh and colleagues, who hypothesised that it could be required for post-translational protein modifications, such as phosphorylation and ubiquitination. Indeed, human NANOG contains 11 phosphorylation sites in its N-terminal domain (Brumbaugh et al., 2014; Ho et al., 2012; Wang et al., 2019; Xie et al., 2014), and one of them, Y35, is absent from the predicted NANOGP1 protein isoforms as it falls within the N-terminal deletion. The consequences of deleting Y35 in NANOG in hPSCs are currently unknown. However, in cancer cells, substituting Y35 with T inhibited its interaction with a tyrosine kinase FAK, involved in regulation of cell migration, and ultimately prevented the formation of the expected filopodia-formation phenotype in *NANOG* overexpression studies (Ho et al., 2012). This study showed that another phospho-site mutation, Y174F, present in CD1, had the same effect in *NANOG* overexpression assays; coincidentally, Y174 also falls within a NANOGP1-1 deletion, but this time in CD1. This could imply that in hPSCs, the absence of just one or both of these phosphorylation sites in NANOGP1 could also be sufficient to impair one or more downstream phosphorylation pathways.

Four other NANOG phosphorylation sties, Ser/Pro-52, Ser/Pro-65, Ser/Pro-71 and Thr/Pro-287, which are conserved among mouse, human and rat *Nanog* orthologs, were shown to supress ubiquitination when phosphorylated and, therefore, to stabilise mouse NANOG (Moretto-Zita et al., 2010). This study demonstrated that in mouse, the four sites interact with prolyl isomerase PIN1, which in turn stabilised NANOG by supressing its ubiquitination. Mutating those phosphorylation sites led to inhibition of their interaction with PIN1 and resulted in impaired cell self-renewal and increased

NANOG protein degradation. More specifically, replacing Ser with Ala in just two phospho-sites out of four was sufficient to disrupt NANOG-PIN1 interaction. The inability of NANOG to interact with PIN1 (also shown in PIN1 direct inhibition) led to its instability and level reduction.

A study by Ramakrishna and colleagues, performed in primed hPSCs, demonstrated an opposite result: deletion of all amino acid residues in NANOG positions 47-72 (PEST motif; contains conserved Ser/Pro-52, Ser/Pro-65, Ser/Pro-71) led to the reduction of NANOG proteasomal degradation (Ramakrishna et al., 2011). They, however, did not test individual phospho-sites.

Coincidentally, out of the four phosphorylation sites mentioned above, one is mutated in NANOGP1 (Ser/Pro-65 -> Ser/His-65). Unfortunately, Moretto-Zita et al., 2010 did not analyse the consequences of mutating Ser/Pro-65. Instead, they used Ser/Pro-71 site as an example for a single phosphorylation site mutation, which caused partial impairment of NANOG turnover. This study mutated Ser/Pro-71 by replacing Ser with Ala, and did not test the role of proline in +1 position, which is mutated in NANOGP1. Therefore, it is not possible to use that experiment as direct evidence that another single phosphorylation site mutation would have the same effect on NANOGP1. It is reasonable to assume, though, that mutating the +1 proline in the phosphorylation site could affect the downstream processes. Indeed, +1 position prolines do not get phosphorylated themselves, but are required for correct phosphorylation reactions to occur, as was demonstrated in a study of another mammalian kinase, KIS (Maucuer et al., 2000). Taken together, a single Ser/Pro-65 phosphorylation site mutation could have an effect on NANOGP1 turnover in hPSCs; however, this speculation is based on secondary evidence and needs to be tested experimentally. A potential future study could test this by studying NANOGP1 interaction with proteins binding PEST motif. If any effect is seen when compared to NANOG, a rescue protein variant could be created (i.e., by replacing the NANOGP1-specific mutations with NANOG WT nucleotides) and tested for the ability to restore the impaired/changed phenotype.

In summary, based on our proposed protein structure for NANOGP1, the NANOGP1 N-terminal deletion, as well as NANOGP1-1 CD1 deletion could hypothetically affect its downstream phosphorylation pathways involved in ubiquitination and protein turn over. Similarly, the N-terminal deletion could cause altered co-repressor binding. However, both of these models would require additional protein-protein interaction assays in hPSCs to be confirmed or rejected. Other mutations found in the CDS of *NANOGP1* are less likely to have any significant effect on transactivation, dimerisation, or DNA binding, with the exception of the 2ΔW amino acid deletion that would need to be tested in future studies.

3.3.3 Divergent expression patterns of *NANOGP1* and *NANOG* in human embryos and hPSCs

One of the major findings of this chapter is that *NANOG* and *NANOGP1* have overlapping but distinct expression patterns in the developing embryo. *NANOGP1* exhibited a more temporally restricted expression window and was first detected in ICM, while *NANOG* was first detected at the 8-cell stage, thereby demonstrating a more prolonged duration of expression. Strikingly, expression of the two duplicates was also detected in the extraembryonic lineages where their patterns differed as well. *NANOGP1* RNA was present in a subpopulation of primitive endoderm cells, while some trophectoderm cells expressed *NANOG* transcripts. Notably, both of the duplicates exhibited noticeably lower transcriptional levels in the extraembryonic lineages compared to the epiblast.

Studying NANOGP1 in the context of primitive endoderm was beyond the time scale of this project. However, in future studies, *NANOGP1* overexpression and expression downregulation assays in hPSC-to-primitive endoderm differentiation protocol (Linneberg-Agerholm et al., 2019) could help to reveal *NANOGP1*-specific functions, if any are present. Researching its potential activity in primitive endoderm would be particularly interesting, since primitive endoderm does not express *NANOG*; *NANOGP1* would therefore be investigated as 'individually' as possible, without the risk of having *NANOG* function masking/affecting it.

NANOG protein expression in human trophectoderm, but not in primitive endoderm, has already been shown by immunostaining (Cauffman et al., 2009). In fact, expression of NANOG and GATA4 were found to be mutually exclusive in the developing human epiblast and primitive endoderm (Roode et al., 2012), similar to *Nanog* and *Gata6* in mouse epiblast and primitive endoderm, respectively (Chazaud et al., 2006; Plusa et al., 2008; Rossant et al., 2003). Therefore, detecting *NANOGP1* transcripts in some primitive endoderm cells was unexpected. One of the possible explanations for this could be that *NANOGP1* may have an impaired 'ICM -> epiblast/primitive endoderm' resolving mechanism, and therefore remains in both epiblast and primitive endoderm. Alternatively, *NANOG* and *NANOGP1* could have different developmental timings for becoming restricted to epiblast. In this way, our observations could be interpreted as *NANOGP1* being downregulated in primitive endoderm, but not yet being completely silenced. The role of NANOG expression in trophectoderm and whether *NANOGP1* has any unique functions in primitive endoderm both await future investigations.

In addition to the non-identical embryo expression patterns, *NANOGP1* is also expressed at a significantly lower level in the primed hPSCs compared to *NANOG*. This result does not imply that *NANOGP1* is a naïve-specific marker, since some transcripts were still present in the primed cell lines, but it serves as further evidence for a more temporally restricted expression of *NANOGP1*. Furthermore, similar behaviour is seen in the developing epiblast of cultured embryos, where

NANOGP1 expression drops by Day 14, while *NANOG* is highly expressed for a longer period of time and stays upregulated on Day 14. Collectively, these findings present *NANOGP1* as a slightly more transient factor than *NANOG*, implying that the potential function of the former could be more stage-specific. This could mean that high expression levels of *NANOGP1* are either not required in late epiblast, and, similarly, in primed hPSCs, or that *NANOGP1* has to be downregulated there not to interfere with some, currently unknown, process(es).

The reason why *NANOGP1* expression level is lower than that of *NANOG* in the primed hPSCs and late epiblast could be related to the duplicated putative regulatory regions, found upstream of *NANOGP1*. It was clear that the *NANOGP1* promoter and proximal enhancer regions had much stronger active markers in the naïve cells compared to primed. Identification of the precise mechanism, responsible for establishing the expression difference between the naïve and primed state remains an interesting topic and awaits future investigation. Overall, the putative promoter regions of *NANOG* and *NANOGP1* exhibited mostly similar pluripotency factor binding and histone modification profiles in naïve and primed hPSCs.

3.3.4 Human naïve pluripotency and pseudogene expression

Finally, I uncovered that in addition to *NANOGP1*, other key pluripotency factors, *POU5F1* and *DPPA3,* have highly expressed pseudogenes in naïve hPSCs, namely, *POU5F1P1-3* and *DPPA3P2*, respectively. Thus far, these pseudogenes have mostly been studied in human cancers and adult stem cells (Cheng et al., 2018; Jez et al., 2014; Mostert et al., 2000; Poursani et al., 2016; Schneider et al., 2002), with little knowledge of their properties in hPSCs. For example, *DPPA3P2* was found to encode a protein, STELLAR, and is a marker of testicular cancer and developing germ cells (Bouckenheimer et al., 2018; Cheng et al., 2018; Kossack et al., 2013; Oosterhuis and Looijenga, 2005), but was not studied in the context of human pluripotency. *POU5F1* pseudogenes have received more attention and thus were analysed in more depth. For instance, Jez et al., 2014 and Poursani et al., 2016 show that *POU5F1* pseudogenes exhibit expression in multiple somatic, cancer and even primed hPSCs lines. Judging by the sequence conservation, *POU5F1P1* and *POU5F1P3* were initially assumed to possess an ancestral function as its CDS is almost identical to that of *POU5F1*, while *POU5F1P2* is truncated and had only partially preserved its last exon while all the others are missing (Poursani et al., 2016). Quite unexpectedly though, *POU5F1P1* and *POU5F1P3* could not be translated into a protein, even in the cell lines where they are highly expressed at the mRNA level; the reason for this remains unclear (Poursani et al., 2016). Taken together, high expression of *POU5F1* pseudogenes in naïve hPSCs, reported in this thesis, could mean that they have developed a naïve-specific role, with *POU5F1P1* and *POU5F1P3* being the most likely candidates, likely as lncRNA, as hypothesised by Prousani and

colleagues. Their role, as well as a potential function of the *DPPA3* pseudogene, would require further investigation.

Analysing gene expression data from naïve hPSCs, I identified expression of two processed *NANOG* pseudogenes, *NANOGP4* and *NANOGP8*. According to the published literature, *NANOGP8* and *NANOGP4* are also expressed in several human cancer lines, however, whether they have any functions remains unclear (Ambady et al., 2010; Palla et al., 2014; Zhang et al., 2006). *NANOGP8* is one of the most conserved processed *NANOG* duplicates (Booth and Holland, 2004) and additionally, it was shown to be able to mediate somatic-to-primed reprogramming in mouse and human fibroblasts (Palla et al., 2014). The latter was likely possible due to the complete conservation of its DNA-binding homeodomain, which is known to be sufficient in mediating the reprogramming function of *NANOG*, as shown in Theunissen et al., 2011. *NANOGP4* CDS is less similar to that of *NANOG*: the pseudogene was predicted to encode three truncated reading frames, with only one of them encoding the functionally important homeodomain, whereas the N-terminus and C-terminus encoding regions are either completely missing or significantly shortened in all three ORF versions (Booth and Holland, 2004). Therefore, provided that the NANOGP4 and NANOGP8 are in fact expressed by naïve hPSCs, I speculated that *NANOGP8* could have the potential to contribute to naïve pluripotency and to have a conserved, *NANOG*-like role. *NANOGP8* and *NANOGP4,* however, are transcribed at a lower level than *NANOG* and *NANOGP1* and presumably lack regulatory sequences, so it is not clear whether these pseudogenes are expressed in a regulated way. *NANOGP4* would also most likely have diverged or have a limited functional conservation, if it has any role in the naïve context at all.

Among other highly expressed pseudogenes in naïve hPSCs, I identified copies of ribosomal genes and genes related to mitochondrial function that are involved in protein and energy synthesis, respectively. Interestingly, the high abundance of ribosomal and mitochondrial pseudogenes in the human genome attracted the attention of scientists almost three decades ago, and since then they have been hypothesised to contribute to their ancestral function, as well as tissue-specific expression, carcinogenesis and/or ageing (Dutta et al., 2011; Mamoor, 2020; Shay and Werbin, 1992; Tonner et al., 2012; Woischnik and Moraes, 2002; Yuan et al., 1999; Zhang, 2002). However, very few functional studies have been performed and this topic remains poorly understood. It is possible that in naïve hPSCs, as well as in other cell types where such pseudogenes are highly expressed, they still perform their ancestral function and/or ensure that such crucial processes as energy and protein synthesis can still occur if mutations disrupt other copies involved in the process. In my opinion, this reasoning could also explanation the high expression of pluripotency factor pseudogenes in naïve human cells; this and other hypotheses for *NANOGP1* expression in naïve pluripotency are discussed in more detail in Chapter 6.

To conclude, this chapter described several important aspects of *NANOGP1* organisation and structure. *NANOGP1* was presented as a conserved copy of *NANOG*, both within its CDS and its putative regulatory elements. Importantly, *NANOG* and *NANOGP1* had very similar overlapping expression patterns both *in vivo* and *in vitro*. NANOGP1 protein had several mutations, deletions and substitutions, that could in theory affect its activity, but none were located in the important domains. Based on this knowledge, I hypothesise that *NANOGP1* has the potential to possess both conserved and/or novel functions and properties. This is subject, however, to one important condition: whether *NANOGP1* can be translated into a protein. Indeed, high mRNA expression levels and conservation do not ensure that a pseudogene will be translated into a stable protein form, like in the case of *POU5F1P3*, discussed above. This crucial question is addressed in the next chapter.

4 Characterising NANOGP1 protein: expression in naïve hPSCs, chromatin binding and dimerisation

4.1 Introduction

4.1.1 Background

In this chapter I address a key question that has not been answered until now: can *NANOGP1* RNA be translated into a protein in the context of human pluripotency?

To answer this, a correct system for the protein expression must first be chosen. *NANOGP1* is highly similar to *NANOG* at the sequence level and, as was shown in the previous chapter, their expression patterns also overlap. Hence, I hypothesised that NANOGP1 and NANOG protein expression patterns would also at least partially overlap. Finally, based on the NANOGP1 protein structure prediction analysis (Section 3.2.5), I also suggested that if NANOGP1 existed, it would likely have had similar properties to NANOG.

What do we know about the NANOG protein?

NANOG protein is expressed in the human embryo and by hPSCs. In cell culture, NANOG can be detected in primed and naïve hPSCs, as well as during the hPSC formative capacitation (Guo et al., 2017; Hyslop et al., 2005; Rostovskaya et al., 2019; Theunissen et al., 2014). In human embryo, NANOG protein expression can be seen in the ICM and preimplantation epiblast, while some reports also mention its presence in the developing trophectoderm (Cauffman et al., 2009; de Paepe et al., 2013; Gerri et al., 2020; Guo et al., 2016; Kimber et al., 2008; Kuijk et al., 2012; Niakan and Eggan, 2013; Roode et al., 2012). Finally, its expression can also be detected in the developing germ cells (Gkountela et al., 2015; Kerr et al., 2008; Kuijk et al., 2010).

Several key studies on mouse pluripotency showed that NANOG functions as a dimer, and can also dimerise with other transcription factors via its tryptophan-rich domain (Gagliardi et al., 2013; Mullin et al., 2008; Mullin et al., 2017; Mullin et al., 2020; Torres and Watt, 2008; Wang et al., 2008a). The same ability to form homodimers by human NANOG has also been demonstrated in Chang et al., 2009.

The chromatin binding profile of human NANOG has been investigated extensively both in primed and naïve hPSCs; examples of the NANOG chromatin binding studies are Barakat et al., 2018, Boyer et al., 2005, Chovanec et al., 2021 and Gifford et al., 2013. A very recent study by Chovanec and colleagues presented a comparison of NANOG binding profiles between the two pluripotent states. This study demonstrated that in hPSCs NANOG can be found not only in promoter regions, but also in naïve, primed and shared (between naïve and primed) enhancers and superenhancers, which NANOG co-binds with the other two key pluripotency factors, OCT4 and SOX2.

Considering that *NANOGP1* exhibits high RNA expression in naïve hPSCs, where *NANOG* is also highly expressed and is capable of binding chromatin in its protein form, I chose naïve cell culture as the most optimal cell system to investigate potential NANOGP1 protein expression. I also hypothesised that NANOGP1 protein, if it indeed exists, is capable of binding chromatin, since the DNA-binding homeodomain was predicted to be conserved between NANOG and NANOGP1 at the amino acid level (Section 3.2.5). I also suggested that since the two homeodomains were predicted to be completely identical in their protein-encoding DNA sequence, the binding profiles of the two proteins are also likely to be similar, if not identical. Finally, I proposed that NANOGP1 protein is capable of dimerising as its predicted tryptophan-rich region is intact in two predicted isoforms, and bears only one small deletion in the third isoform (Section 3.2.5).

In this chapter, I discover that *NANOGP1* pseudogene indeed can be translated into a stable protein in naïve hPSCs. I demonstrate that, as expected, NANOG and NANOGP1 share a relatively small number of chromatin binding regions. I also uncover 84 unique NANOG-independent chromatin binding regions, although this data requires further clarification. Finally, I demonstrate that, in line with my hypothesis, recombinant NANOGP1 is capable of forming homodimers as well as heterodimers with NANOG *in vitro*.

In summary, this chapter shows for the first time that endogenous *NANOGP1* is translated into a stable protein. Additionally, it describes that at least some properties of the protein are conserved, such as the ability to form dimers and the existence of shared chromatin binding domains with NANOG sites.

This chapter contains results obtained in collaboration with a scientist from Babraham Institute Bioinformatics facility. Here, ChIP-sequencing analysis was performed with help from Dr. Christel Krueger (Figures 4.15-4.24). All the other figures in this chapter are entirely my own work.

All figures in this chapter were adapted and/or made by myself.

4.1.2 Aims

1. Identify whether *NANOGP1* can be translated into a stable protein in naïve hPSCs.
2. If NANOGP1 protein exists, study its chromatin binding in naïve hPSCs.
3. Investigate whether recombinant NANOGP1 protein can form homodimers, as well as heterodimers with NANOG.

4.2 Results

4.2.1 Characterising the expression of epitope-tagged NANOGP1 protein in naïve hPSCs

In this section, I describe generation of naïve hPSC lines in which the endogenous *NANOGP1* locus is epitope-tagged to produce a tagged NANOGP1 protein. Epitope tagging was chosen by me because there were no currently available antibodies capable of distinguishing NANOGP1 from NANOG due to their high structural similarity. In addition to the epitope-tagged NANOGP1, I also attempted to tag *NANOG* gene in a separate cell line, as a control experiment.

To my knowledge, until this study, there have been no published reports in which endogenous gene tagging was performed directly in naïve hPSCs. As a result, I had to ensure that each step of the experiment was optimised, which is detailed below.

In order to introduce an epitope tag, I utilised two different CRISPR-mediated genome editing approaches, namely, those using Cas9 and Cas12a proteins (Deltcheva et al., 2011; Hsu et al., 2013; Zetsche et al., 2015). These two proteins recognise different PAM sequences (NGG and TTTV, respectively) and create different types of ends after the DNA cut (blunt and sticky, respectively), therefore offering more targeting options and increasing the possibility of successful gene editing. Additionally, Cas12 only requires a crispr RNA (crRNA) sequence for targeting, unlike Cas9 which also needs trans-activating crispr RNA (tracrRNA) (Deltcheva et al., 2011; Hsu et al., 2013; Zetsche et al., 2015). Firstly, I designed crRNA molecules that would guide the CRISPR complex to their target site. Initially, I aimed to tag both the N- and C-terminal ends, and therefore designed crRNAs that would individually target both 5' and 3' ends of a gene. crRNA cut sites were positioned <30bp away from the anticipated knock-in site for increased HDR efficiency (Quadros et al., 2017; Renaud et al., 2016). I also designed single strand oligonucleotides (ssODNs) that contained an epitope tag sequence flanked by homology arms, serving as homology directed repair (HDR) templates. In this experiment, I chose using ssODN over double-stranded templates since they exhibit higher HDR efficiency and are more likely to generate a successful knock-in edit (Codner et al., 2018; Miura et al., 2015; Yoshimi et al., 2016). To ensure that the genome is being cut only once by the CRISPR-Cas complex, the PAM sites and crRNA-complementary regions within the HDR templates had single-nucleotide mutations. All ssODN templates also had phosphorothioate bonds connecting the last three nucleotides on each end in order prevent the DNA from degradation (Papaioannou et al., 2009).

For the ssODN design, I chose two different tags, V5 and 3xFLAG, which are widely used in the field in such methods as immunostaining, immunoprecipitation, ChIP and ChiP-seq (Kidder et al., 2011; Lobbestael et al., 2010; Wang, 2009). Ideally, my plan to have two differentially tagged cell lines would have ensured that the potential signal detected by only one tag is not an artefact and has a biological

meaning. Additionally, in case one of the tagging strategies was to fail at any stage, I would have had a back-up option remaining.

Here I designed reagents for 5'-*NANOGP1* and 3'-*NANOGP1* tagging to be used with CRISPR-Cas9, as well as for a 5'-*NANOGP1* strategy to be used with CRISPR-Cas12a; similarly, I designed reagents for tagging *NANOG* using CRISPR-Cas9 (Figure 4.1). For simplicity, all diagrams in this section depict *NANOGP1-1* CDS, while the designed reagents were suitable for tagging all three isoforms.

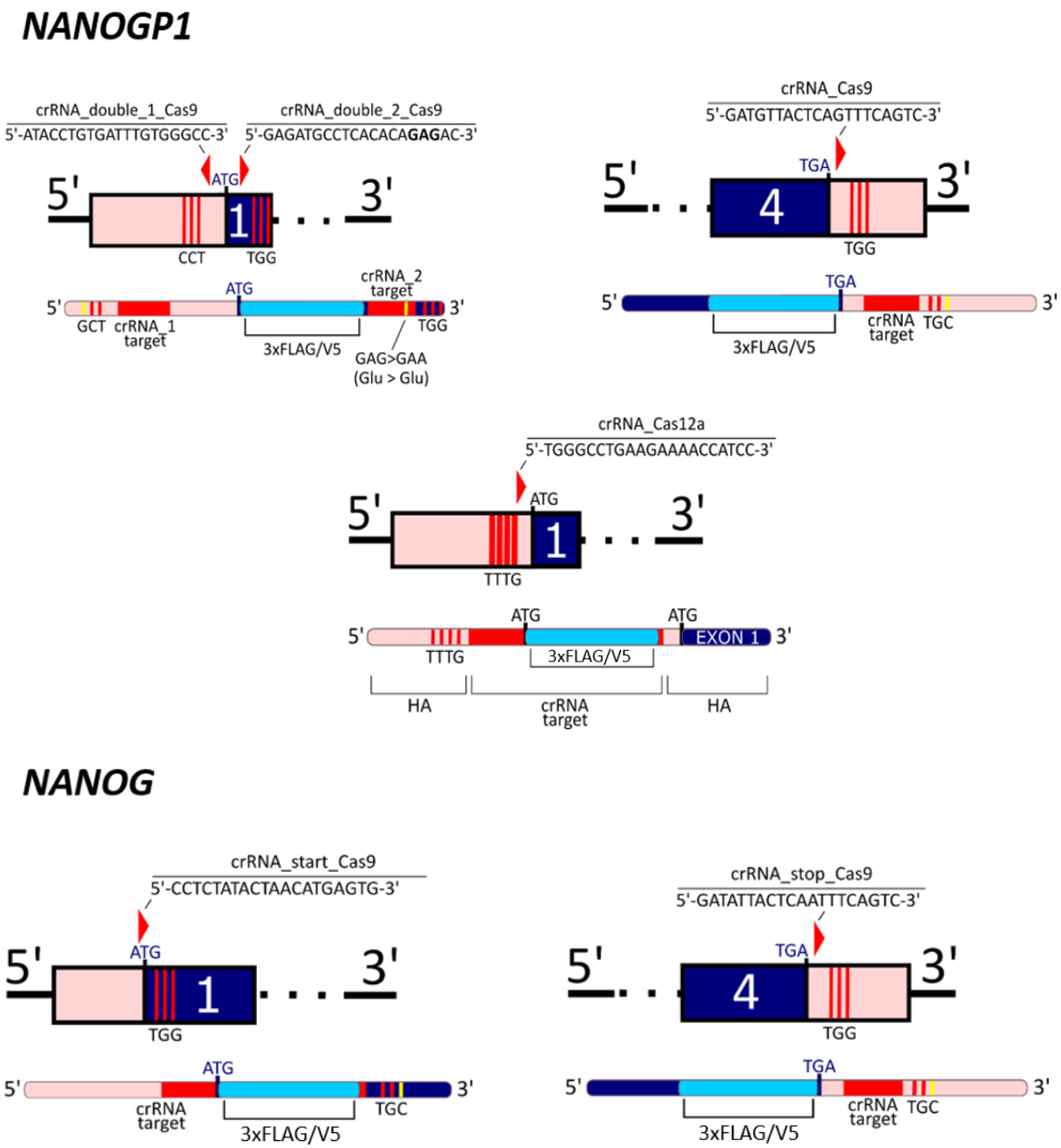


Figure 4.1 Diagram showing crRNA and ssODN template designs for the *NANOGP1* and *NANOG* epitope tagging experiments. Three and two CRIPSR-Cas gene editing strategies are shown for *NANOGP1* and *NANOG*, respectively**.** Each experimental design is described in a separate panel which contains a crRNA schematic (top) and a matching ssODN template (bottom). '5'' and '3'' label the 5'

and 3’ ends of the nucleotide sequence, respectively. Each crRNA schematic includes its 5’-3’ target sequence, and shows its positioning in relation to the gene of interest and PAM. crRNA target sequences are indicted by red arrows (top), and by red blocks (bottom). 5’-UTR and 3’-UTR are shown as pink blocks. Exons are shown as dark blue blocks, labelled with their number (exon 1, exon 4). Red vertical lines indicate PAM. Yellow vertical lines indicate single nucleotide mutations in the PAM or crRNA target sequence. 3xFLAG and V5 tag sequences are shown as light blue blocks. ATG – start codon. TGA – stop codon. ‘Start’, ‘stop’ – crRNA target is located near the start or stop codon, respectively. ‘Double’ – experimental design includes two crRNAs. HA – homology arm. Glu – glutamate.

After the design was complete, the double stranded DNA cutting efficiency of each crRNA was tested using an *in vitro* CRISPR/cut method (NEB, M0386). This method required combining PCR-generated DNA molecules (cutting templates) with their respective CRISPR complex components (pre-assembled crRNA:tracrRNA:Cas9, or crRNA:Cas12a). First, I designed primer pairs that would amplify crRNA target DNA regions (cutting templates) from *NANOGP1* and *NANOG* genomic sequences (Figure 4.2).

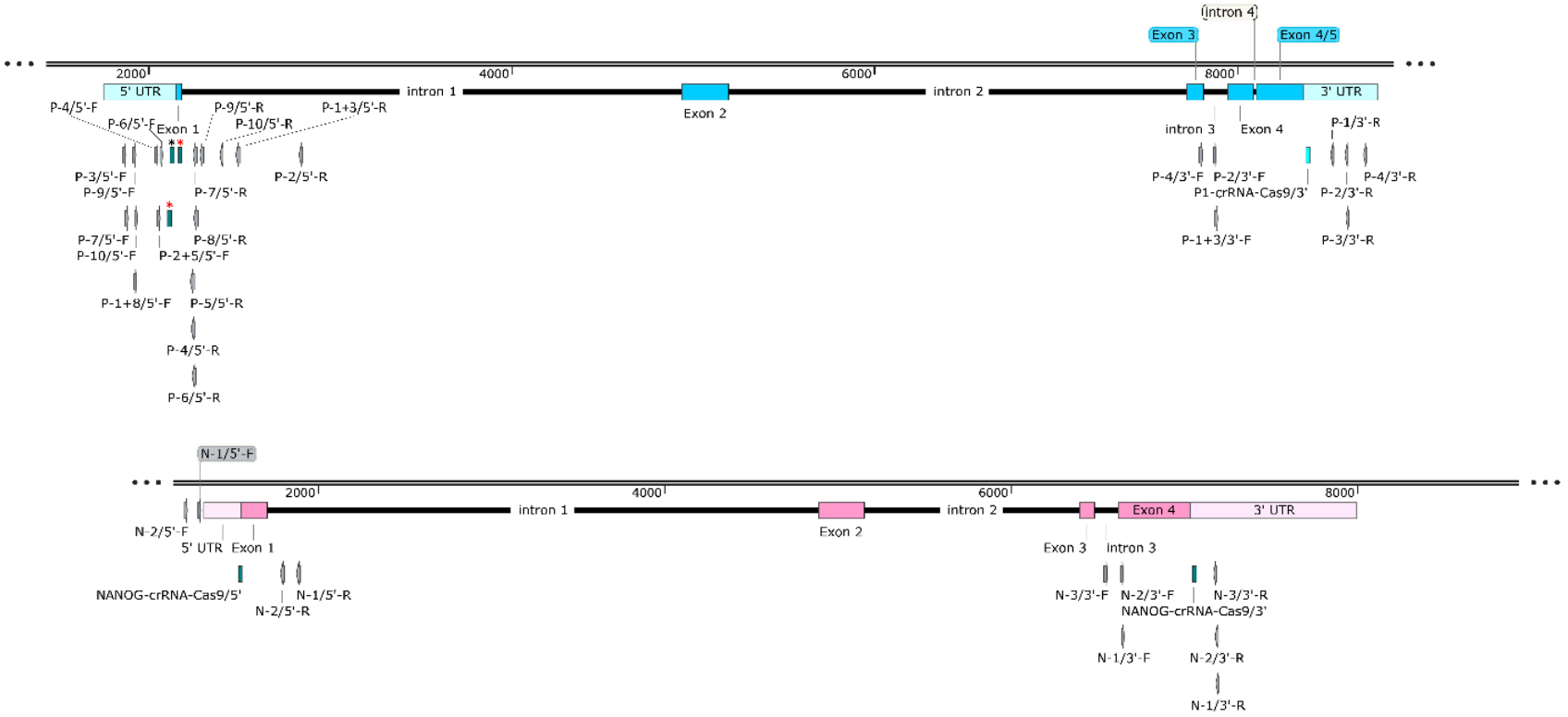


Figure 4.2 Sequence map showing primers used for testing crRNA cutting efficiency. *NANOG* sequence is shown on top (exons and UTRs are light blue), *NANOGP1* is shown at the bottom (exons and UTRs are pink). Scale, bp.

Each primer name contains the following details:

1) a letter, indicating whether it corresponds to *NANOG* or *NANOGP1* (N – *NANOG*, P - *NANOGP1*)
2) primer pair index number (i.e., 1)
3) which end of the gene/pseudogene sequence it binds (3' or 5').
4) primer orientation (F/R).

If a primer was used in more than one primer pair, it is indicated by its index number (i.e., 1+3).
Black asterisk – NANOGP1-crRNA-Cas12a/5'. Red asterisks – NANOGP1-crRNA-1-Cas9/5' and NANOGP1P1-crRNA-2-Cas9/5'.

I then tested the primer pairs and selected those that produced PCR products of the expected length and that were subsequently validated by Sanger sequencing (using the 'forward' primer from each primer pair). As a result, one primer pair was selected for each tagging experiment (**Figure 4.3**). See primer sequences in Table 2.7 Primers used for genotyping, cloning validation and Sanger sequencing. F, R – forward and reverse primer orientation.

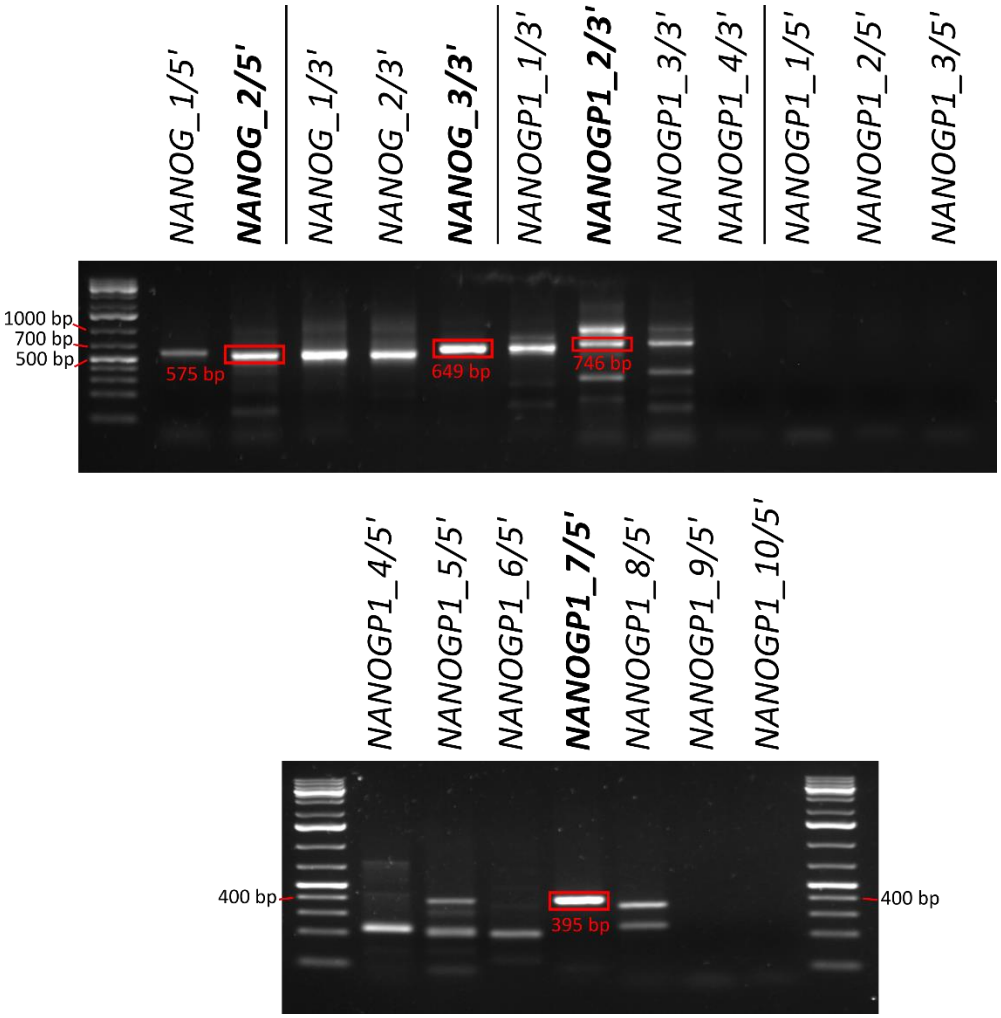


Figure 4.3 Gel electrophoresis images showing results of the genotyping primer screen. Names of the PCR products indicate whether they were used for *NANOGP1* or *NANOG* genotyping screen, as well as the primer pair index number and whether they bind the 5' or 3' end of the gene/pseudogene sequence**.** Names of the PCR products that passed selection are shown in bold. The red rectangle indicates the band of the expected size that was excised and sequenced. Expected band size is in red: *NANOG*_2/5' – 575 bp, NANOG_3/3' – 649 bp, *NANOGP1*_2/3' – 746 bp, NANOGP1_7/5' – 395 bp.

After suitable DNA templates were obtained, the crRNA *in vitro* cutting protocol had to be tested using an efficient crRNA as a positive control. For this purpose, I used crRNA designed to target *RING1B,* which was previously designed and validated by a different lab member in their project. This involved combining a purified *RING1B* PCR product containing the crRNA target sequence, with *RING1B* crRNA, preassembled with tracrRNA and Cas9. After the protocol was shown to be working with the control *RING1B* reagents, I validated the cutting ability of crRNA molecules designed in my

project using the same strategy (Figure 4.4). All crRNA molecules were shown to cut their respective PCR-amplified regions. Notably, Cas12a *NANOGP1* crRNA was the most efficient, not leaving any original DNA product uncut.

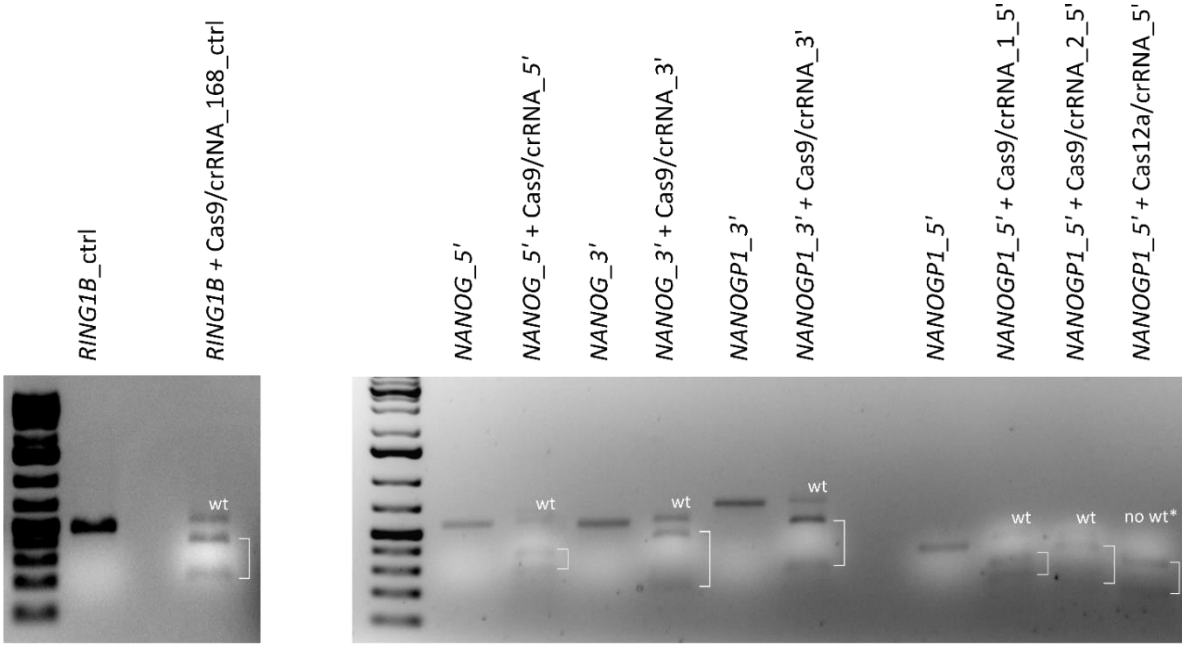


Figure 4.4. Gel electrophoresis images showing results of the *in vitro* crRNA efficiency assay. A positive control (left) and the *NANOG/NANOGP1* crRNA screen (right) are shown. Lane names indicate what PCR product, crRNA and Cas protein were used. '5'' and '3'' indicate 5' and 3' ends of a nucleotide sequence. Wt – uncut PCR product. No wt* - no uncut PCR product remained. White bracket indicates two products of the crRNA-mediated cutting. Each lane contained ~25-40 ng of DNA (a full volume of one CRISPR/Cas reaction, or the same amount of uncut control), therefore the bands produced in the *in vitro* cutting reactions are fainter than the uncut control ones.

To further analyse and quantify the efficiency of DNA cutting, I performed a similar crRNA test *in vivo*. Human primed H9 hPSCs were transfected with the *NANOGP1*_5', *NANOGP1*_3', *NANOG*_5' and *NANOG*_3' CRISPR reagents: Cas9 or Cas12a protein, preassembled with its corresponding crRNA:tracrRNA (Cas9) or crRNA (Cas12a) molecules. HDR templates were not used since this assay was only testing the ability of crRNA to cut DNA *in vivo*. After transfection, the cells were expanded and used for genomic DNA extraction. The DNA was further used to generate PCR products with the four selected primer pairs, described above and in **Figure 4.3**. *NANOG*_5', *NANOG*_3' and *NANOGP1*_3' PCR products were sequenced (using primer *NANOGP1*_7/5'-F) and analysed using Genewiz Amplicon EZ NGS service, which analyses the ratio and composition of edited:unedited amplicons in a sample. *NANOGP1*_5' PCR products were sequenced and analysed with CRISPResso online tool (Clement et al., 2019), which uses a similar approach as the Genewiz service. The efficiency of crRNA cutting was assessed by quantifying the amount of sequencing reads that contained deletions. The presence of a deletion indicated that DNA cutting took place and was subsequently repaired by the NHEJ mechanism (Cong et al., 2013; Mali et al., 2013). As a result, I concluded that

only one reaction had generated high efficiency of cutting (~70%) was suitable for epitope tagging HDR; this reaction was mediated by the *NANOGP1*_5' CRISPR/Cas12a reagents (Figure 4.5, Figure 4.6, Table 4.1).

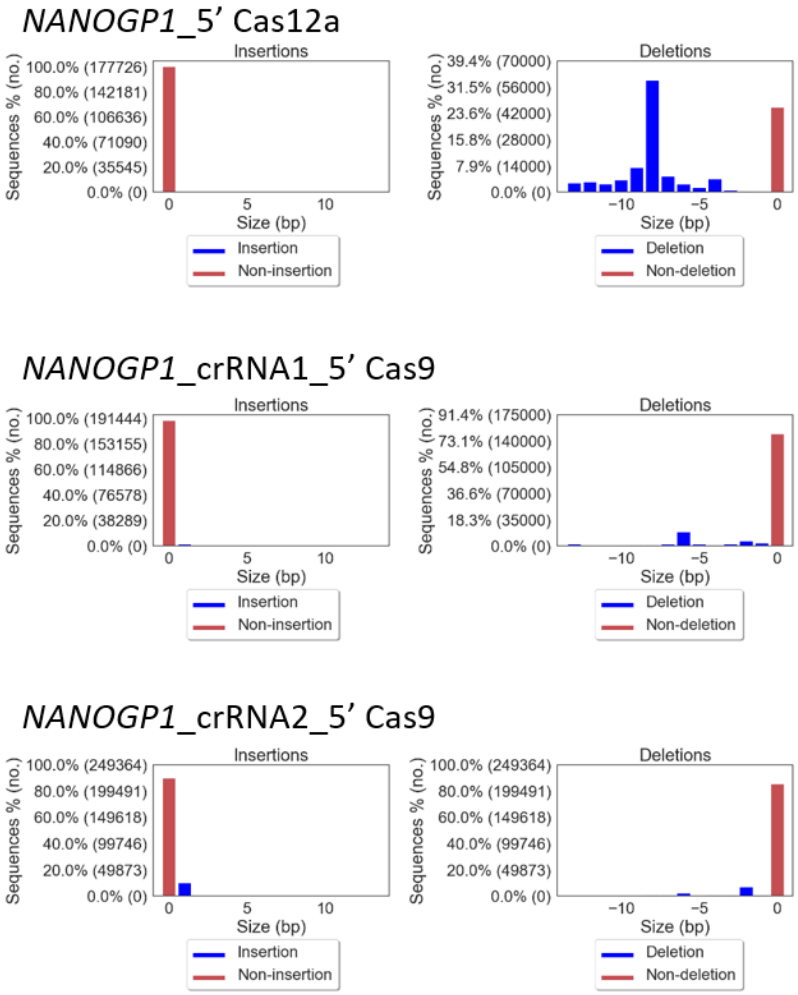


Figure 4.5 Histograms showing indel profile across sequenced PCR amplicons in the *in vivo* crRNA cutting assay. Y-axis shows percentage of sequences/reads as well as the total number of reads generated for each type of analysis. X-axis indicates changes in the amplicon length caused by a mutation, bp. Histograms were generated as part of CRISPResso report.

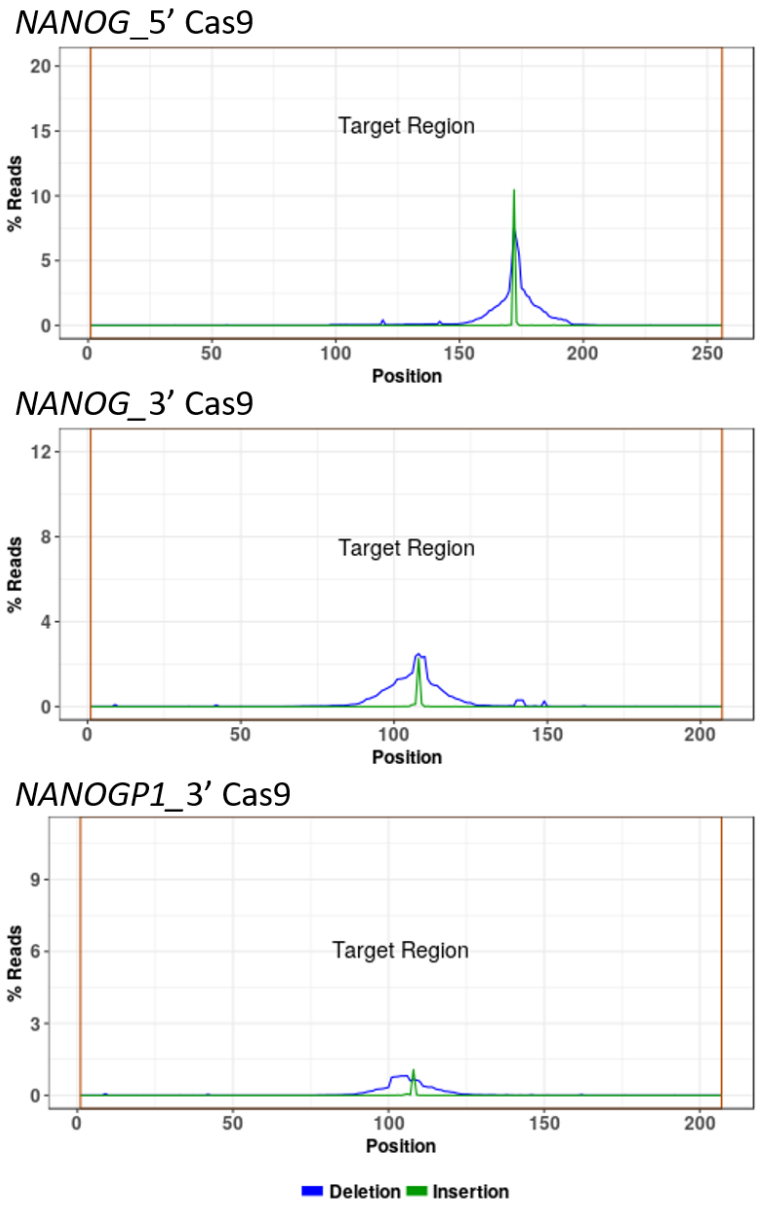


Figure 4.6 Graphs showing indel profile across sequenced PCR amplicons in the *in vivo* crRNA cutting assay. Y-axis – percent of reads generated by Amplicon EZ sequencing. X-axis – position of deletion/insertion within the amplicon. Vertical brown lines indicate the borders of the amplicon. Graphs were generated as part of Amplicon EZ report.

Table 4.1 Table showing *in vivo* crRNA cutting assay results

crRNA	*In vivo* DNA cutting efficiency (% of amplicons with deletions per sample)
*NANOGP1*_5' Cas12a	70%
*NANOG*_5' Cas9	30%
*NANOGP1*_crRNA1_5' Cas9	18%
*NANOGP1*_crRNA2_5' Cas9	15%

*NANOG*_3' Cas9	3%
*NANOGP1*_3' Cas9	9%

These crRNA cutting efficiency results were in line with the *in vitro* test, described above (FIG), where the *NANOGP1*_5' Cas12a crRNA did not leave any uncut DNA behind, unlike when using all other test crRNA molecules.

A detailed composition of the *NANOGP1*_5' Cas12a cut site and all deletions found in the sequencing sample are shown in Figure 4.7. In the bottom part, this figure shows all variants of deleterious mutations created in the target DNA by *NANOGP1*_5' Cas12a crRNA, once again demonstrating high efficiency of the crRNA reagent. The top part of the figure contains a summary of all deleterious mutations in the sample, showing that these reagents have a high enough efficiency for the knock-in reaction to occur (based on the high cumulative amount of the deleterious mutations in the predicted cut site). This figure also highlights that the cut site is in the close vicinity of the *NANOGP1* ATG codon, implying that it will be beneficial for the knock-in HDR reaction.

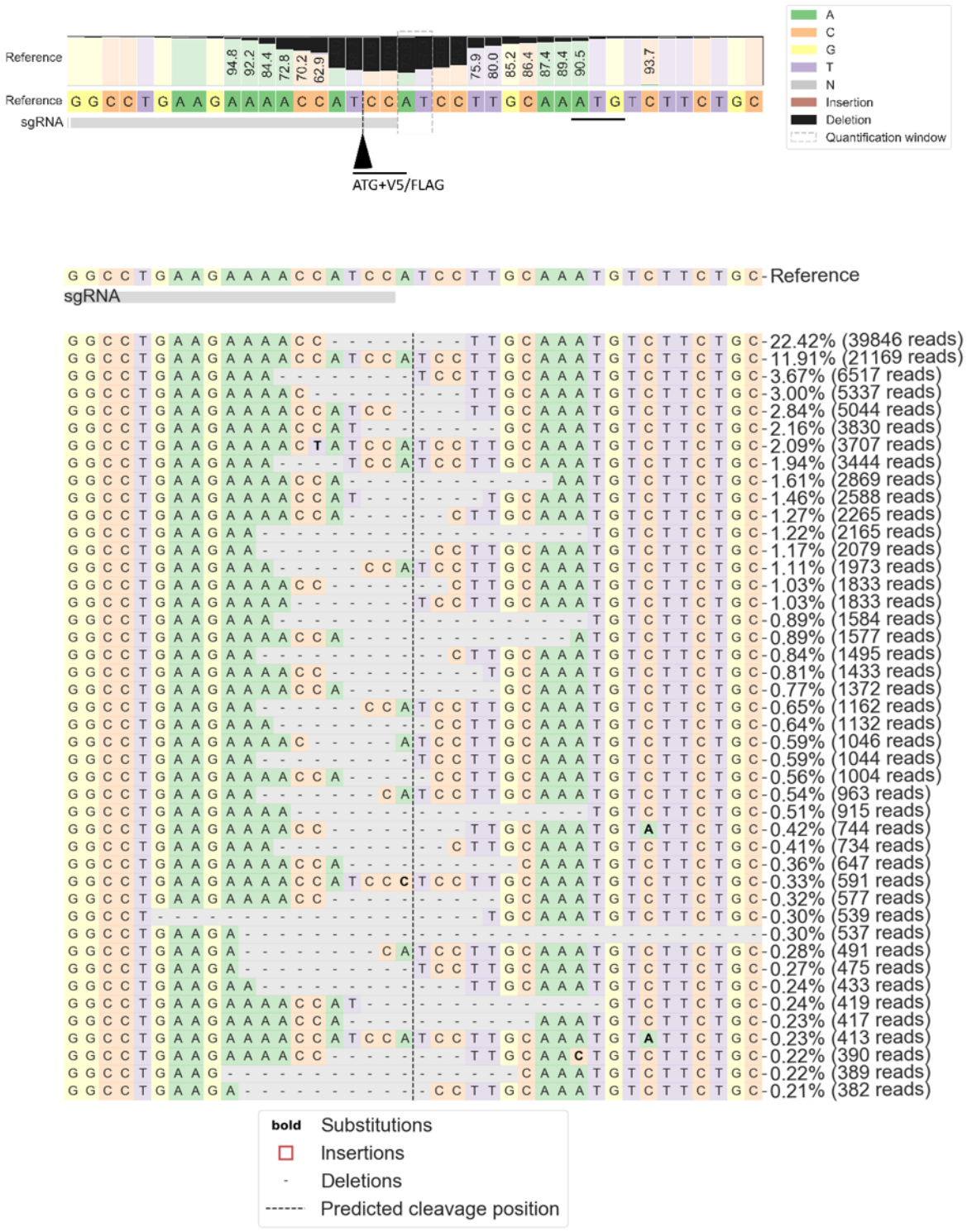


Figure 4.7 Diagram showing nucleotide percentage quilt (top) and alleles frequency table (bottom) around *NANOGP1*_5'_Cas12a crRNA.

Top: Predicted crRNA cut site and the tag knock-in site are indicated with a black arrow. *NANOGP1* start codon is underlined. Numbers above nucleotides indicate the percentage of reads, where a respective nucleotide was not deleted. *NANOGP1*_5'_Cas12a crRNA specificity was validated using the crRNA design software, which confirmed that it does not target the *NANOG* sequence.

Bottom: *Reference* shows the uncut control amplicon sequence. Percentage and number of reads, corresponding to a specific allele type, are shown on the right.

Here crRNA is called sgRNA, which is a universal title that CRISPResso gives to all guide RNA molecules.

At the end of this screen, I selected the *NANOGP1*_5' Cas12a crRNA to be used in the epitope tagging. None of the other CRISPR reagents were used.

Next, I optimised the hPSC transfection reaction protocol. I had to ensure that the transfection reaction is as efficient as possible, so that in the final experiment a sufficient number of cells in the population would contain epitope-tagged NANOGP1. During the optimisation, among other experimental factors I had to take one important limitation into account. The ssODN templates were designed to be as short as possible to ensure the most efficient HDR; they therefore only contained the tag sequence and homology arms, but no selectable reporter. Consequently, I tested whether co-transfecting CRISPR reagents with a vector encoding GFP protein under a constitutive promoter could improve identification of the transfection efficiency. In this experiment, I nucleofected naïve HNES1 hPSCs with a pCAG-eGFP vector and two different amounts of ssODN using Amaxa™ Nucleofector™ 4D. After 24 h, the transfected hPSCs were analysed by flow cytometry. The analysis showed that combining multiple components in the transfection reaction significantly reduced its efficiency (Figure 4.8).

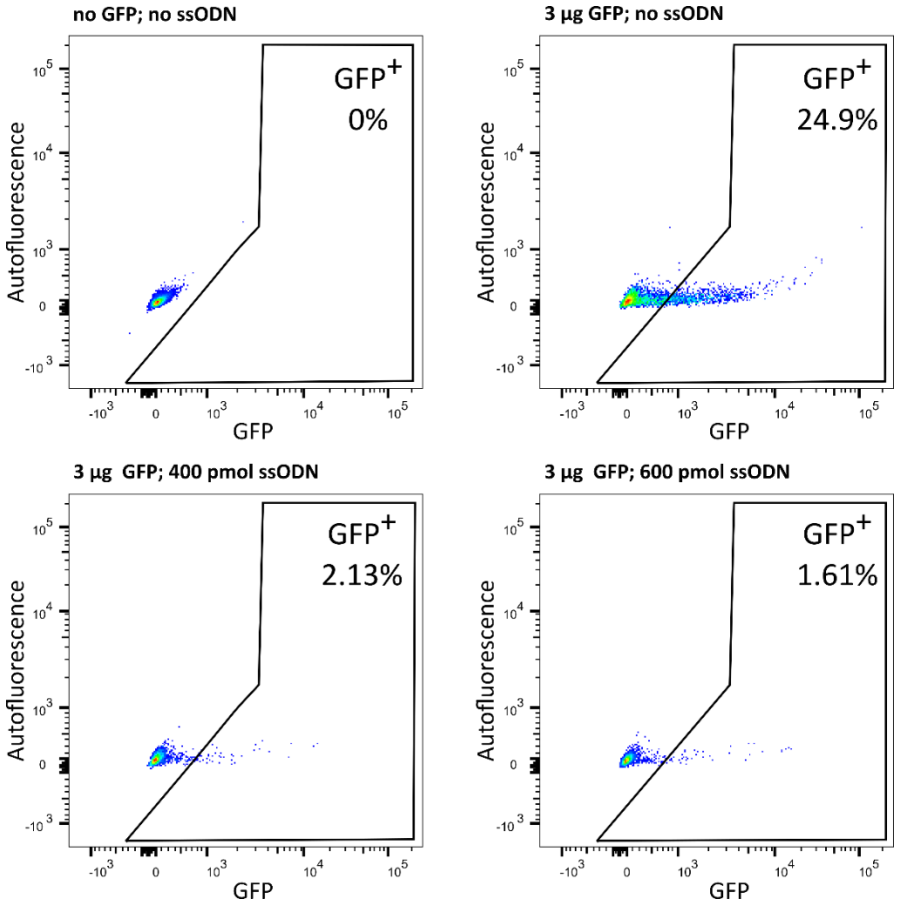


Figure 4.8 Flow cytometry scatterplots showing efficiency of co-transfecting ssODN reagents into hPSCs with a constitutive GFP-encoding vector. Title of each scatterplot indicates the amount of plasmid/ssODN used in a transfection. X-axis shows GFP fluorescence. Y-axis shows autofluorescence.

Percent of GFP-positive cells is indicated in the top right corner of each scatterplot. This experiment was performed twice, one representative replicate is shown here.

Cells transfected with the GFP vector alone had almost 25% GFP^+ cells, while addition of ssODN decreased the positive population to 1.6% and 2.1%. Such low transfection efficiency was not suitable for my future experiments, as other CRISPR reagents would have been added for the reaction to occur, also potentially decreasing the transfection efficiency. Also, not all transfected cells would undergo HDR, affecting the experimental efficiency even further. Therefore, this strategy was not pursued, and no additional reporter was added to the tagging protocol. Instead, I decided to assess the efficiency of gene editing after transfection and cell expansion by genotyping and immunostaining.

For the final epitope tagging transfection I used the Neon™ Transfection system, which is known to be an efficient system for nucleofecting naive hPSCs (Guo et al., 2021). A detailed description of the final nucleofection protocol can be found in Chapter 2. Briefly, naïve CR-H9 hPSCs were transfected with ssODN encoding V5 or 3xFLAG, *NANOGP1_5'* crRNA and Cas12a. After the nucleofection, cells were seeded into PXGL medium (Bredenkamp et al., 2019b; Rostovskaya et al., 2019), supplemented with 10 µM Y-27632 for increased cell survival and 2 µM M3814, a NHEJ-inhibitor to increase the HDR efficiency (Riesenberg et al., 2019). The latter two components were kept in the medium for 72 h. Additionally, after the nucleofection, cells were incubated at 32°C for 24 h ('cold shock') to further improve the HDR (Guo et al., 2018; Skarnes et al., 2019).

This experiment was performed in two biological replicates for each epitope tag, and all replicates showed the presence of the tags in naïve hPSC genomes after expanding the cells for a week, as shown in gel electrophoresis images (Figure 4.9). Successful epitope tag insertion was additionally validated by Sanger sequencing (Genewiz), using primers HA-F and P1-tag-seq-F. Epitope-tagged naïve hPSC lines were further analysed as a bulk heterogeneous population, as they did not survive the procedure of clonal selection.

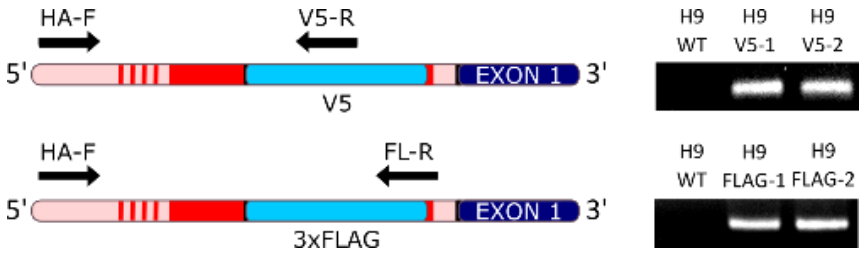


Figure 4.9 Diagram (left) and gel electrophoresis images (right) showing the strategy and outcome of the *NANOGP1* epitope tagging genotyping, '5'' and '3'' indicate 5' and 3'-ends of the nucleotide sequence. 5'-UTR is shown as pink blocks. Exon 1 is shown as a dark blue block. crRNA target sequences are shown as red blocks. Vertical red lines represent PAM sequence. 3xFLAG/V5 tag is indicated as light blue blocks. HA-F, V5-R, FL-R – PCR primers used for genotyping. HA-F primer binds upstream the HDR repair template. H9 – name of the cell line. WT – untransfected wild-type cell lines.

V5-1, V5-2 – two NANOGP1-V5 hPSC lines (two biological replicates). FLAG-1, FLAG-2 – two NANOGP1-3xFLAG hPSC lines (two biological replicates).

Naïve hPSCs containing V5-epitope were also analysed using immunostaining using an anti-V5 antibody. The assay demonstrated the presence of nuclear-localised V5-tagged protein in the cells. Additionally, the V5-signal overlapped with a nuclear OCT4 signal, showing that the tagged cells were pluripotent. Also, as expected, the proportion of cells with a V5 signal (cells that had undergone transfection, successful DNA cutting and HDR) was very low (Figure 4.10).

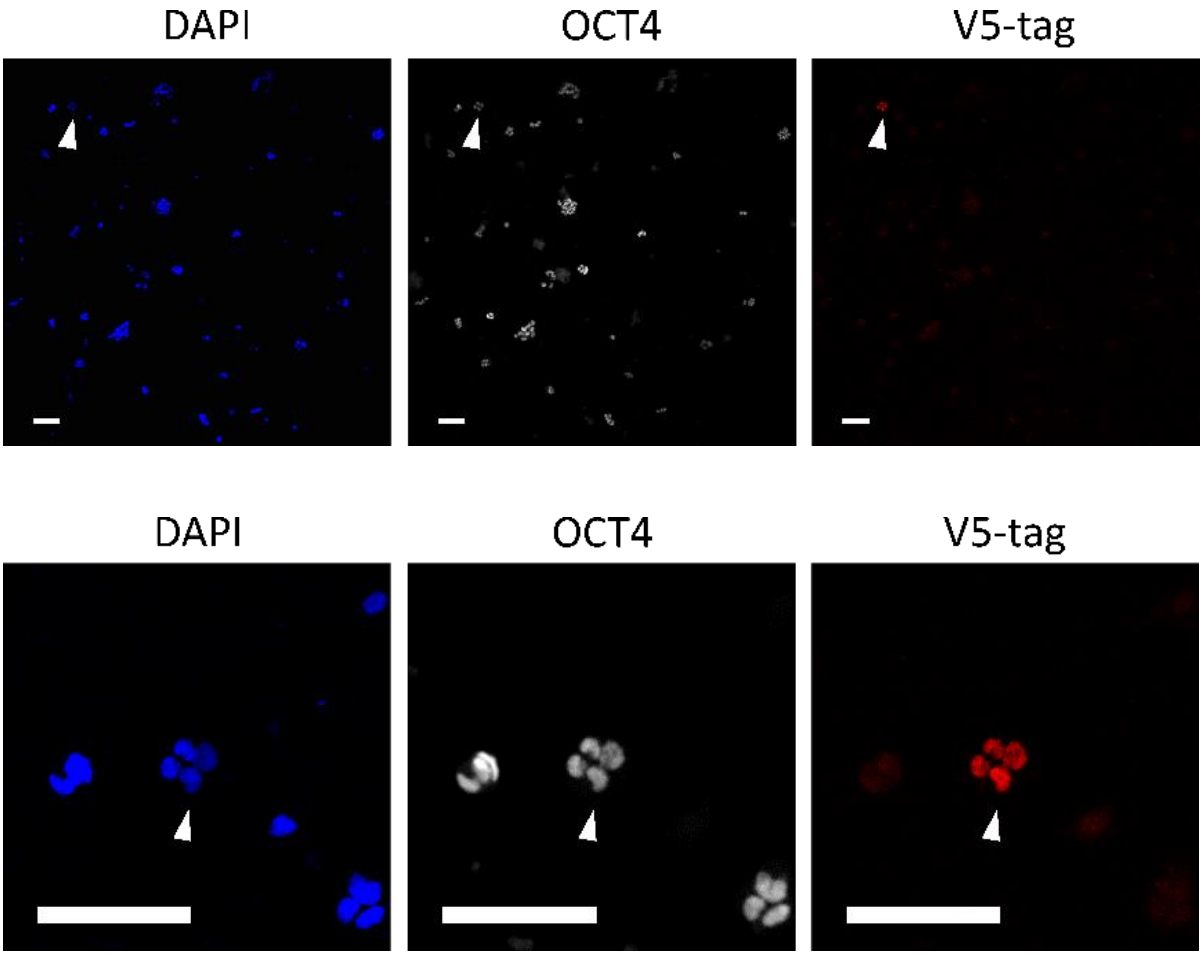


Figure 4.10 Immunofluorescence images showing co-localised nuclear V5-tag, OCT4 and DAPI signal in NANOGP1-V5 naïve hPSCs. Bottom panel is a zoom-in image of the V5-tag positive colony in the top panel. White arrows indicate the V5-positive colony. Scale bar, 100 µm.

A similar immunostaining experiment was attempted for the NANOGP1-3xFLAG tagged hPSCs. However, the antibody did not exhibit the required specificity and as a result, images produced by the wide-field and confocal microscopes had a very high background signal (not shown).

Due to a low number of tagged cells in the population, I decided not to analyse the cells in bulk by Western blotting. Instead, I performed a co-immunoprecipitation experiment using anti-V5 or anti-FLAG antibodies, which enriched the amount of the tagged protein in the sample, and then

performed a Western blotting assay using anti-NANOG antibodies. To validate that the V5 and FLAG signal corresponds to epitope-tagged NANOGP1, I used two different anti-NANOG antibodies that can distinguish between NANOG and NANOGP1 due to their predicted protein sequence differences (Figure 4.11). The NANOG antibody from R&D recognises the C-terminus of NANOG, and should therefore also detect NANOGP1. The NANOG antibody from Abcam binds to the N-terminal domain of NANOG that is absent in NANOGP1.

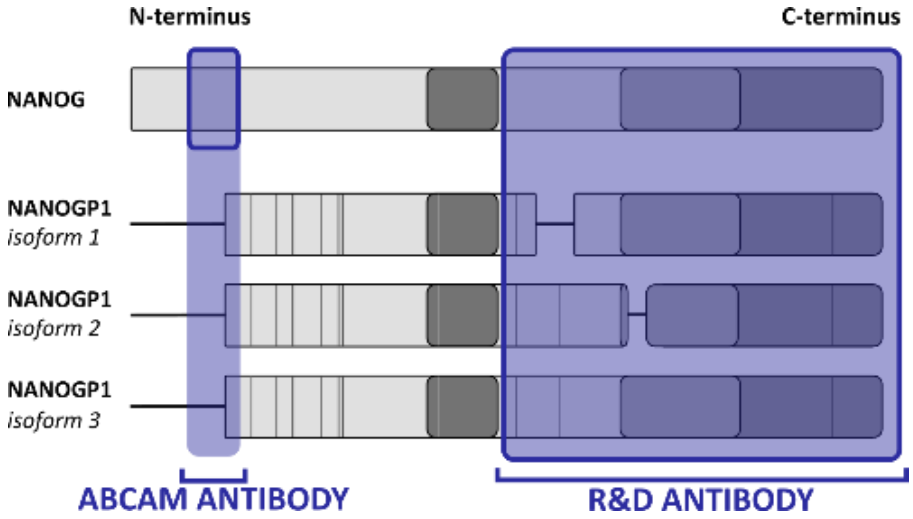


Figure 4.11 Diagram showing specificity of the two NANOG antibodies used in this project. NANOG and NANOGP1 domains are in grey. Antibody binding regions are in purple.

The results showed that epitope-tagged NANOGP1 protein was detected in naïve CR-H9 hPSCs (Figure 4.12). As expected, NANOGP1 was not detectable in the input samples, presumably because very few cells in the population were correctly tagged, but the NANOGP1 signal was present in the immunoprecipitated samples. A protein of the same molecular weight as the V5 and FLAG bands was detected by the C-terminal but not the N-terminal NANOG antibody, providing strong evidence that the immunoprecipitated protein corresponded to NANOGP1. NANOG was present in the input and detected by both R&D and Abcam antibodies, and was not detected in the pull-down sample.

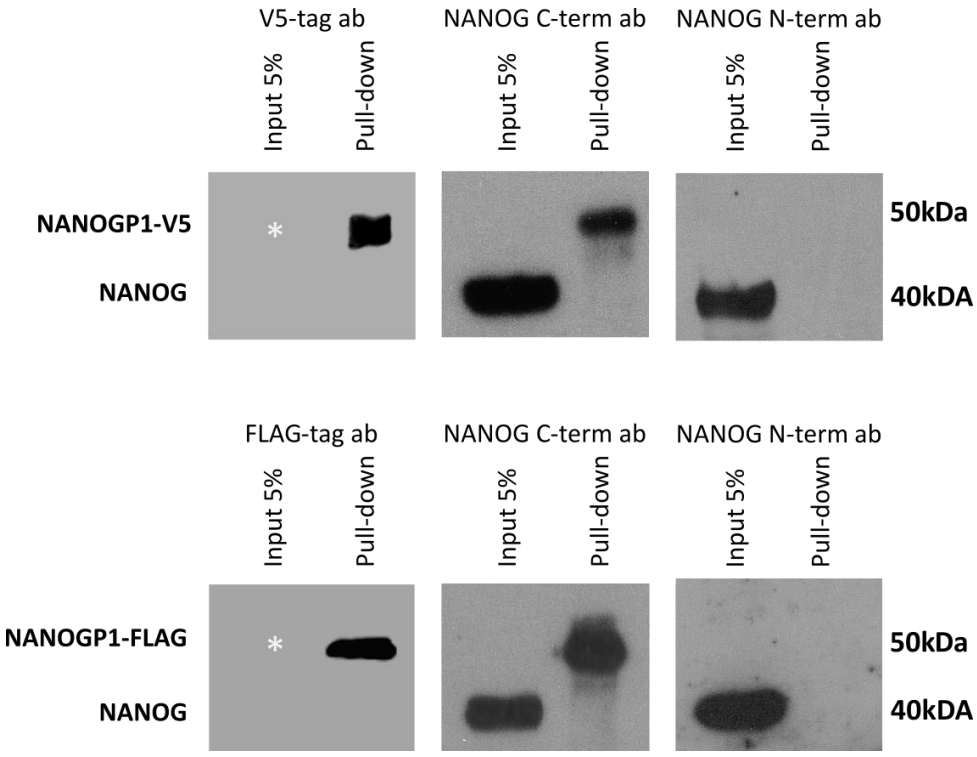


Figure 4.12 Western blotting image showing proteins pulled down in the V5 and FLAG Co-IP experiments compared to their corresponding input samples. * - due to low number of NANOGP1-epitope tagged cells in the population, the proteins were not detected in the input samples. 50kDa, 40kDA – approximate size of the pulled-down proteins.

In summary, in this section, I have shown for the first time that endogenous *NANOGP1* pseudogene is translated into a stable, nuclear-localised protein. This finding resolved conflicting reports in literature (described in Section 1.4.4-1.4.5) and publicly available databases (i.e., Ensembl) that for many years had not been able to agree whether *NANOGP1* is protein-coding or not. Having discovered that *NANOGP1* in fact encodes a protein, it has now become possible to investigate its chromatin binging profile, comparing it to the chromatin binding of NANOG. This enabled the first functional characterisation of *NANOGP1* in the context of human pluripotency, which is described in the following section.

4.2.2 Characterising NANOGP1 chromatin binding in naive hPSCs by ChIP-sequencing

After discovering that NANOGP1 protein is expressed by naïve hPSCs, I analysed the chromatin binding of the epitope-tagged protein using ChIP-sequencing (ChIP-seq). Of note: the results presented here are preliminary, because in the current form they raise several concerns from a quality control (QC) point of view. In some cases, it was difficult to determine if the ChIP signal was true or afflicted by technical artefacts. Nevertheless, the results obtained here are interesting and are worth describing. In the Discussion, I will also describe potential ways to investigate this topic further.

The ChIP-seq experiment was performed in two biological replicates, using two separately generated cell lines for each tag. A detailed description of the protocol can be found in Section 2.5.

Briefly, pellets were double-fixed using Di(N-succinimidyl) glutarate and formaldehyde (Nowak et al., 2005), and lysed. This was followed by chromatin sonication that produced fragment size distributions of predominantly ~500 bp, as seen by gel electrophoresis (Figure 4.13).

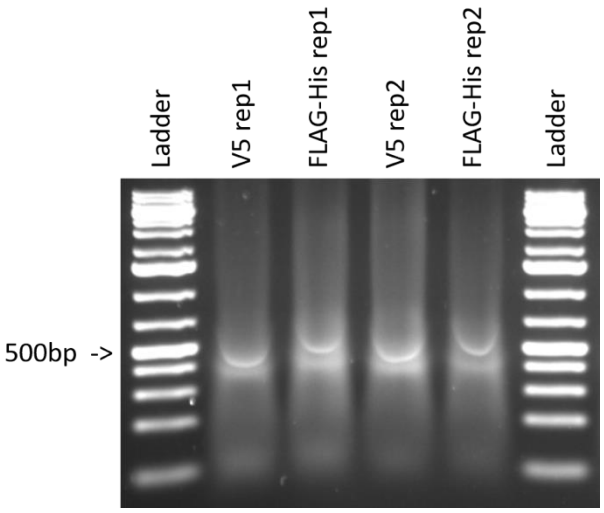


Figure 4.13 Gel electrophoresis image showing fragment size distribution of the sonicated chromatin used in ChiP-seq. Lane titles indicate which cell line the chromatin was extracted from. 500 bp – average fragment size. Rep1/2 – biological replicate (here, individual hPSC lines).

ChIP-seq reactions were performed using the same V5 and FLAG antibodies that had been used in co-immunoprecipitation, described above. Additionally, as a positive control, I did ChIP-seq using an anti-NANOG C-terminal antibody with the NANOGP1-V5 and NANOGP1-FLAG naive hPSC lines. After immunoprecipitation and stringent washing, the chromatin was reverse cross-linked, RNA and proteins were degraded, and the isolated DNA used for generating ChIP-seq libraries. After library preparation, the absence of contaminants and library fragment size distributions were tested with help of an Agilent Bioanalyser (Figure 4.14).

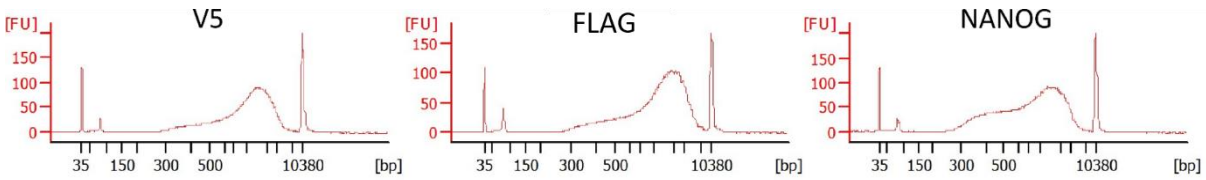


Figure 4.14 Spectra showing ChIP-seq library fragment size distribution, produced by Agilent Bioanalyzer. Broad library peaks are positioned between the two size standards located at 35 bp and 10,380 bp. A small peak near the 35 bp standard marker is a remaining adapter (insignificant amount). Y-axis – arbitrary fluorescence unit. X-axis – fragment size scale, bp.

In total, I generated nine libraries: 2xNANOGP1-V5, 2xNANOGP1-3xFLAG, their respective input samples, and a NANOG sample. After generating the libraries and performing the QC, the libraries were sequenced by the Babraham Institute Next Generation Sequencing facility, which generated 20-43 million uniquely mapped reads per library. Initial data analysis demonstrated by Principal Component Analysis (PCA) showed that the biological replicates were grouped together, as were all of the inputs, which suggests good quality ChIP-seq replicates (Figure 4.15).

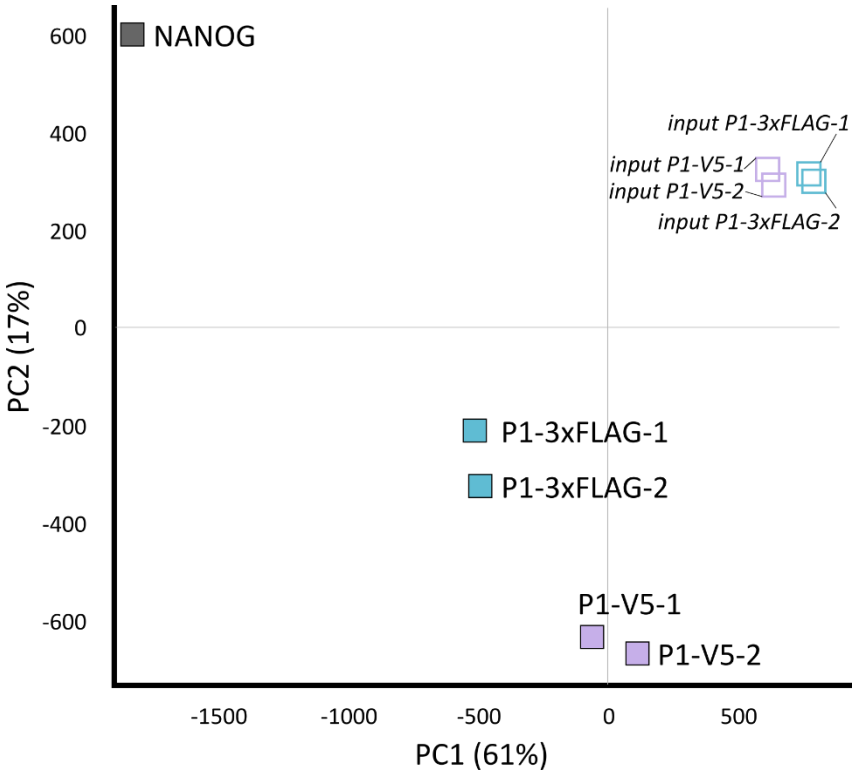


Figure 4.15 PCA plot showing comparison of ChIP-seq samples. X-axis – PC1, Y-axis – PC2. PCA was performed over all probes. *Adapted with permission from Dr. Christel Krueger.*

Further analysis revealed that there were no obvious areas of enrichment in sequencing reads above the input for the two NANOGP1-V5 samples, indicating that the V5 ChIP had failed. In contrast, the NANOG and NANOGP1-3xFLAG were noticeably different compared to their corresponding input samples and showed regions of strong enrichment (peaks); four sequencing data track windows are shown in Figure 4.16 as an example. Therefore, only the NANOGP1-3xFLAG libraries were used further in the analysis.

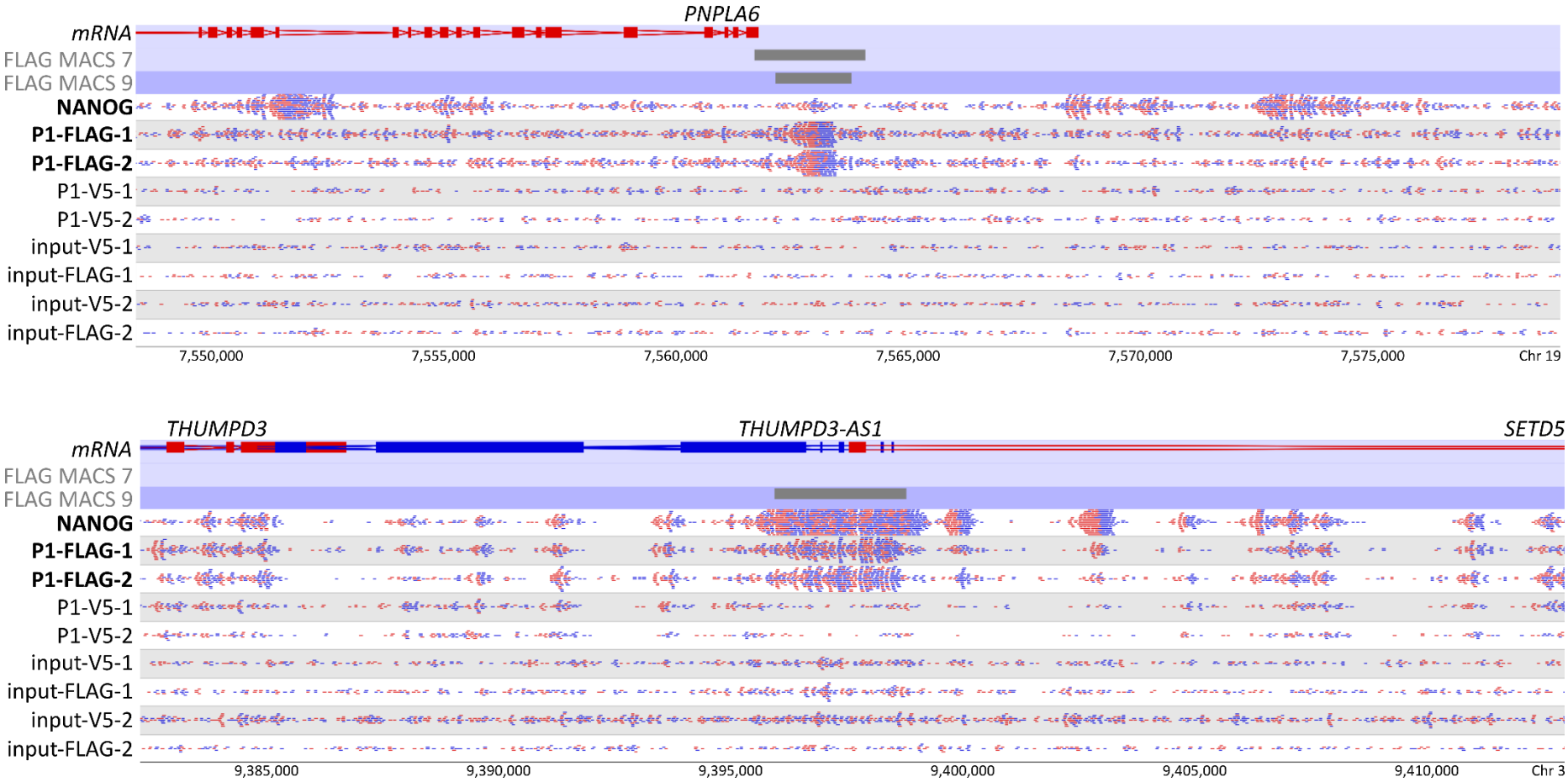


Figure 4.16 ChIP-seq track, showing comparison of sequencing reads enrichment between all the samples – continued on the next page. Sequencing reads are mapped against the published genome sequence, GRCh38_v10.2; mRNA for selected genes is shown (top row). Mapped reads and mRNA are in red and blue, corresponding to the two opposing DNA strands. NANOGP1-3xFLAG peaks are shown as grey blocks; peak calling used $p<10^{-7}$ and $p<10^{-9}$ parameters in 'FLAG MACS 7' and 'FLAG MACS 9', respectively. Scale at the bottom of the diagram represents position of the locus within the chromosome. Chr – chromosome. Scale, 5 kb. *Adapted with permission from Dr. Christel Krueger.*

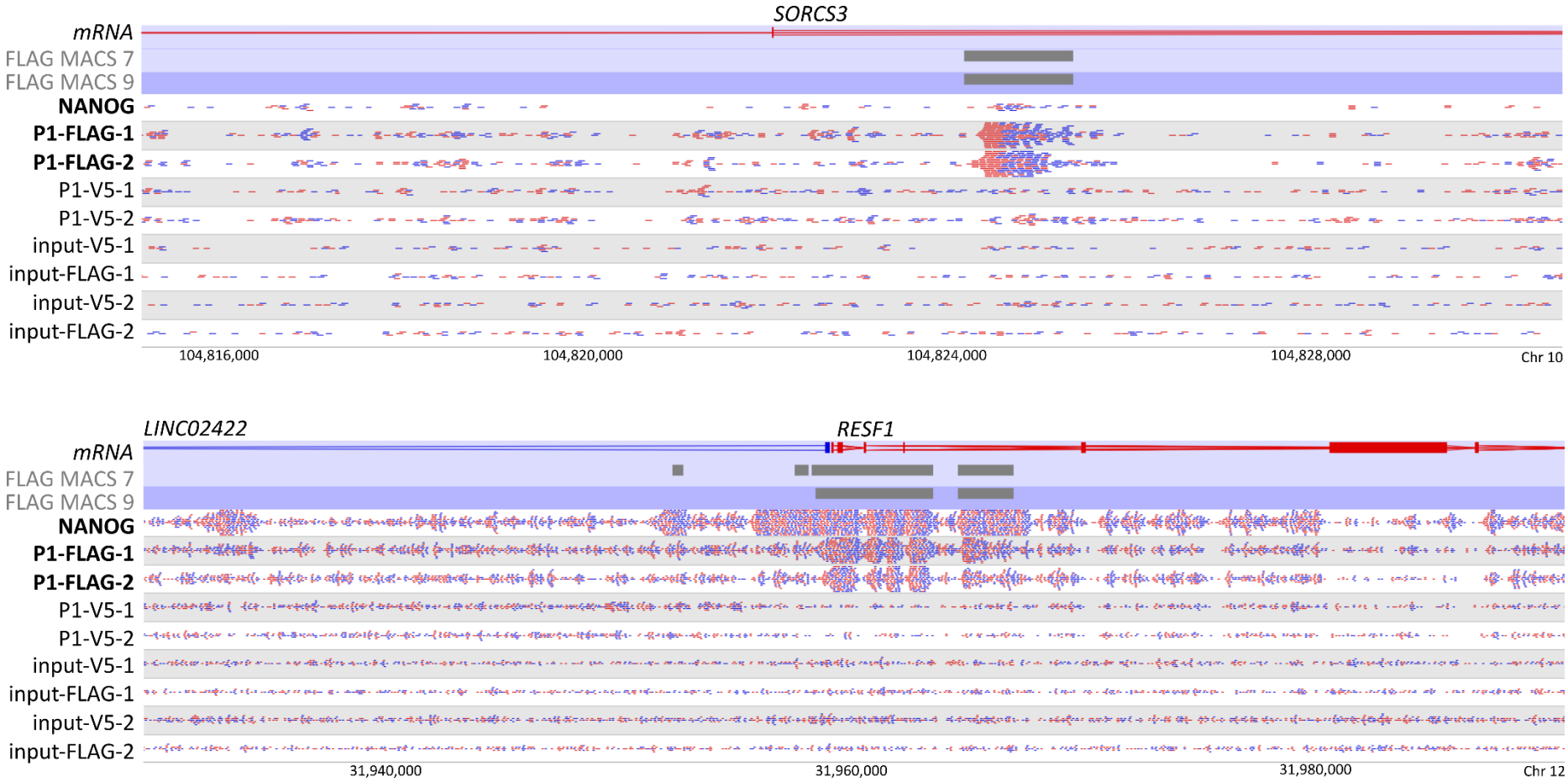


Figure 4.17 ChIP-seq track, showing comparison of sequencing reads enrichment between all the samples. Sequencing reads are mapped against the published genome sequence, GRCh38_v10.2; mRNA for selected genes is shown (top row). Mapped reads and mRNA are in red and blue, corresponding to the two opposing DNA strands. NANOGP1-3xFLAG peaks are shown as grey blocks; peak calling used $p<10^{-7}$ and $p<10^{-9}$ parameters in 'FLAG MACS 7' and 'FLAG MACS 9', respectively. Scale at the bottom of the diagram represents position of the locus within the chromosome. Chr – chromosome. Scale, 4 kb (top), 20 kb (bottom). *Adapted with permission from Dr. Christel Krueger*

This ChiP-seq experiment contained only one replicate of NANOG ChiP-seq library, which was intended to be a positive control for the experimental procedure. In order to be able to compare NANOGP1-3xFLAG chromatin binding to that of NANOG, a published NANOG ChIP-seq dataset in naïve hPSCs was used as it had replicated data (Chovanec et al., 2021). The correlation between the published ChIP-seq dataset and the one produced in this study was strong (R=0.598), as shown in the scatterplot in Figure 4.18.

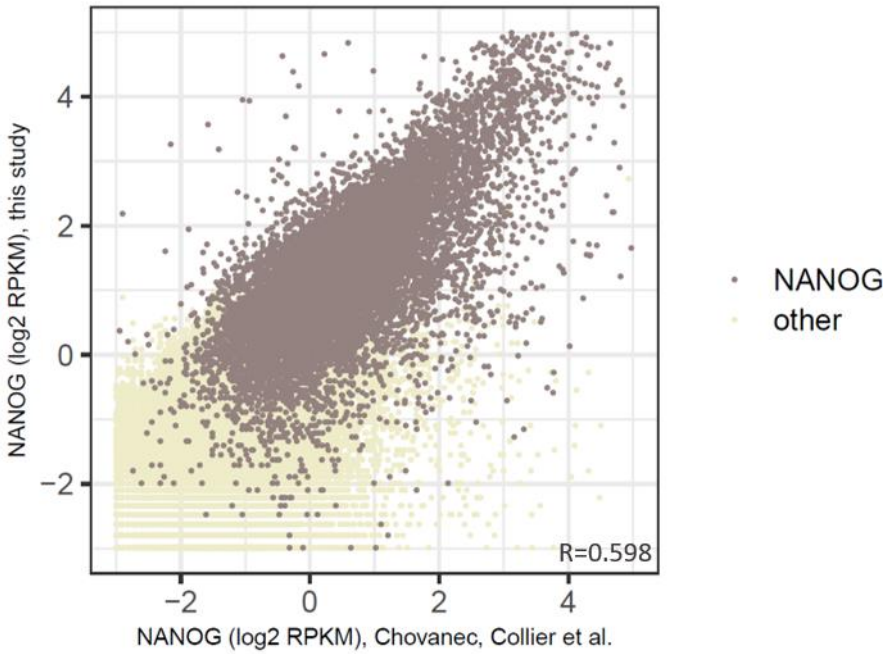


Figure 4.18 Scatterplot showing log2 RPKM quantitation of 1 kb genome tiles for NANOG ChIP-seq data. NANOG ChIP-seq in this study was performed on naïve H9 NANOGP1-3xFLAG hPSCs; NANOG ChIP-seq from Chovanec et al., 2021 was performed on naïve H9 NK2 hPSCs. Tiles overlapping the NANOG peaks from Chovanec et al., 2021 are shown in dark grey. *Adapted with permission from Dr. Christel Krueger.*

Taking the intersect of two independent biological replicates, 467 NANOGP1 only peaks were identified using MACS (Zhang et al., 2008), $p<10^{-9}$. Out of the 467 peaks, 383 (82%) overlapped with NANOG peaks from Chovanec et al., 2021. These 'shared' peaks represent a small fraction (0.533%) of the total number of all NANOG-bound sites (n=71923). Surprisingly, the remaining 84 NANOGP1-bound sites did not overlap with NANOG, thereby defining NANOGP1-specific peaks (Figure 4.19).

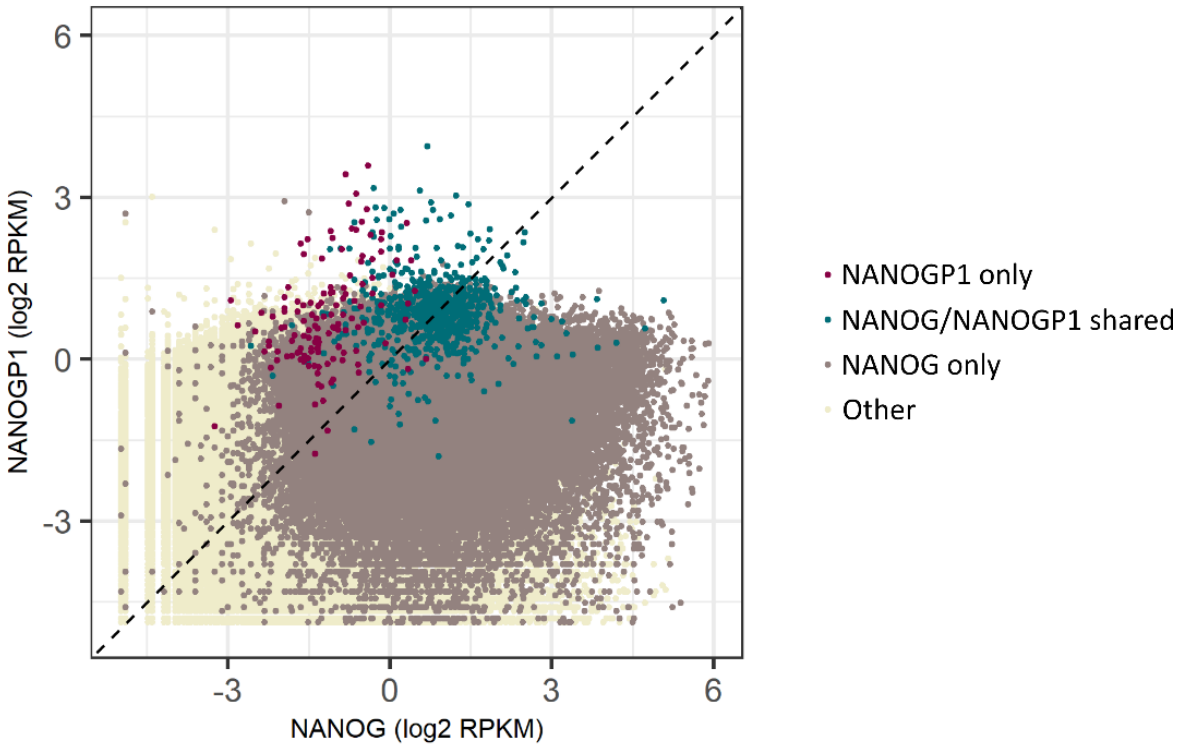


Figure 4.19 Scatterplot showing log2 RPKM quantitation of 1 kb genome tiles for NANOG (x-axis) and NANOGP1 (y-axis) ChIP-seq data. Tiles overlapping NANOG only peaks are shown in dark grey, NANOGP1-3xFLAG only peaks in dark violet and shared peaks in teal. *Adapted with permission from Dr. Christel Krueger.*

Overall, NANOGP1 ChiP-seq helped to identify chromatin regions bound by both NANOG and NANOGP1, which was expected based on the conservation of the DNA-binding homeodomain between the two duplicates. Presence of NANOGP1 only peaks, however, was potentially concerning, and is discussed below.

This thesis and Chovanec et al., 2021 used the same NANOG C-terminal antibody for ChiP-seq, which was expected to bind both NANOG and NANOGP1 C-terminus. Therefore, it was concerning that NANOGP1 only peaks were detected in our dataset but not in Chovanec et al., 2021. To test whether NANOG was indeed absent from the NANOGP1 only sites, ChIP-seq signal over a 4kb signal at NANOGP1 peaks was analysed. The figure below shows that in fact, a very small level of NANOG can be detected in the NANOGP1 only peaks Figure 4.20, and the location of peaks follows the same pattern in the NANOG ChiP-seq library generated in this study and the one made by Chovanec et al., 2021. Some areas in the NANOGP1 only peaks, however, appear to be NANOG negative. Based on this analysis, it is possible that existence of NANOGP1 only peaks reflects real biology, but it could also be related to a technical artefact. Therefore, in my opinion, this dataset should be analysed further, but with caution and treated as preliminary.

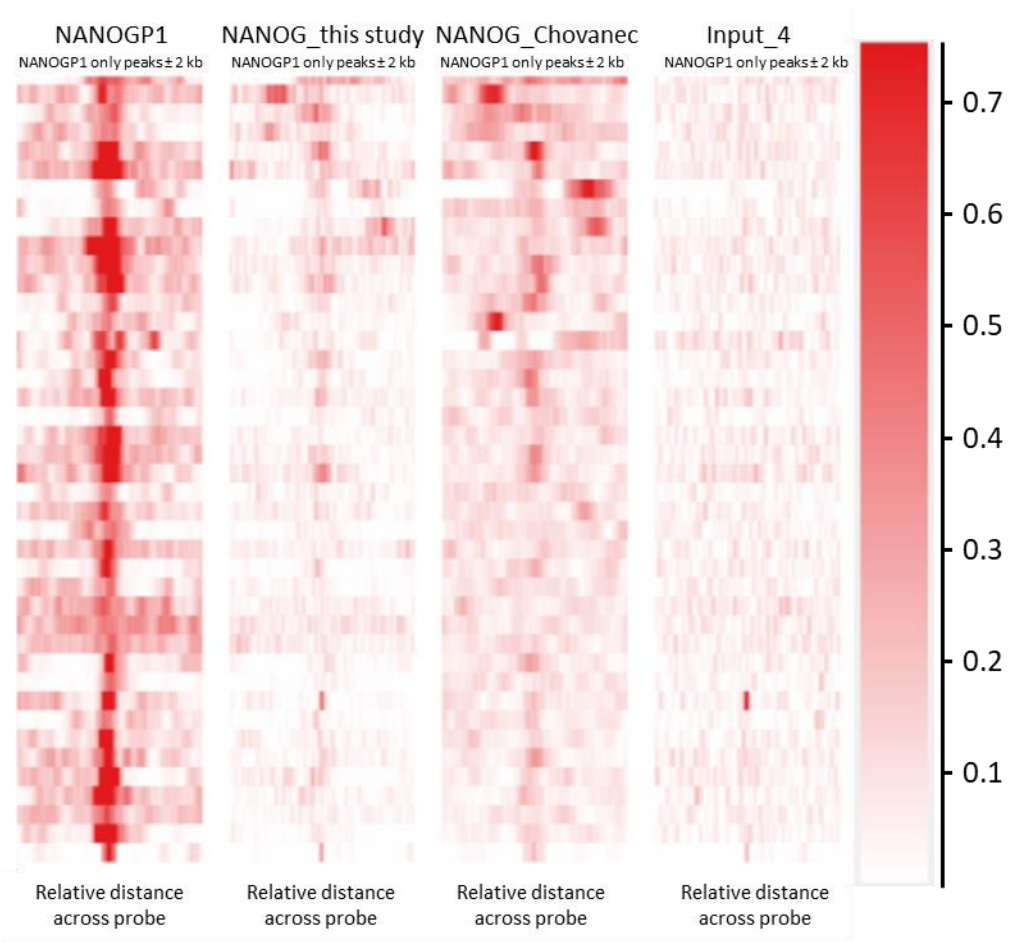


Figure 4.20 Heatmap showing reference adjusted ChIP-seq signal across a 4kb window, centred on the NANOGP1 peaks. Peaks were MACS called (Zhang et al., 2008) in naive hPSCs from this study across NANOGP1 only, NANOG only and a control sample 'Input_4' datasets from this study, as well as NANOG datasets from Chovanec et al., 2021. *Adapted with permission from Dr. Christel Krueger.*

In a further attempt to characterise NANOG only, NANOGP1 only and NANOG/NANOGP1 shared peaks, and to understand whether they could be different functionally, chromatin state analysis was performed with help of ChromHMM software (Ernst and Kellis, 2012). In this analysis, peak location data was compared with the annotated chromatin features such as, for instance, distinct regulatory regions (active/inactive enhancers), inactive chromatin, or background. In this analysis, annotation categories were taken from Chovanec et al., 2021. An additional control group was added containing 1000 randomly selected 900 bp genomic windows, which mostly fell within the background. As a result, the overall chromatin state profiles within NANOG, NANOGP1 and NANOG/NANOGP1 shared peaks were found to differ between the systems (Figure 4.21). Compared to a set of randomly-assigned peaks, NANOGP1 only and NANOG/NANOGP1 shared peaks were much more likely to overlap with active promoters than other defined categories, whereas NANOG only peaks were likely to be found in active and inactive enhancers (31% and 25%), as well as active promoters (18%). Notably, 74% of NANOGP1 only peaks corresponded to the background, which could

not be annotated by any particular regulatory pattern, while only 1% of the shared peaks was found in the background, and the majority, 89% corresponded to active promoters. Similarly, only 7% of the NANOG only peaks were found in the background, highlighting the general difference in the chromatin state of the NANOGP1 only peaks compared to NANOG and NANOG/NANOGP1 categories. However, it is important to note that in Chovanec et al. 2021, only a few histone marks were used to define chromatin states. It is possible, therefore, that the NANOGP1 'background' can be defined by other markers that were not included in the original analysis.

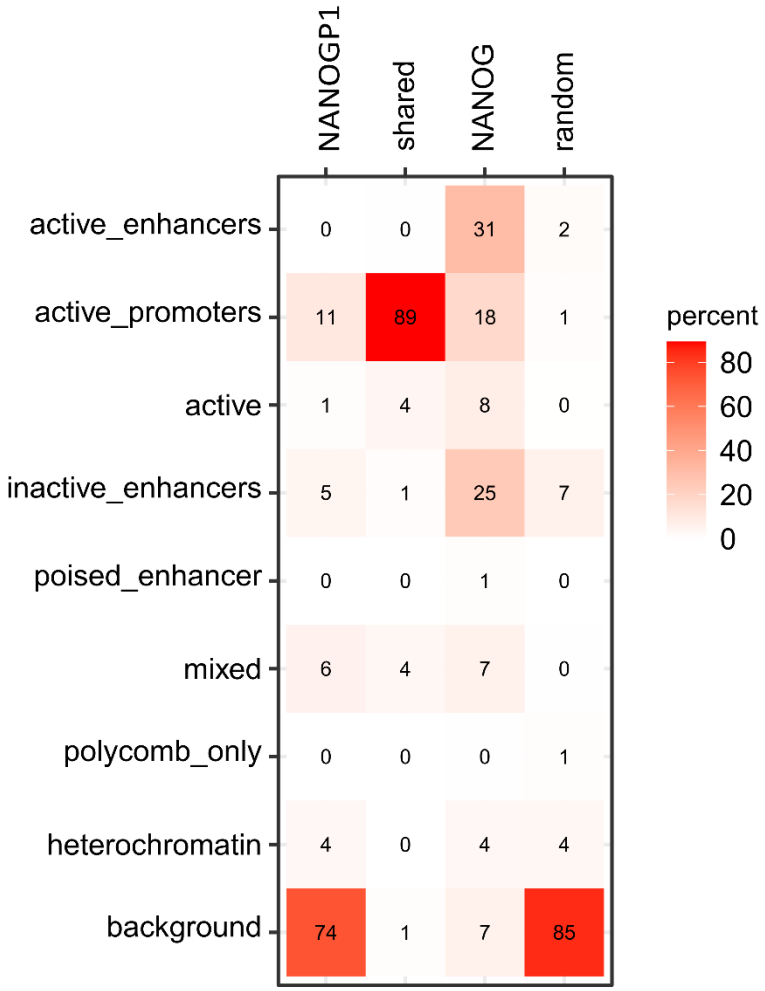


Figure 4.21 Heatmap showing peak locations with respect to chromatin states. Categories are from the ChromHMM genome annotation for naïve hPSCs in Chovanec et al., 2021. Numbers represent percent peak centres falling into each chromatin state. Random – 1000 x 900bp randomly selected genomic windows. *Adapted with permission from Dr. Christel Krueger.*

After analysing chromatin state profiles of the ChiP-seq data I concluded (provided that the peaks are biologically meaningful and not artefacts) the possibility that NANOG only, NANOGP1 only and NANOG/NANOGP1 shared binding may indeed have different functional properties. To further characterise protein binding within these three peak categories, *de novo* motif search was performed using Homer software. This analysis revealed one particular motif, highly represented among the

NANOGP1 only and NANOG/NANOG1 peaks, compared to both NANOG only peaks and a random set of 1000x900 bp genomic windows. NANOGP1 and shared motifs were strongly and significantly enriched in the top motif for an overlap with the binding motifs of REST protein ($p<1x10^{-42}$ each; Chi-squared test was performed) (Figure 4.22)**.** REST is a repressor of neuronal fate in the non-neuronal lineages (Schoenherr and Anderson, 1995).

Figure 4.22 Diagram showing the resulting motif produced in de novo NANOGP1-3xFLAG motif search. This coincides with the REST protein motif (Gertz et al., 2013). Letters represent the four DNA nucleotides, A – adenine, T – thymine, G – guanine, C – cytosine. Size of each letter reflects its abundance within the indicated position among all found motifs. *Adapted with permission from Dr. Christel Krueger.*

All the other motifs bound by NANOGP1 found were insignificant ($p>0.05$; Chi-squared test), with the exception of one other NANOGP1 only motif which had a significant p-value ($p=0.01$; Chi-squared test), but the motif data was found in a ChiP-seq dataset that was not published or peer-reviewed and therefore was not included here.

Since NANOGP1 only and NANOG/NANOGP1 peaks were overlapping with REST motif at a significant level, it was important to check whether any other peak categories had a similar (or different) REST motif representation. To do this, the percentage of peaks overlapping the REST motif was quantified for each category. As a result, more than 35% of all NANOGP1 peaks, and 5% of shared peaks, overlapped with REST motifs, compared to ~1% of NANOG-only peaks and <1% of random genomic regions. Notably, 11% of REST ChIP-seq peaks (Gertz et al., 2013) overlap with REST motifs, indicating that NANOGP1 is more often bound at sites containing a REST motif than REST itself (Figure 4.23).

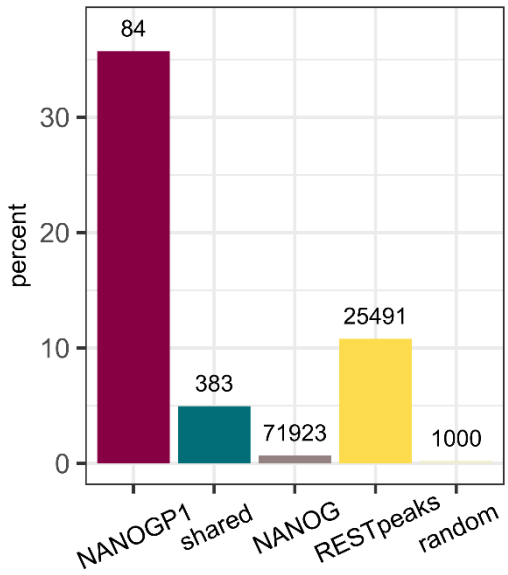


Figure 4.23 Bar plot showing percentage of peaks overlapping the REST motif. This bar plot integrates data from two different cell types: NANOGP1 and NANOG binding to REST motifs in naïve hPSCs (this study), and REST binding REST motifs in primed hPSCs (Gertz et al., 2013; ChIP-seq). Total peak numbers are indicated above the bars. Peak overlap data for each category was compared to a random set (1000x900 bp genomic regions) using the Chi-squared test, p-values are $p<1x10^{-42}$ (NANOGP1 and shared) and $p<1x10^{-12}$ (NANOG); significance accepted at $p<0.05$. *Adapted with permission from Dr. Christel Krueger.*

Interestingly, when REST ChIP-seq binding in primed H1 hPSCs was analysed (Gertz et al., 2013), REST log2 RPM signal was also higher in the NANOGP1 binding sites, not only compared to the shared and NANOG sites, but also to the majority of REST peaks themselves.

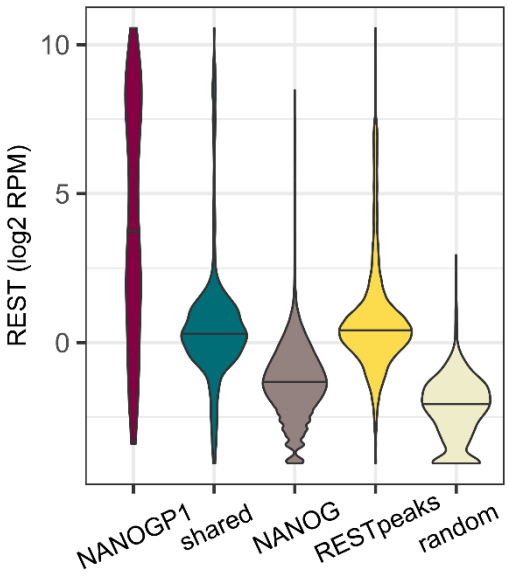


Figure 4.24 Violin plot showing REST protein binding in primed H1 hPSCs. Values shown are log2 RPM quantitation of 1 kb tiles overlapping ChIP-seq peaks. ChIP-seq data is from Gertz et al., 2013. REST

peaks and a random set of 1000x900 bp regions were chosen for comparison. Median values are shown as horizontal lines. *Adapted with permission from Dr. Christel Krueger.*

So far, the data demonstrated that NANOGP1 peaks were highly associated with REST motif and REST chromatin binding, and also exhibited unique chromatin state profiles. If NANOGP1 only and NANOG/NANOGP1 shared peaks were indeed functionally different from the NANOG only peaks, it could have become visible during analysis of their respective target genes. To characterise their potential target genes and to investigate whether they are united by any particular gene ontology (GO) categories, GO analysis was performed for the NANOG only, NANOGP1 only and NANOG/NANOGP1 shared predicted target genes, using Enrichr online tool.

At first, potential target genes were identified on the basis of their location within 10 kb of a peak. From that list, genes were only included if a peak overlapped with the gene sequence or was in its close proximity (0-500 bp). As a result, it was found that the majority of NANOGP1 only (86% out of n=84) and NANOG/NANOGP1 shared (96% out of n=383) peaks overlapped with or were near genes. In the case of NANOG only peaks, the starting number of them was so large (n=71,923), that a randomly selected subset of 200 NANOG target genes had to be generated (also subsequently applying 'peak overlapping or near a gene, 0-500 bp' filter). Notably, random n=200 lists of NANOG-only targets were generated five times, and the results of their GO analysis were consistently similar; therefore, only one list is shown below. As a result, NANOG only (n=200), NANOGP1 only (n=72) and shared (n=366) predicted target genes were used in the GO analysis.

The GO analysis revealed that NANOGP1 only gene targets were found in three neural lineage-associated terms, namely, Purkinje neurons, pyramidal cells and interneurons (Figure 4.25). These three terms were (a) top ranked by p-value <0.05 and (b) the only terms identified that had adjusted p-value <0.05.

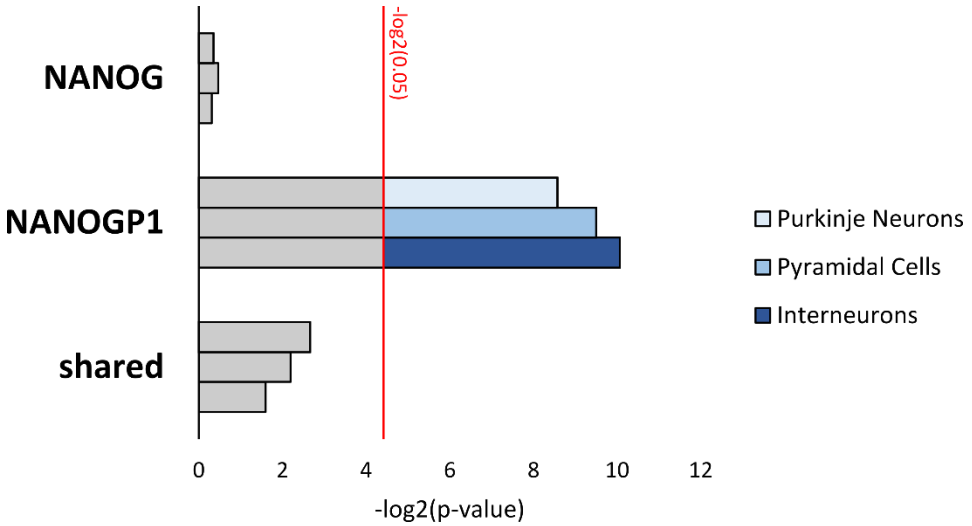


Figure 4.25 Bar chart showing neural cell types discovered in gene ontology (GO) analysis of predicted NANOGP1-3xFLAG gene targets. X-axis – -log2-transformed p-values produced in the GO analysis. Benjamini-Hochberg statistical test was applied, significance was accepted at $p<0.05$. Y-axis – predicted targets of NANOG, NANOGP1-3xFLAG as well as their shared target genes.

GO analysis of the NANOG/NANOGP1 shared peaks identified genes that are overrepresented in PSCs, dividing germ cells and, surprisingly, with a very low p-value, foetal natural killer T-cells (Figure 4.26).

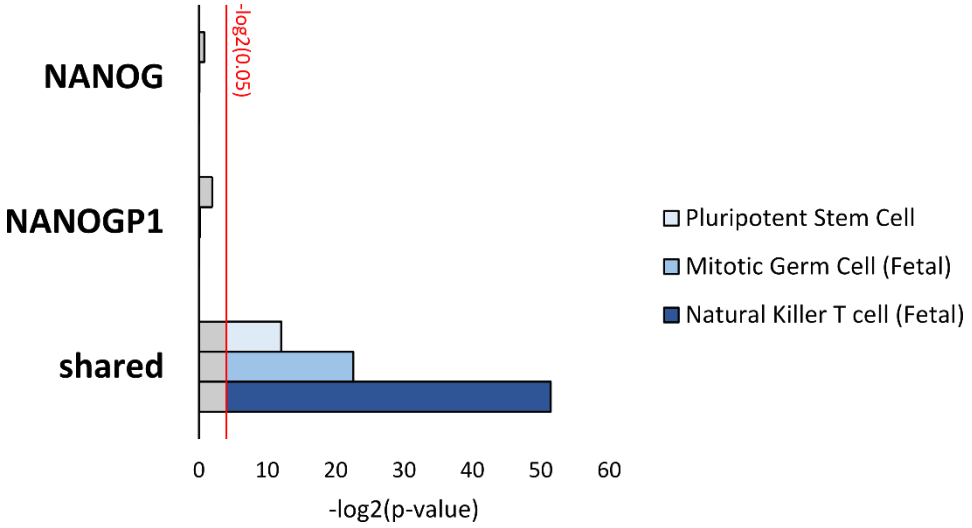


Figure 4.26 Bar chart showing cell types discovered in gene ontology (GO) analysis of predicted NANOG/NANOGP1 shared gene targets. X-axis – -log2-transformed p-values produced in the GO analysis. Benjamini-Hochberg statistical test was applied, significance was accepted at $p<0.05$. Y-axis – predicted targets of NANOG, NANOGP1, as well as their shared target genes**.** NANOG only target GO analysis did not produce any specific groups.

I also attempted to perform GO analysis on a list of predicted NANOG only target genes (n=200). However, no meaningful GO categories were identified, likely due to such a high and versatile number of NANOG only peaks.

GO analysis results allowed me to hypothesise that NANOGP1 only binding could be potentially functionally different from the protein binding in NANOG only and NANOG/NANOGP1 shared binding regions, based on such a noticeable difference between their predicted downstream targets.

To further analyse the gene clusters described above, the Enrichr Clustergram tool was used to check which genes and, more importantly, how many of them, drove the clustering. The 'Neural cell type' cluster that was associated with the genes near to NANOGP1 peaks was found to be driven by a group of eight genes (out of 72 in the gene set) (Figure 4.27).

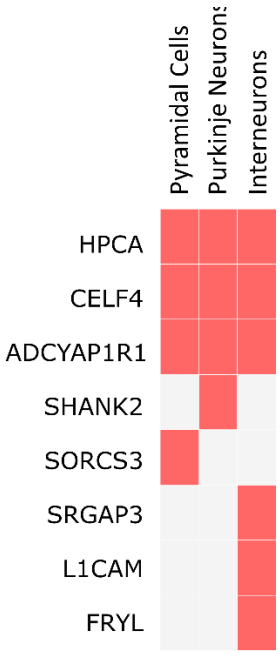


Figure 4.27 Heat map showing predicted NANOGP1 target genes that contributed to formation of specific cell types in the gene ontology analysis.

Interestingly, the 'Natural Killer T-cell group' that was identified in genes near to shared peaks were driven by a very large number of genes. The list contained more than 150 unique names, which was almost half of all unique gene targets (n=366) of the shared NANOG/NANOGP1 peaks (
Figure 4.28).

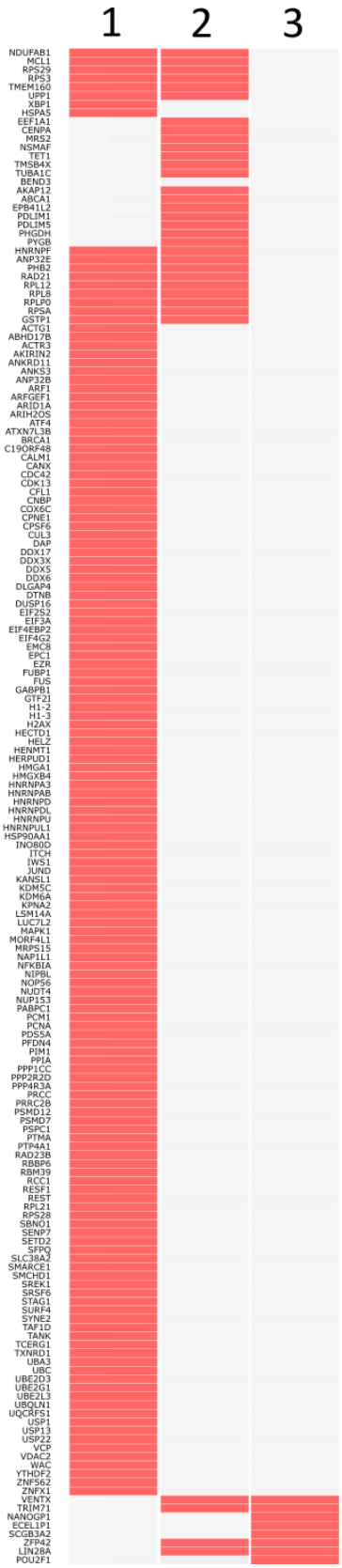


Figure 4.28 Heat map showing predicted NANOG/NANOGP1 target genes that contributed to formation of specific cell types in the GO analysis. 1 – NKT cell (foetal), 2 – Mitotic germ cell (Foetal), 3 – PSC.

Finally, using the same gene lists as in the GO analysis above, NANOG, NANOGP1 only and NANOG/NANOGP1 shared target gene expression was analysed in the naïve t2iLGo hPSC RNA-seq dataset, produced in this study (see Chapter 5). Here, shared gene targets were found to be expressed at a much higher level compared to the NANOG-only and NANOGP1-only target genes (Figure 4.29).

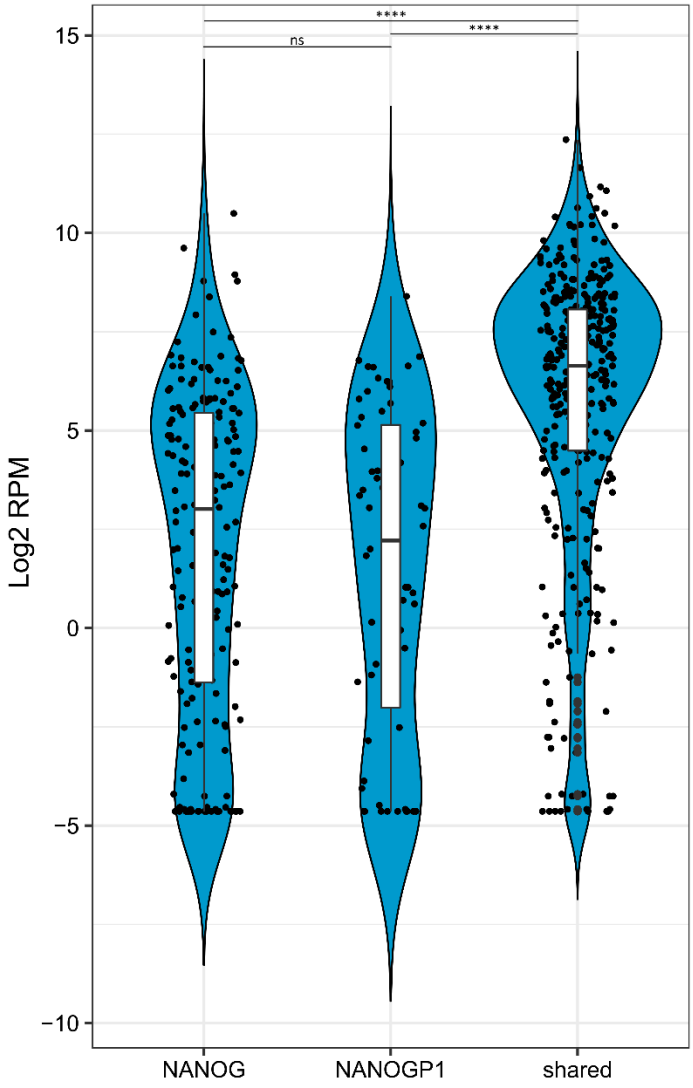


Figure 4.29 Violin plots showing expression of predicted NANOG, NANOGP1 and shared NANOG/NANOGP1 gene targets in the naïve hPSCs. Data was taken from the RNA-seq dataset produced in this study. Violin plots show mean gene expression for 72 NANOGP1-only, 366 shared and a randomised set of 200 NANOG-only targets. Gene expression values are in Log2 RPM (read per million). Individual values represent average of three biological replicates. Median and quartiles are shown. ANOVA statistical test with Dunnett's multiple comparisons was performed (****p < 0.0005, ns – not significant).

In summary, in this section, I used ChIP-seq in naïve hPSCs to show that epitope-tagged NANOGP1 has a small set of unique chromatin binding sites that do not overlap with NANOG

occupancy, as well as a greater number of sites that are targeted by NANOGP1 and NANOG. Sequences at NANOGP1 only peaks were strongly enriched for REST motifs, and re-analysis of published data corroborated this finding by showing that a large proportion of NANOGP1 only peaks were bound by REST in primed hPSCs. Predicted genes targets have roles associated with neuronal cells, in keeping with the known function of REST to target neuronal genes in non-neural cell types. Most NANOGP1 binding sites were found in 'background' chromatin and active promoters. In contrast, the majority of shared binding sites were found in active promoters, and their target genes were also significantly higher expressed in naïve hPSCs, when compared to NANOGP1 only and NANOG only target genes. The data analysis here had certain limitations; for example, it was not clear whether the NANOGP1 only signal was absent in Chovanec et al, 2021, which used the same antibody. This will have to be investigated further. If the data is indeed meaningful, it opens up a plethora of exciting new research avenues. For instance, this would mean that NANOGP1 could have developed a NANOG-independent way of regulating the neuronal differentiation, due to its high association with REST. The role of shared peaks appeared to be associated with pluripotency and not differentiation, although presence of a highly represented NKT-cell category in the GO analysis could mean that some other functional divergence may become possible when NANOG and NANOGP1 interact. In any case, this section demonstrated a valuable preliminary dataset, showing that NANOGP1 could be capable of binding chromatin and therefore regulating gene expression, either in a manner similar to NANOG, or in a new unknown way.

4.2.3 Investigating NANOGP1 homodimerisation and NANOGP1/NANOG heterodimerisation

NANOG is known to act as a dimer, and is also known to be capable of forming heterodimers with other proteins (Section 1.3). Here, I aimed to investigate whether NANOGP1 could form dimers as well. As mentioned above, there was no antibody available that would pick up endogenous NANOP1 in the naïve context. Additionally, the epitope tagged lines (Section 4.2.1) contained only a small proportion of cells with the epitope tagged NANOGP1. To study protein dimerisation, I needed a more robust and high scale method that would allow extraction of large amounts of protein to be analysed by size exclusion chromatography. To achieve this, I produced recombinant proteins in Sf9 insect cells (Figure 4.30), followed by Ni^{+}-NTA protein pull-down and AKTA size exclusion chromatography, and Western blotting to assess protein interactions.

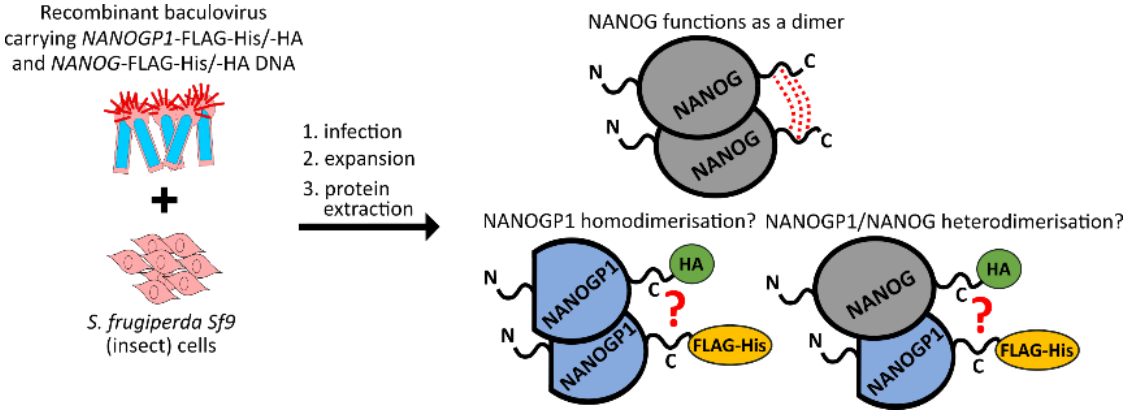


Figure 4.30 Summary of the recombinant protein assay, used to study the ability of NANOGP1 to form homodimers and heterodimers with NANOG. N, C – N-terminus, C-terminus. HA, FLAG-His – epitope tags. Dimerisation is indicated by red dotted lines.

The first goal was to clone *NANOGP1* (isoform 1) and *NANOG* CDSs in frame with an epitope tag (two constructs each, with HA and FLAG-His tags) into the bacmid DNA. The choice of tags was based on the method itself: the protein labelled with a FLAG-His tag was to be pulled-down by a Ni^{+}-NTA column, and the protein labelled with HA could only be detected in the final Western blotting if the proteins dimerise. Therefore, CDSs were first inserted into pBluescript plasmids to add the tags. Tagged CDSs were then ligated into an entry vector, pEntr3c, and then into a constitutive expression plasmid based on the destination vector pDest8. The expression vector was transformed into DH10Bac *E.coli* where it was inserted into the bacmid DNA with help of an enzyme transposase. The approach was based on standard Bac-to-Bac recombinant protein synthesis (Invitrogen). A detailed description of this experiment can be found in Chapter 2. A cloning summary is shown in Figure 4.31.

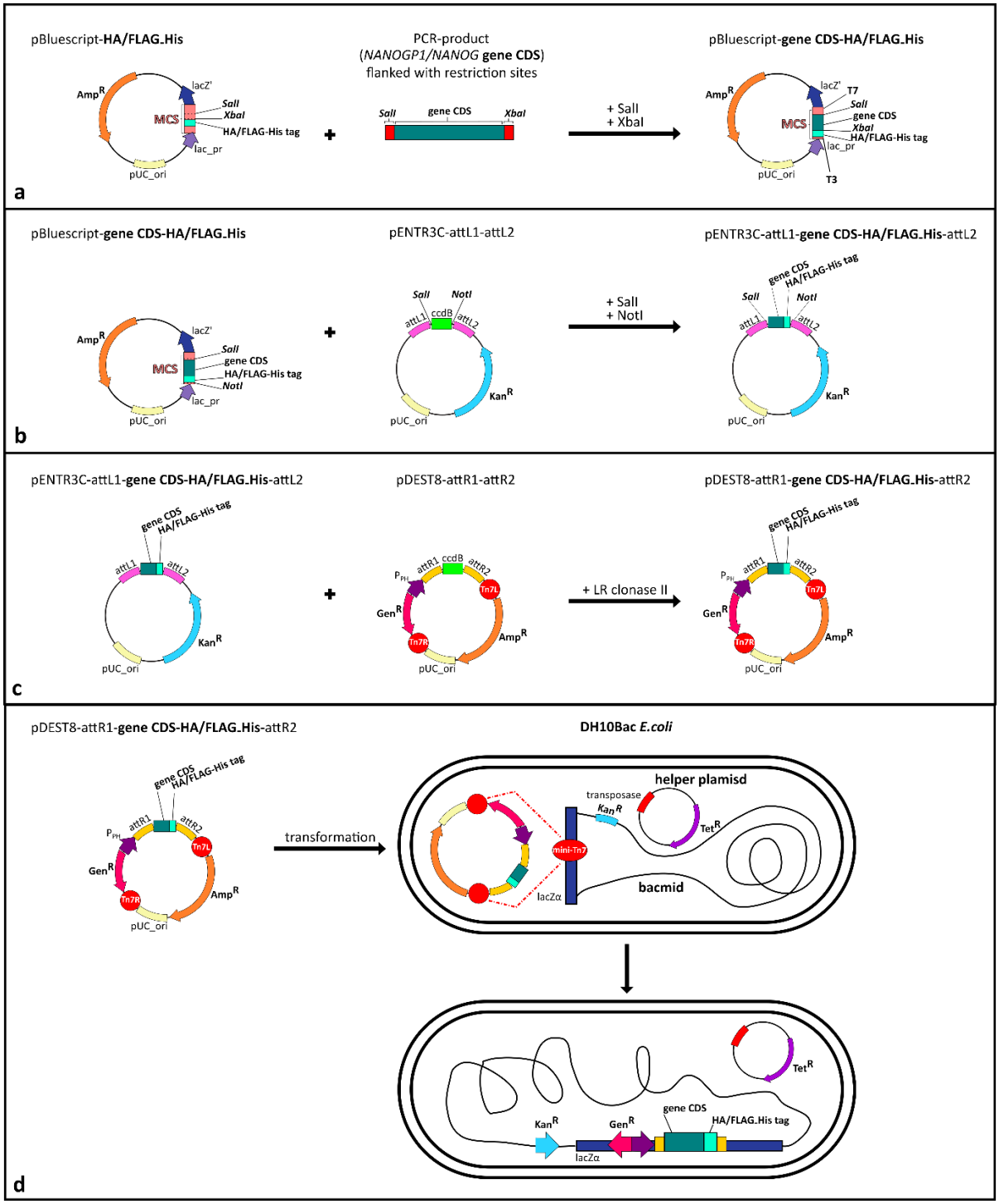


Figure 4.31 Summary of the cloning performed to generate epitope-tagged NANOGP1 and NANOG protein. See Chapter 2 for detailed cloning description.

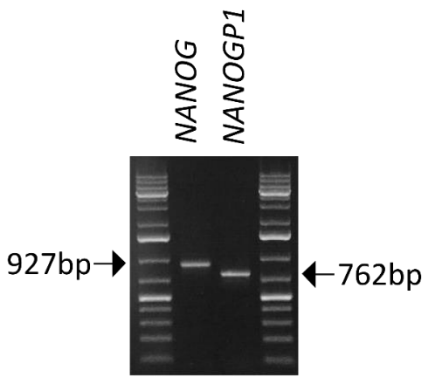
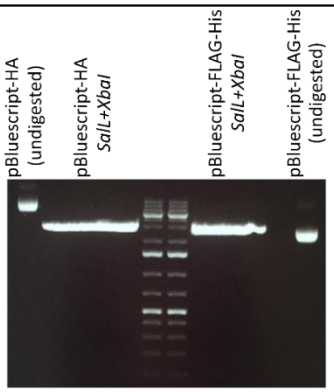
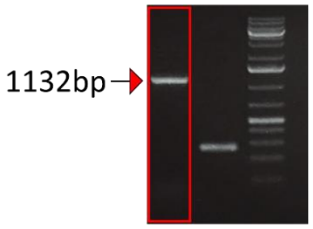
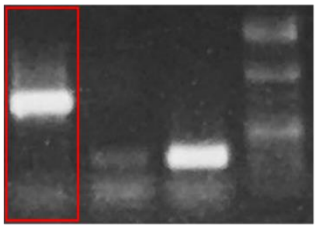
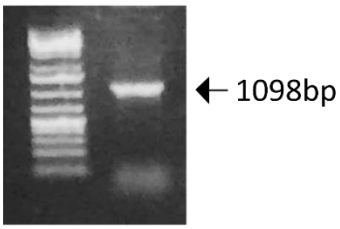
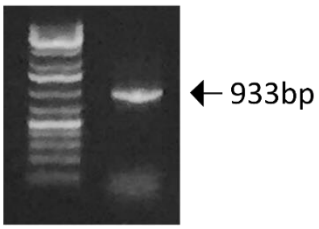
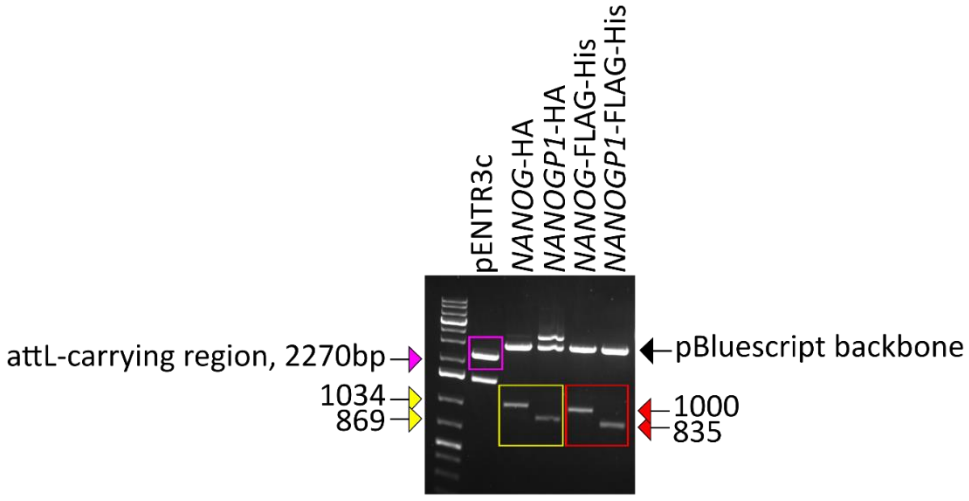
SalI-CDS-XbaI, undigested	pBluescript-tag, SalI+XbaI digested
	
pBluescript-CDS-tag, bacterial colony genotyping with T7_F +T3_R primers PCR	
pBluescript-*NANOG*-HA 	pBluescript-*NANOGP1*-HA 
pBluescript-*NANOG*-FLAG-His 	pBluescript-*NANOGP1*-FLAG-His 
pENTR3c and pBluescript-CDS-tag, SalI+NotI digested	
	

Table 4.2 – continued on the next page

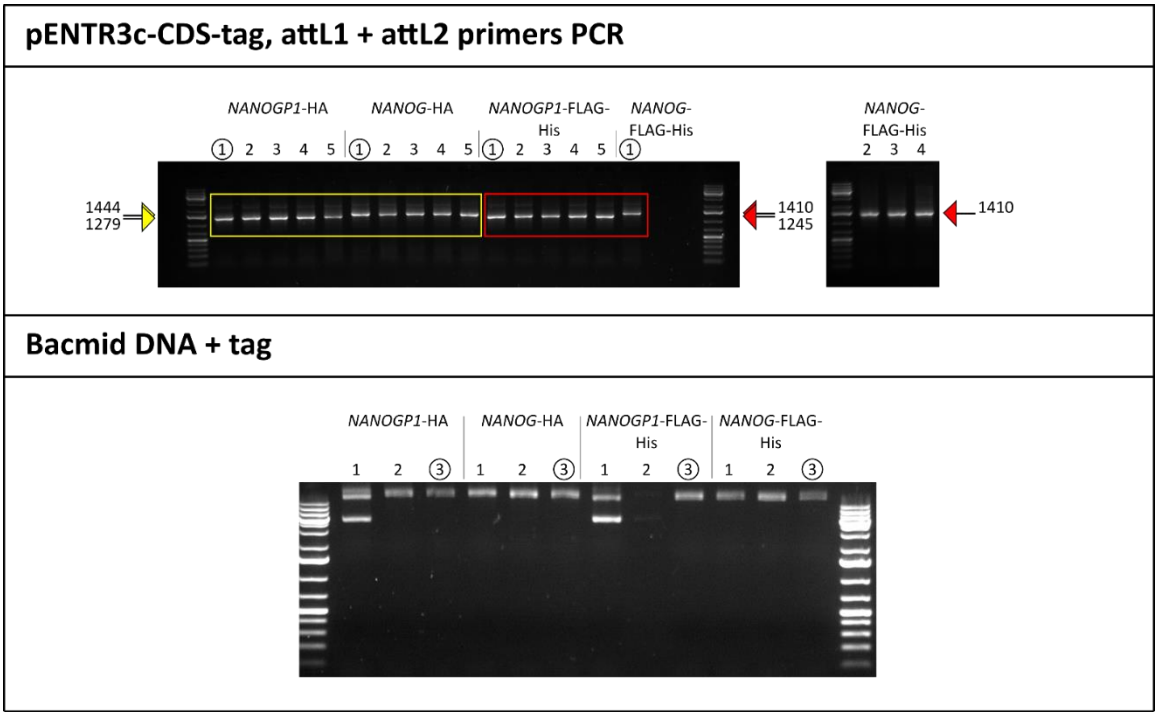


Table 4.2 Gel electrophoresis images showing sequential cloning validation steps for the recombinant protein syntheis. See Chapter 2 for detailed cloning description.

After the bacterial culture selection, each bacmid DNA construct (*NANOG*-HA, *NANOG*-FLAG-His, *NANOGP1*-HA and *NANOGP1*-FLAG-His) was individually used to transfect Sf9 insect cells. Each of the four insect cell lines were expanded and the first baculovirus stock was extracted 5 days later (P1), using culture centrifugation and supernatant filtering. Fresh insect cultures were then used to increase the titer of the filtered baculovirus in two more rounds of infection (P2 -> P3), additionally validating the synthesis of the recombinant protein in the insect cells using the Coomassie staining (Figure 4.32).

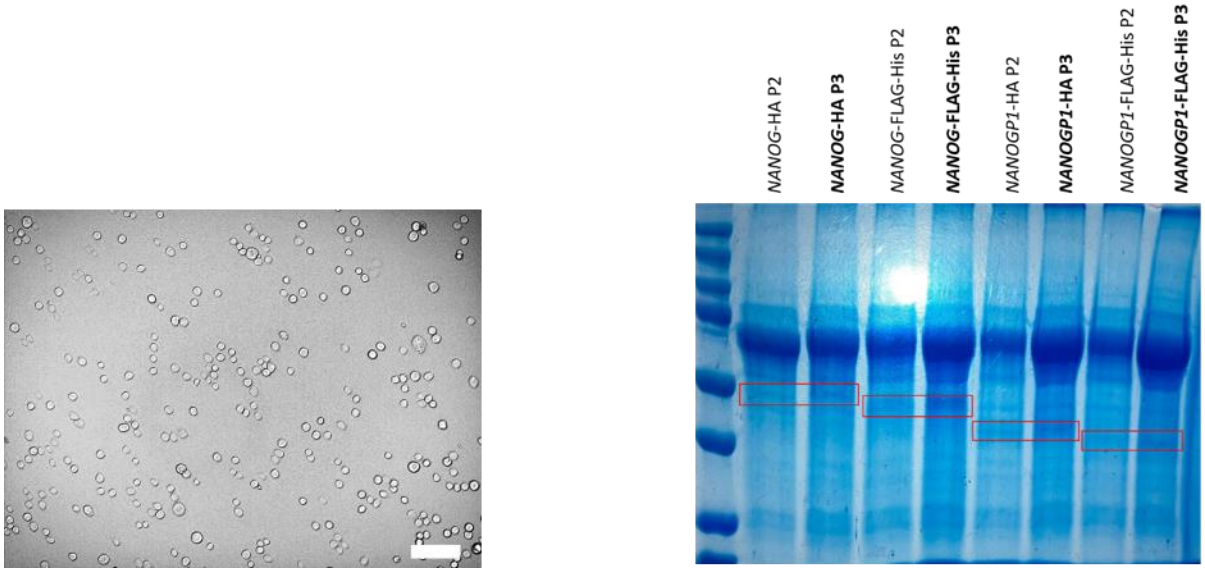


Figure 4.32 Bright-field image of the Sf9 insect culture used in recombinant protein synthesis (left).

Coomassie staining of a PAGE-SDS gel, showing presence of the recombinant protein in insect cultures used for producing of P2 and P3 viral stock (right). Scale, 100 μm. Proteins of interest are indicated by red rectangles.

Baculovirus P3 stocks were then used in different combinations (described below) to infect a large volume of fresh insect cell culture. In this step, I aimed to produce a large quantity of co-expressed recombinant proteins. Therefore, cell pellets were collected and snap frozen 2.5 days post-infection, prior to the virus-induced cell lysis that would have destroyed the cells carrying the protein-encoding constructs, as well the protein itself.

To create the different combinations, equal volume fractions of the P3 viral stock were combined to produce cell cultures that synthesised two proteins at the same time. The protein combinations used here were:

1. NANOG-HA + NANOG-FLAG-His
2. NANOGP1-HA + NANOGP1-FLAG-His
3. NANOG-HA + NANOGP1-FLAG-His
4. NANOG-FLAG-His + NANOGP1-HA

Cell culture #1 was used as a positive control sample, validating whether NANOG protein can homodimerise in insect cells, as would be expected. Cell culture #2 was used to determine the ability of NANOGP1 to homodimerise. Cell cultures #3 and #4 were used to reciprocally analyse any potential heterodimerisation between NANOGP1 and NANOG.

Proteins from the pellets were extracted and passed through a gravity column with Ni2+-NTA beads to bind His-tagged proteins. If dimerisation was occurring between the pair of proteins co-expressed by the insect cells, then the HA-tagged protein would also have been pulled-down during the His-tag isolation, and both proteins would bind to the column. After a series of washes, the proteins in the column were eluted in imidazole/glycerol-based buffer.

Eluted protein samples were analysed by size exclusion chromatography using an AKTA Pure system. The output of this experiment is typically a chromatogram with multiple peaks. Depending on the size and elution time of each peak I was able to detect (a) presence of the protein of interest (POI) in the sample, (b) its size, (c) its amount and (d) formation of POI dimers or other aggregates. A POI at approximately the size of tagged NANOG and NANOGP1 proteins would be expected to generate a peak in the ~70 ml fraction as a monomer (Figure 4.33). Dimerised proteins would produce a larger peak that is separated from the monomeric peak in earlier fractions. The imidazole peak at the end of the extraction is used as a counter ligand in size exclusion chromatography, binding to metal ions and later being displaced by His-tagged proteins. Eventually, imidazole leaves the column, in the ~120 ml

fraction. Larger peaks shown in the POI example graph below (specifically the 41-57 ml fractions) represent experimental artefacts, likely involving protein aggregates that are stuck in the column or larger non-specific protein complexes.

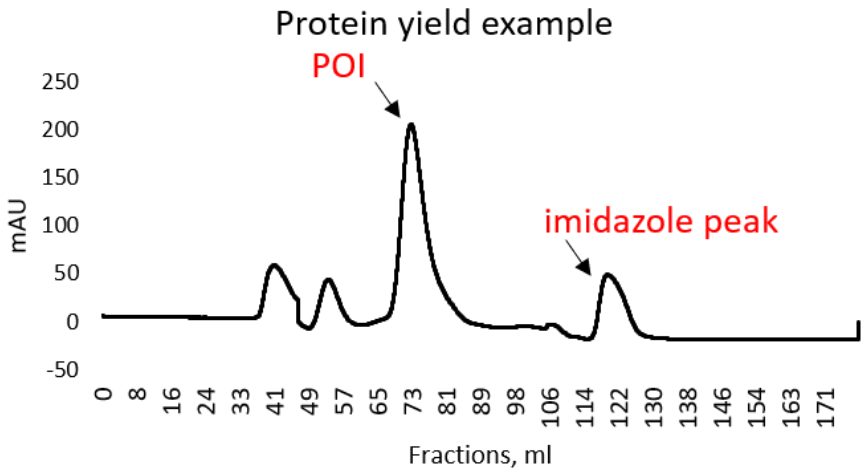


Figure 4.33 Example of a size exclusion chromatogram for purification of cell lysate solution, using the AKTA Pure system. X-axis, fractions produced by the AKTA system, in ml. Y-axis, absorbance, in mAU (milli-absorbance unit). POI – protein of interest. Data generated with a collaborator, Katie Mullholland (unpublished).

Unfortunately, when purifying the protein samples containing tagged NANOG and NANOGP1, an unexpected technical obstacle was encountered: even though each pellet had been produced from 2 L of log-phase insect cell culture, very little protein was bound to the beads within the column. Instead, most of the protein aggregated and remained on top of the column. In an attempt to eliminate this technical obstacle, the following adjustments were made: increasing/decreasing protein sample incubation time, adding/removing triton in the lysis buffer, as well as attempting to adjust the amount of resin used in the experiment. However, none of the above improved the protein yield. Therefore, the experiment was continued albeit with a smaller amount of protein than expected initially. At the AKTA purification step, another issue was encountered: the proteins that were bound to the column also seemed to aggregate while being passed though the size exclusion system, leaving most of the protein sample as an aggregate that was separated in one of the early fractions and is visible as one large peak on the spectrum (37-45 ml fractions). Despite these technical issues, in the NANOGP1-FLAG-His + NANOGP1-HA protein pull-down sample it was possible to distinguish two small peaks at the expected sized of a monomer and dimer (Figure 4.34),

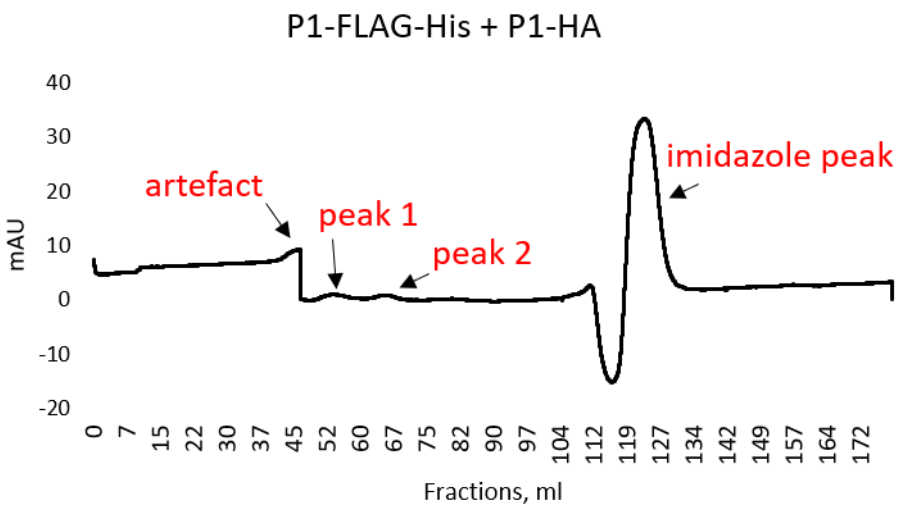


Figure 4.34 Size exclusion chromatogram showing two protein peaks obtained in analysing NANOGP1-FLAG-His+NANOGP1-HA lysate. Peak 1 and peak 2 indicate two protein peaks of different size. Artefact represents a protein aggregate that was not separated during the size-exclusion chromatography. X-axis, fractions produced by AKTA system, in ml. Y-axis, absorbance, in mAU (milli-absorbance unit).

Fractions corresponding to these two peaks in the NANOGP1-FLAG-His + NANOGP1-HA co-expression samples were collected, purified and analysed separately by Western blotting (Figure 4.35). The protein represented by Peak 2 (smaller size protein/protein complex) produced a signal when probed with the C-terminal NANOG antibody, and by the FLAG antibody, but not with the HA antibody. This protein was therefore concluded to be a NANOGP1 monomer (a faint HA signal in Peak 2 is due to the slight overlap between Peaks 1 and 2 as it was impossible to identify a clean boundary between them). The protein represented by Peak 1 (larger protein/protein complex) reacted with antibodies against NANOG, FLAG and, importantly, HA, and was therefore concluded to represent the NANOGP1 homodimer. This result showed that recombinant NANOGP1 was capable of homodimerising when expressed in insect cells.

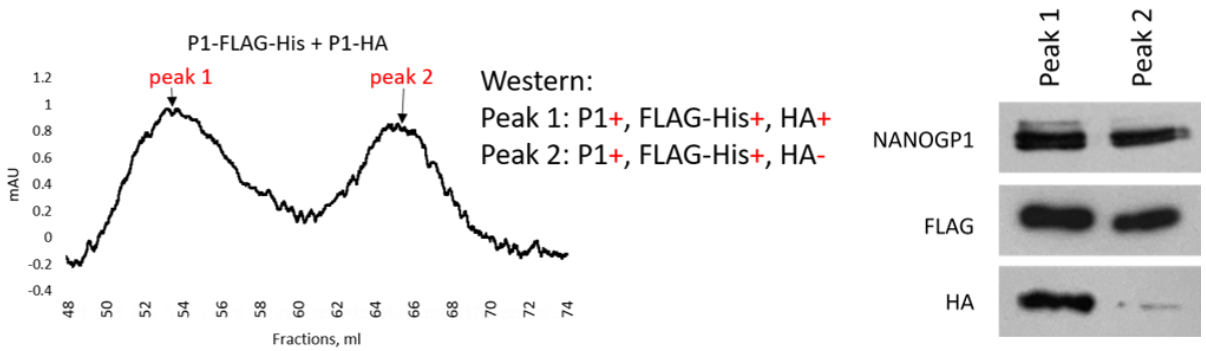


Figure 4.35 Zoomed in view of the the size exclusion chromatogram (Figure 4.34), showing two separate protein peaks (left) and a Western blotting image of the two samples corresponding to

these peaks (right). X-axis, fractions produced by AKTA system, in ml. Y-axis, absorbance, in mAU (milli-absorbance unit).

The quality of AKTA size exclusion chromatography performed on the remaining three samples was lower than that of NANOGP1-FLAG-His + NANOGP1-HA sample. Due to the very low amount of protein obtained in the pull-down step and a relatively large aggregate forming in one of the early fractions, it was impossible to distinguish two or more separate peaks (Figure 4.36).

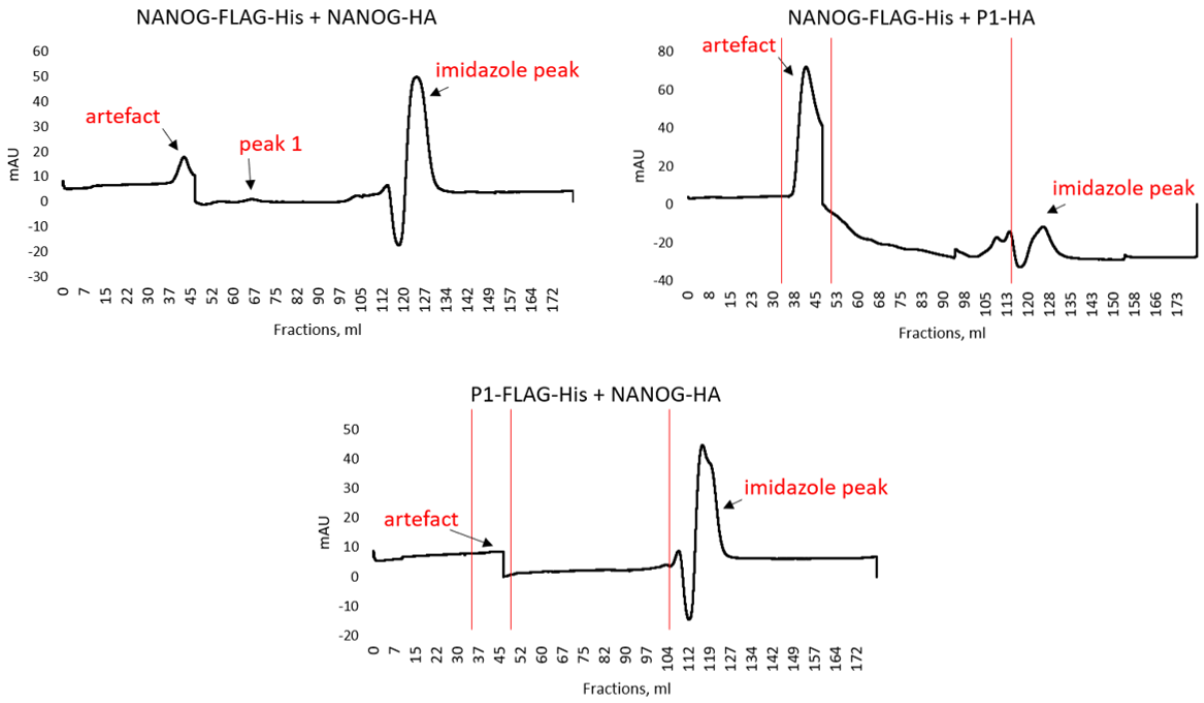


Figure 4.36 Size exclusion chromatogram of NANOGP1-FLAG-His + NANOG-HA, NANOG-FLAG-His + NANOG-HA and NANOG-FLAG-His + NANOGP1-HA lysates. Peak 1 - possible individual protein peak. Artefacts represent protein aggregates. X-axis, fractions produced by AKTA system, in ml. Y-axis, absorbance, in mAU (milli-absorbance unit). Vertical red lines separate the artefact peak from the area were the protein peaks separated by size were expected to be.

Nevertheless, the fractions where the peaks would have been present, and the artefact aggregate peak fractions, were collected and analysed by Western blotting. As a result, I was able to detect the following:

1. The NANOG-FLAG-His + NANOG-HA sample reacted with NANOG, FLAG and HA antibodies, confirming the expected pattern of NANOG homodimerization.
2. The NANOGP1-FLAG-His + NANOG-HA samples: Peak 1 reacted with NANOG and FLAG antibodies, and Peak 2 reacted with NANOG and HA antibodies.
3. The NANOG-FLAG-His + NANOGP1-HA sample: Peak 1 reacted with NANOG and FLAG antibodies, and Peak 2 reacted with NANOG and HA antibodies.

These results demonstrated that, in all cases, an HA-tagged protein was pulled-down with the FLAG-His-tagged protein, showing that recombinant NANOGP1 is capable of homodimerisation and also of heterodimerisation with NANOG. Due to the technical complications described above, it was not possible to analyse all four samples on the same Western blotting gel (as shown in Figure 4.37) and therefore the data obtained from the four separate blots were combined and summarised in Figure 4.37. All protein bands were of expected size and in line with the predicted protein separation shown in the Western blotting schematic below.

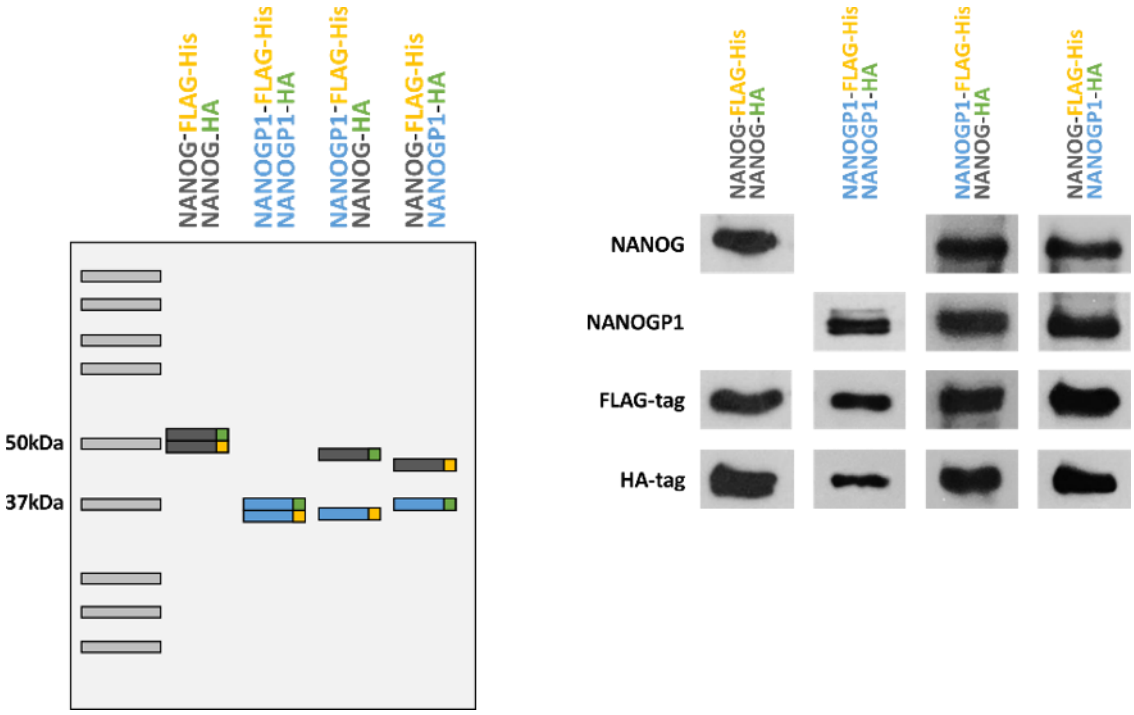


Figure 4.37 Summary of NANOGP1 and NANOG recombinant protein experiment; western blot schematic (left). Western blotting results obtained for the four insect cell lysates (right). 50 kDa and 37 kDa – two main marker bands used for distinguishing NANOG and NANOGP1 proteins by size. NANOG, NANOGP1, FLAG-tag, HA-tag were detected by NANOG C-term, NANOG C-term, anti-FLAG and anti-V5 antibodies, respectively.

In summary, in this final section I demonstrated that recombinant NANOGP1 protein (isoform 1) is capable of forming homodimers, as well as heterodimers with NANOG. Due to technical complications, it was not possible to perform size exclusion chromatography on all samples or quantify the data. Nevertheless, these results led to a conclusion that NANOGP1 has preserved the functional property to form homo- and heterodimers, supporting my main hypothesis that it could function and dimerise with other proteins in the same way NANOG does.

4.3 Discussion

Here I discovered that endogenous *NANOGP1* can form a stable, full-length protein. This finding rejects the model for *NANOGP1* activity and potential function that was proposed by Booth and Holland in 2004. According to their study, *NANOGP1* was most likely to use a truncated ORF, which would terminate translation only several amino acids into the protein synthesis. Based on this, databases, such as Ensembl, classified *NANOGP1* as non-protein-coding. My results suggest that this classification should now be updated.

In order to be able to detect endogenous NANOGP1, it had to be ectopically tagged, since no antibodies would have been able to distinguish it from NANOG. The successful tagging was replicated by using two different HDR templates, V5 and 3xFLAG, each additionally repeated twice in independent cell lines. Importantly, the crRNA sequence was chosen to provide good specificity to targeting the start codon of NANOGP1 compared to NANOG. However, the epitope tagging experiment produced another unexpected result: as seen in Figure 4.12 the molecular weight of the NANOGP1 protein was larger than originally expected. It is possible that the protein had undergone post-translational modifications. Also, as mentioned in Section 3.2.5, it was not clear where the *NANOGP1* mRNA would terminate. Thus, it was only assumed that the stop codon would be the same as that of NANOG, while the *NANOGP1* 3'-end in fact contained six additional stop codons downstream of the predicted one. Overall, this could affect the size of the NANOGP1 protein and where its final domain would terminate. Nevertheless, the ability to be detected by the C-terminal NANOG antibody and the inability to be detected by the N-terminal NANOG antibody collectively means that the overall predicted domain structure of the NANOGP1 protein was correct. Therefore, I have confidence that despite *NANOG* and *NANOGP1* having very similar structure and sequences, the tagged protein described here is NANOGP1, not NANOG.

After discovering that *NANOGP1* encodes a protein, I proceeded to investigate its chromatin binding profile in naive hPSCs. The dataset produced had several uncertainties, such as very strong association of NANOGP1 only peaks with REST binding, which was not detected by the NANOG C-terminal antibody, expected to bind both NANOG and NANOGP1. Additionally, it was not clear whether NANOGP1 and NANOG were co-binding (as dimers) in the shared peaks, or that both were capable of binding the shared motif but did not interact with each other. Finally, less than 500 NANOGP1 peaks vs. 70,000+ NANOG peaks were identified. The low efficiency could be attributed to the heterogeneous cell population and/or the target detection method (antibody targeting an epitope tag vs. endogenous protein, respectively). To improve the quality of NANOGP1 ChiP-seq, I propose several possible approaches. First, it would be highly beneficial to tag NANOG and perform its ChIP-seq the same way as the ChiP-seq of NANOGP1 was conducted in this study. This way we would be

able to eliminate the bias of comparing a signal produced by a tag antibody vs a signal produced by a protein-specific antibody. This could also improve the quality of statistical analysis as it might improve the ratio between NANOG and NANOGP1 peaks. Moreover, another tagging experiment could be performed, using another sequence - HA, for instance - followed by testing if the same REST-specific peaks could be observed. Finally, it would also be interesting to attempt to 'cut out' the *NANOGP1*-tag locus, use these cells in ChiP-seq and check whether the REST-associated peaks disappear. It would also be interesting to perform a mass spectrometry and/or co-immunoprecipitation analysis of the NANOGP1 pull-down sample and identify what proteins it interacts with, i.e., can we actually detect REST and/or NANOG there?

Nevertheless, we can still hypothesise that the ChiP-seq data had biological significance. If so, it could mean that the presence of unique NANOGP1 binding with a divergent chromatin state profile reflected that NANOGP1 had developed new protein interactions and properties that are not present for NANOG.

What could those novel protein interactions be? To answer this question, it is first important to discuss the known, as well as predicted, protein interactions of NANOG. In primed hPSCs, NANOG directly interacts with a SMAD2/3 complex and guides it to the promoters of pluripotency-associated genes, which eventually become transcriptionally activated, in response to the NANOG/SMAD binding (Vallier et al., 2009). In the alternative scenario, when NANOG is absent, SMAD2/3 instead binds to the promoters of differentiation-associated genes, which promotes stem cell differentiation. How NANOG interacts with SMAD2/3 proteins is poorly understood, although one study identified a SMAD4-like domain in the N-terminus of NANOG (Hart et al., 2004). This let me hypothesise that the interaction could be occurring via that region and, interestingly, this region overlaps with the NANOGP1 N-terminal deletion. However, more importantly, it is not currently known whether this interaction is two- or one-sided. Does the SMAD2/3 complex also affect where NANOG can bind? If it does, then NANOGP1 protein would not be regulated in the same manner because the SMAD4-like domain is absent from its truncated N-terminal sequence. Collectively, this could mean that NANOGP1 was excluded from an important (hypothetical) expression regulation pathway and, possibly, has other proteins regulating its expression.

In my opinion this hypothesis is promising, but is still only an assumption, mostly based on secondary evidence and, therefore, this would have to be tested in future experiments. One such experiment could involve testing whether SMAD2/3 proteins overlap with the NANOGP1-only peaks. This could be performed by cross-comparing NANOGP1, NANOG and SMAD2/3 ChiP-seq datasets.

It is also important to note that the NANOG/SMAD2/3 interaction, on which the hypothesis is based, was studied in primed hPSCs (Vallier et al., 2009), and not much has been known for a while

about TGF-β/ACTIVIN/NODAL/SMAD signalling in naïve hPSCs. However, a recent study discovered that the TGF-β/ACTIVIN/NODAL signalling pathway, required for maintenance of human primed pluripotency, is active and also important in the naïve state. A study by Osnato and colleagues demonstrated that downstream effector proteins of TGF signalling, SMAD2/3, have common binding sites in naive and primed hPSCs, including at the promoters of pluripotency-associated genes. Additionally, most TGF pathway components are also expressed in naïve and primed hPSCs at a similar level (Osnato et al., 2021). Therefore, it is possible that the SMAD2/3 complex interacts with NANOG (and cannot interact with NANOGP1) in the naïve hPSC context as well, and hence, the hypothesis described above could be applied to the naïve hPSCs too.

The hypothesis is summarised in Figure 4.38.

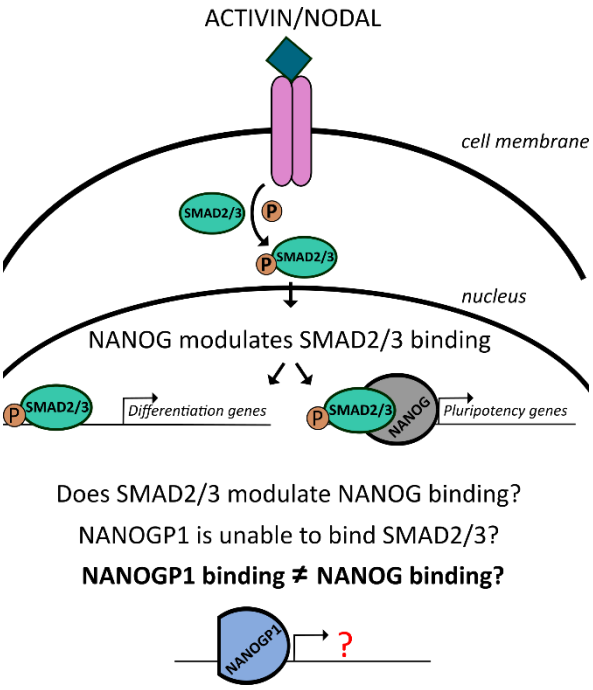


Figure 4.38 Diagram showing the interaction between NANOG and SMAD2/3 complex, and the hypothesis regarding the NANOGP1 interaction with the complex.

Partial deletion of the N-terminus of NANOGP1 could have led to it developing new protein interactions and, consequently, novel chromatin binding patterns. Alternatively, the novel interaction could be influenced by the S285N substitution mutation within the NANOGP1 transactivation domain. Finally, other unknown processes could have caused this protein-protein interaction.

The data shown in this chapter suggests that REST might be one of the major protein interactions NANOGP1 could be involved in. This however does not mean that REST itself is a novel

interactor for NANOGP1, because REST is known to interact with NANOG as well. In primed hPSCs, REST is regulated by OCT4/SOX2/NANOG and although REST is not essential for maintenance of pluripotency, REST is needed for regulating cell survival (Thakore-Shah et al., 2015). Mouse ESCs studies have also demonstrated that REST is regulated by OCT4/NANOG/SOX2, and moreover, Nanog itself is the direct target of Rest (Johnson et al., 2008; Loh et al., 2006). Finally, NANOG and REST were shown to directly interact at the protein level in mouse (Johnson et al., 2008; Wang et al., 2006). Published literature shows that REST had developed the ability to independently interact with a vast array of different proteins, such as CoREST, N-CoR, mSin3A, SCP and others, via both its N-terminal and C-terminal domains (Lunyak and Rosenfeld, 2005). These interactions are normally involved in restricting the neural fate in non-neural lineages.

Collectively, this shows that REST is a protein with a large interactome, and is also deeply integrated into the pluripotency network. Being a protein that developed an ability to interact with a large number of cellular factors, it would not be surprising if it developed an ability to interact with NANOGP1. If that were true, it would open up new possible interactions for NANOGP1, present in a complex, 'gathered' by REST, and potentially not involving NANOG (an instead, for instance, interacting with NANOG in a different complex). Therefore, NANOGP1 could have become involved in the restriction of neural lineage development at a new, NANOG-independent level. This could be supported by the detection of neural-lineage associated genes as potential targets of NANOGP1 in the NANOP1 only binding peaks.

In this chapter, I showed that NANOGP1 can dimerise with itself, as well as form heterodimers with NANOG, outside of the naïve context. This was tested using recombinant and tagged NANOGP1 isoform 1, which means that such a result was not completely unexpected, as the key domain mediating the dimerisation activity is conserved in the predicted NANOGP1 structure. It is still unclear whether NANOGP1 can dimerise with NANOG in the naïve context. This of course could be a technical limitation that would require optimisation of the co-immunoprecipitation experiment in naïve hPSCs. It is also possible that that NANOG/NANOG, NANOG/NANOGP1 and NANOGP1/NANOGP1 dimers have different stability, or that NANOG:NANOGP1 dimers simply cannot form a complex. Overall, this was one of the unexpected results that will have to be investigated in the future.

In conclusion, here I showed that NANOGP1 encodes a protein in naïve hPSCs, and could have overlapping, but not completely identical chromatin binding properties, with NANOG. Recombinant NANOGP1 was also shown to be capable of dimerising with itself and with NANOG, likely due to the conserved dimerisation domain. Overall, this chapter supports the hypothesis that NANOGP1 is a

conserved copy of NANOG. Other potential functional similarities of NANOG and NANOGP1 are tested further in Chapter 5.

5 Investigating *NANOGP1* function in naïve hPSCs

5.1 Introduction

5.1.1 Background

After discovering that NANOGP1 protein is present in naïve hPSCs, I next aimed to investigate whether it has any functional roles in establishing or maintaining naïve pluripotency. The ancestral copy of the pseudogene – *NANOG* – has several known functions in naïve hPSCs, such as, for instance, restricting naïve hPSCs from upregulating some trophectoderm-lineage markers (Guo et al., 2021) and reprogramming primed hPSCs towards the naïve state when overexpressed with *KLF2* (Takashima et al., 2014; Theunissen et al., 2014). Here, these two aspects of *NANOG* functionality were tested in relation to *NANOGP1*. Another interesting property of *Nanog*, shown in mouse, is an ability to regulate its own expression via autorepression (Navarro et al., 2012). Here I also test whether *NANOGP1*, as well as *NANOG*, have the same autorepressive function in naïve hPSCs.

A recent study has shown that cultures of *NANOG*-deficient naïve hPSCs upregulate several trophectoderm lineage marker genes, thereby revealing a potentially crucial role for *NANOG* in maintaining naïve pluripotency (Guo et al., 2021). However, the dynamics of the transcriptional response following *NANOG* perturbation, and the effect on gene expression programmes beyond individual marker genes, has not been examined. Moreover, because NANOGP1 potentially binds to shared and unique genomic sites (Section 4.2.2), and has conserved protein sequence and expression profiles (Section 3.2), it is important to establish whether the loss of *NANOGP1* expression also leads to cell differentiation.

Since the predicted NANOGP1 protein encodes a functional homeodomain that is identical to that of NANOG at the amino acid sequence level (Section 3.2.5), I also hypothesise that NANOGP1 could be capable of inducing primed-to-naïve hPSC reprogramming when ectopically expressed.

Ectopic *Nanog* overexpression in mouse serum-free ESCs (equivalent of naïve hPSCs) leads to the transcriptional repression of endogenous *Nanog* expression by an unknown mechanism that involves NANOG binding within a 5 kb region upstream from its gene sequence (Navarro et al., 2012). Section 3.2.7 demonstrated that the 5 kb region upstream of *NANOGP1* is conserved between *NANOG* and *NANOGP1* and was likely created during the tandem duplication event. This led me to hypothesise that in addition to potentially conserved autorepression function that NANOG and NANOGP1 could have in naïve hPSCs, they could also repress the expression of each other, due to the highly conserved duplicated putative promoter region.

5.1.2 Aims

1. Design and validate RT-qPCR primers to distinguish *NANOG* and *NANOGP1* expression.
2. Study the role of *NANOGP1* in naïve hPSCs using a gene expression downregulation approach.
3. Investigate *NANOGP1* function in human pluripotency using gene overexpression assays.

5.2 Results

5.2.1 Distinguishing between *NANOG* and *NANOGP1* RNA expression in hPSCs

To be able to distinguish between *NANOGP1* and *NANOG* expression in hPSCs, two RT-qPCR primer pairs were designed for both genes. The primer pairs for *NANOGP1* and *NANOG* were designed to bind to the same regions in their respective sequences but that had sufficient sequence diversity. In total, each primer pair had at least five nucleotide mismatches to provide target-specific binding and amplification (Figure 5.1).

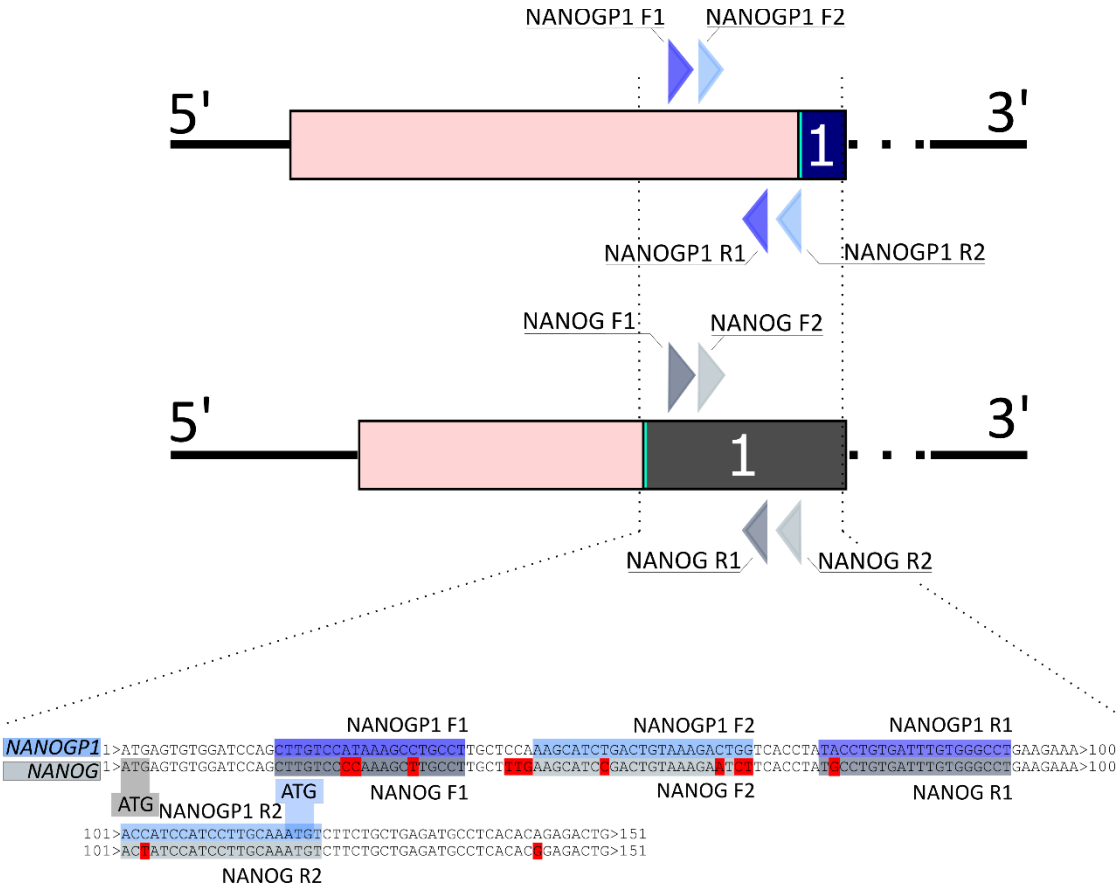


Figure 5.1 Schematic showing *NANOGP1* and *NANOG* RT-qPCR primer design. Binding sites for the primer pairs are indicated by arrows (top) and highlighted in blue (*NANOGP1*) and grey (*NANOG*) at the bottom of the figure. F1/2, R1/2 – forward and reverse primer orientation in their respective primer pairs. Mismatching nucleotides are in red. Exon 1 is shown as blue and black blocks, and 5'-UTR as a pink block. 5' and 3' indicate the nucleotide sequence orientation. Start codon is labelled as a green vertical line.

After the primers were synthesised, they were tested in four different hPSC lines: one primed (H9), and three naïve (CR-H9, NK2-H9 and HNES1) (Figure 5.2). Both primer pairs produced the expected result: *NANOG* expression was ~4-fold higher in the naïve lines compared to the primed, while *NANOGP1* expression was ~20-fold higher in the naïve state, similar to the previously observed differences between the expression patterns of these two genes (Chapter 3). These values were obtained by averaging the gene expression values obtained using the two primer pairs for *NANOG* and *NANOGP1*. RT-qPCR results produced by the primer pairs *NANOG*_PP1 and *NANOGP1*_PP1 were more consistent between the tested biological replicates then their corresponding PP2 pairs, therefore, *NANOG*_PP1 and *NANOGP1*_PP1 were used further in this study. The primers were further validated by gene-specific knockdown and overexpression experiments that are described later in this chapter.

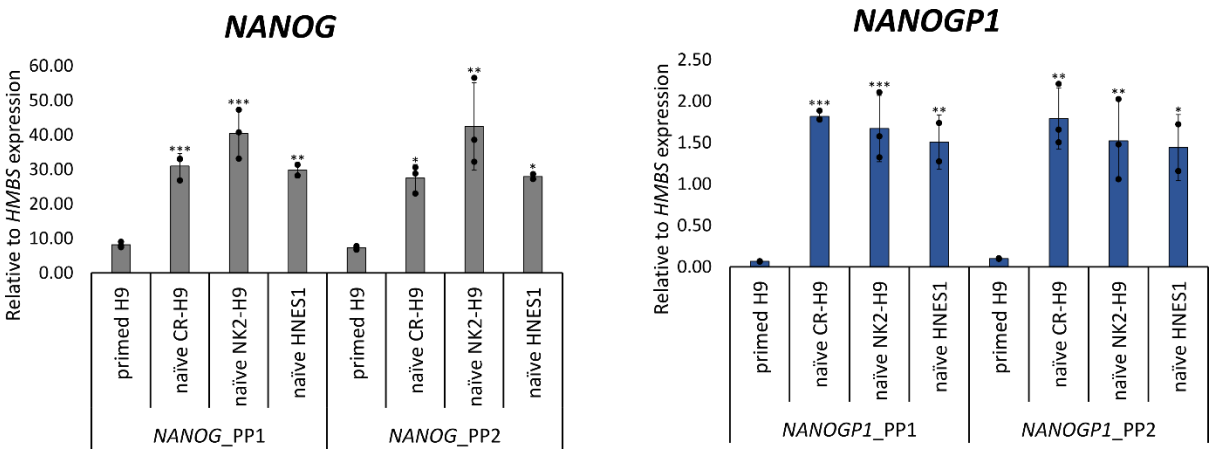


Figure 5.2 Bar charts showing *NANOG* and *NANOGP1* RNA expression in three naïve and one primed hPSC lines. RT-qPCR values are relative to *HMBS* expression values. Individual replicates (n=2 or n=3) and mean ± SD are shown. ANOVA with Dunnett's multiple comparison test was performed ($p < 0.05$ (*), $p < 0.005$ (**) and $p < 0.0005$ (***)); 'primed H9' was used as a control.

5.2.2 Development, validation and application of *NANOGP1* loss of function approach

In this section, gene mis-regulation assay CRISPRi was used to study *NANOGP1* function. This section therefore describes the development and validation of *NANOG* and *NANOGP1* knockdown systems, and their successful application in naïve hPSCs.

5.2.2.1 *NANOGP1* expression downregulation shows that the pseudogene is not required to maintain naïve pluripotency

5.2.2.1.1 CRISPRi gene expression knockdown approach

To investigate whether *NANOGP1* has a role in maintaining the naïve state, the transcriptional response and dynamics following the depletion of *NANOG* or *NANOGP1* in naïve hPSCs were studied using a CRISPR interference (CRISPRi)-based approach.

The CRISPRi gene downregulation system used in this study was developed in Gilbert et al., 2013, Gilbert et al., 2014 and Kearns et al., 2014, and successfully applied in primed hPSCs in Mandegar et al., 2016. Briefly, it involves the destablised Cas9 protein (dCas9) fused with a KRAB domain, mediating transcriptional repression. The destablised nuclease is unable to cut DNA, and instead, is used to recruit the KRAB protein to the TSS of the target gene by site-specific guide RNA (gRNA) (Figure 5.3).

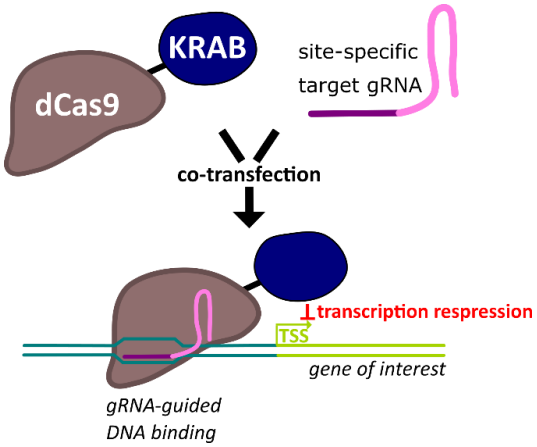


Figure 5.3 Diagram describing the CRISPRi dCas9-iKRAB gene expression downregulation tool. TSS – transcription start site. gRNA – guide RNA. dCas9 – destabilised Cas9.

In the hPSCs that I used (cell line CRISPRi Gen1B), a KRAB-dCas9-P2A-mCherry sequence is integrated into the *AAVS1* safe-harbour locus and is doxycycline-inducible (Mandegar et al., 2016). Additonally, the AAVS1 locus also contained a G418/neomycin resistance-encoding gene, and an reverse tetracycline-controlled transactivator (rtTa) sequence, that were driven by constitutive PGK and CAG promoters, respectively (Figure 5.4).

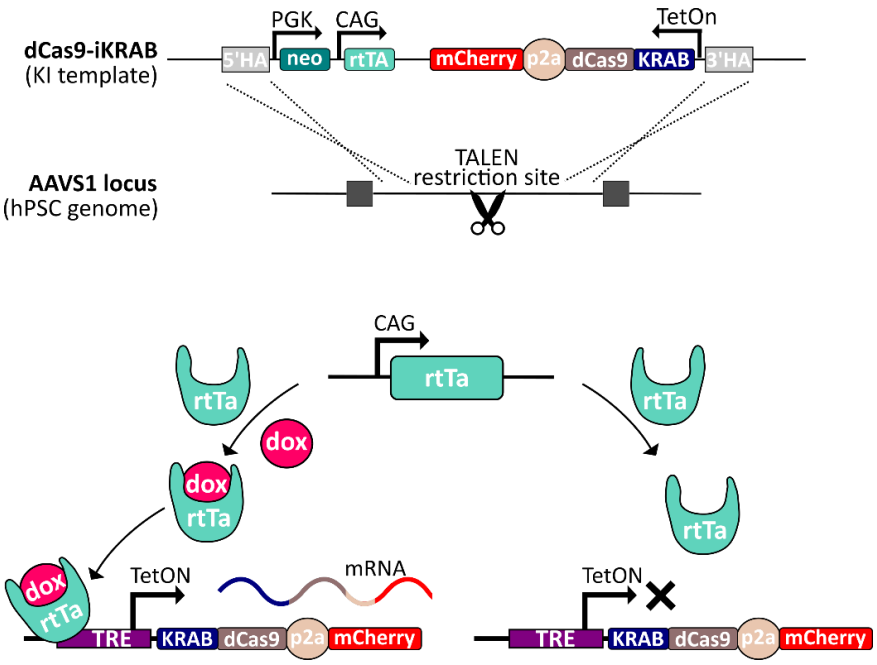


Figure 5.4 Diagram illustrating the AAVS1 locus and the dCas9-iKRAB construct structure (top) and the mechanism of the induced gene expression downregulation (bottom). KI – knock-in. 5'HA, 3'HA – 5' and 3' homology arms. Neo – neomycin. rtTa – transactivator. PGK, CAG, TRE/TetON – promoters. P2a – self-cleaving peptide. Dox – doxycycline.
Adapted from Mandegar et al., 2016.

5.2.2.1.2 CRISPRi *NANOG* and *NANOGP1* gRNA design

Based on Mandegar et al., 2016, not all gRNA molecules are equally efficient at inducing target gene knockdown. Therefore, I generated several gRNA designs, including gRNAs that were specific for either *NANOG* or *NANOGP1*, or that took advantage of sequence conversation to target both genes (Figure 5.5). In this way, the former two gRNAs were designed to create single gene expression knock-downs, and the latter gRNAs to create double gene expression knock-down. To ensure that the efficiency of gRNA is as high as possible, I only tested gRNAs that were designed to bind within the ±150 bp range of their respective TSSs (Mandegar et al., 2016). The designed gRNA sequences were cloned into a pgRNA-CKB constitutive expression plasmid (Mandegar et al., 2016) and were validated by Sanger sequencing. pgRNA-CKB vector contains a constitutive promoter CAG, driving the expression of gRNA, fluorescent reporter mKate2 and a blasticidin-resistance gene (Mandegar et al., 2016).

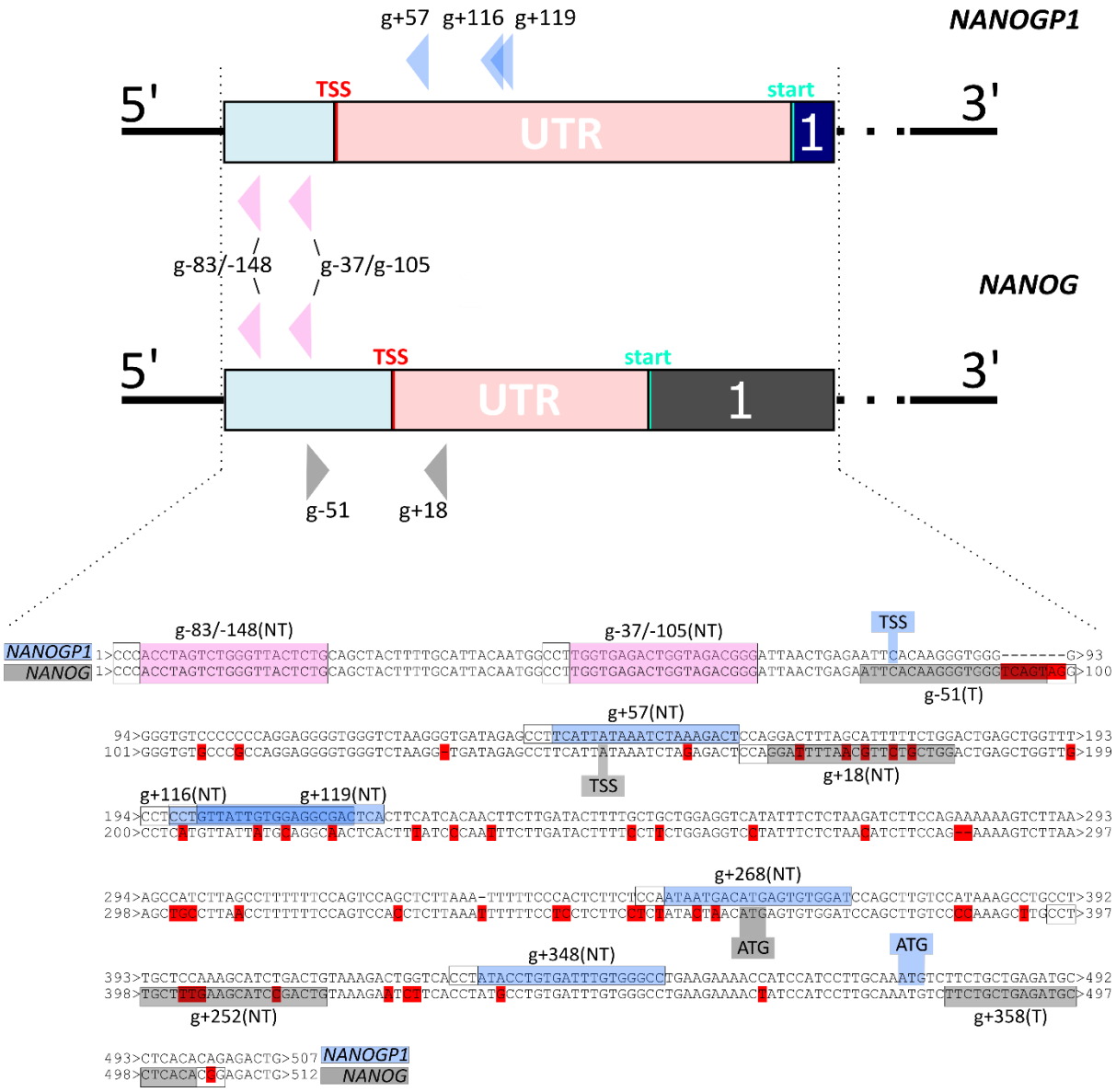


Figure 5.5 Schematic showing *NANOGP1* and *NANOG* CRISPRi gRNA design. gRNA binding sites are indicated by arrows (top) and highlighted in blue (*NANOGP1*), grey (*NANOG*) and pink (shared) at the bottom of the figure. T/NT indicates the coding strand. Mismatching nucleotides are in red. Exon 1 is shown as blue and black blocks, and 5'-UTR as a pink block. 5' and 3' indicate the nucleotide sequence orientation. Start codon is labelled as a green vertical line. TSS – transcription start site (predicted based on the computational analysis, Section 3.2.5). UTR – untranslated region. ATG – start codon. gRNA+N, where N is distance from the gRNA to its target transcription start site.

5.2.2.1.3 CRISPRi: blasticidin kill curve for naïve and primed CRISPRi hPSCs

Before transfecting hPSCs with the pgRNA-CKB plasmids, I had to plan a selection method to be able to detect cells carrying pgRNA-CKB. The study by Mandegar and colleagues used 10 µg/ml blasticidin concentration for the plasmid selection in primed hPSCs, however the required concentration for naïve CRISPRi hPSCs was unknown. Therefore, I optimised the protocol to be used in the naïve cells for the first time. First, to identify the appropriate blasticidin concentration, a

blasticidin kill curve was created for both naïve and primed hPSCs (Figure 5.6). In each case, CRISPRi hPSCs lacking the pgRNA-CKB were treated with blasticidin at five different concentrations for five days, then imaged and counted using Countessa Cell Counter and trypan blue.

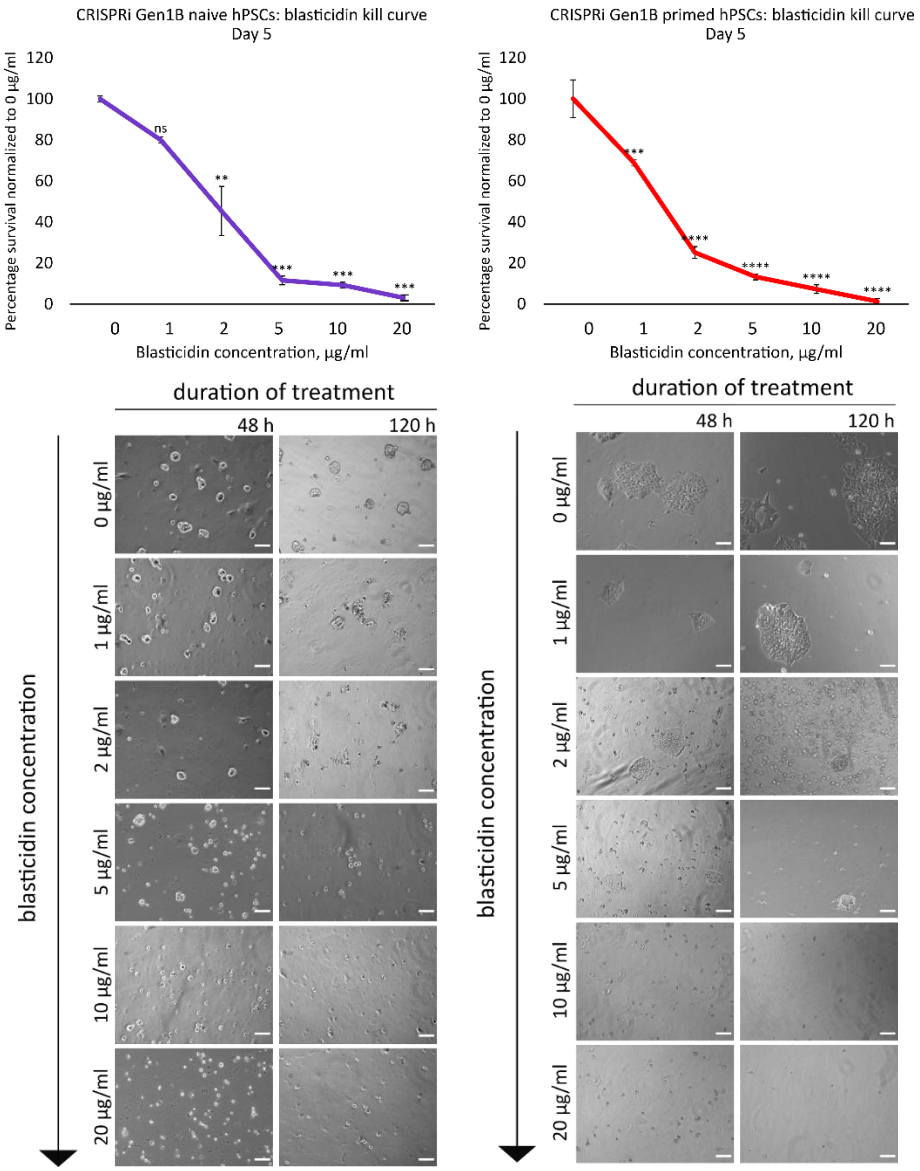


Figure 5.6 Blasticidin kill curve in the naïve and primed hPSCs: linear graph (top) and microscope images (bottom). Mean ±SD is shown, n=2. ANOVA with Dunnett's multiple comparison test was performed (ns – non-significant, $p < 0.005$ (**), $p < 0.0005$ (***), $p < 0.00005$ (****)); '0 µg/ml' was used as a control.
Scale, 100 µm.

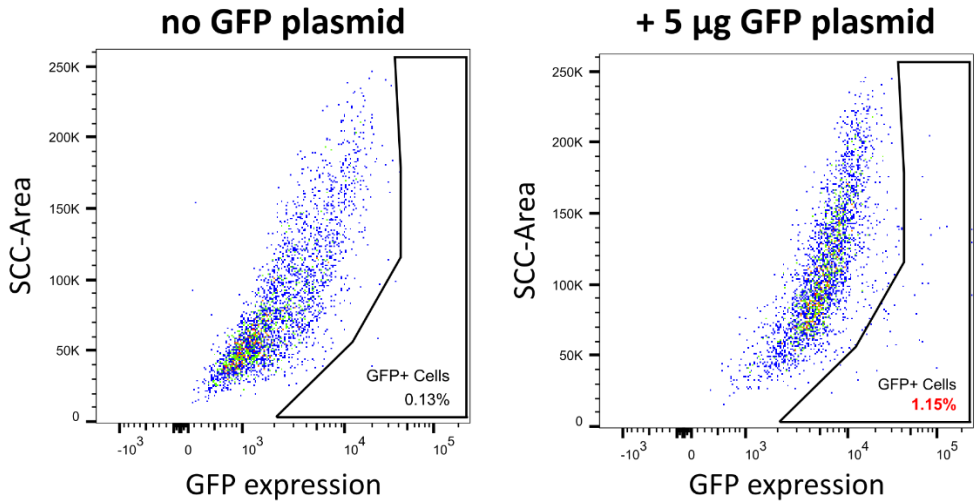
As a result, 5 µg/ml and 10 µg/ml blasticidin concentrations were determined to cause death of >90% of both naïve and primed hPSCs. Therefore, to ensure that the gRNA-transfected naïve cell lines were not overly stressed by the antibiotic selection and also to ensure that the antibiotic resistance system could deal with the treatment, in the experiments described later in this chapter, I

decided to use blasticidin at 7.5 µg/ml concentration for naïve hPSCs in combination with Y-27632. CRISPRi primed hPSCs were treated with 10 µg/ml blasticidin and Y-27632 as described in Mandegar et al., 2016. The resulting selected cell lines were predicted to contain ~10% WT cells and potentially some cells that might silence the plasmid; to eliminate those, I also planned to flow-sort the cells by the expression of mKate2 reporter, encoded by the plasmid.

5.2.2.1.4 CRISPRI: hPSC transfection protocol optimisation

As a next step, I had to decide which WT cell line - naïve or primed - was to be used for the cell transfection. *NANOGP1* was demonstrated to be highly expressed in the naïve state and I therefore aimed primarily to study its function in that context. Therefore, transfecting pgRNA-CKB plasmids into naïve cells would provide a more direct route, but the alternative option was to reprogramme transfected primed cells into the naïve state. Based on a series of transfection tests I performed during the course of this project, naïve hPSCs always exhibited a significantly lower transfection efficiency than the primed lines. An example of such efficiency difference is shown in Figure 5.7.

NAIVE:



PRIMED:

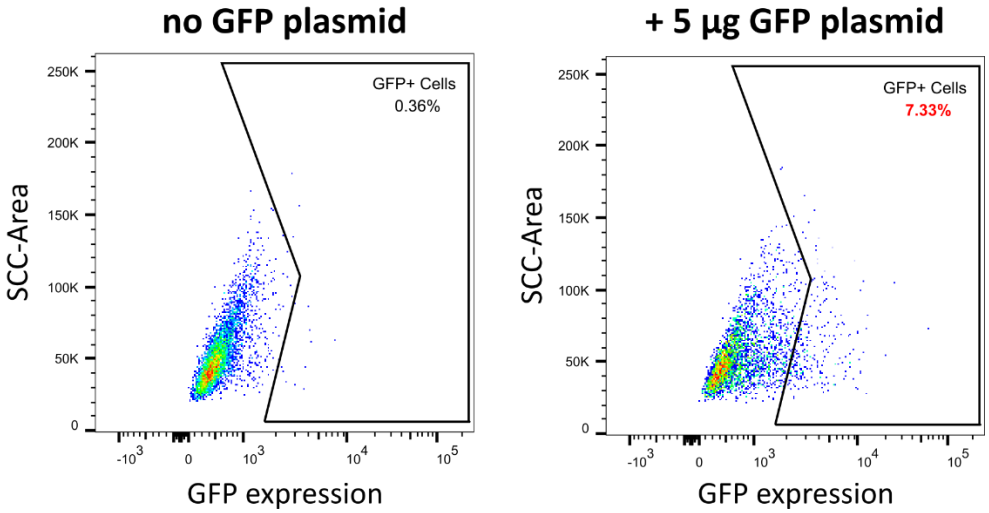


Figure 5.7 Flow cytometry scatterplots showing efficiency of transfecting naïve and primed hPSCs with a constitutive GFP-encoding vector. X-axis shows GFP fluorescence. SSC – side scatter. Percent of GFP-positive cells is indicated in the top/bottom right corner of each scatterplot. This experiment was performed >3 times, one representative example is shown here.

To test whether the same is observed in the naïve and primed CRISPRi hPSC lines, they were transfected with the gRNA-encoding plasmids (Figure 5.8). As before, the efficiency of the naïve CRISPRi hPSC transfection was significantly lower than that of the primed cells, <1% vs. >10%, respectively.

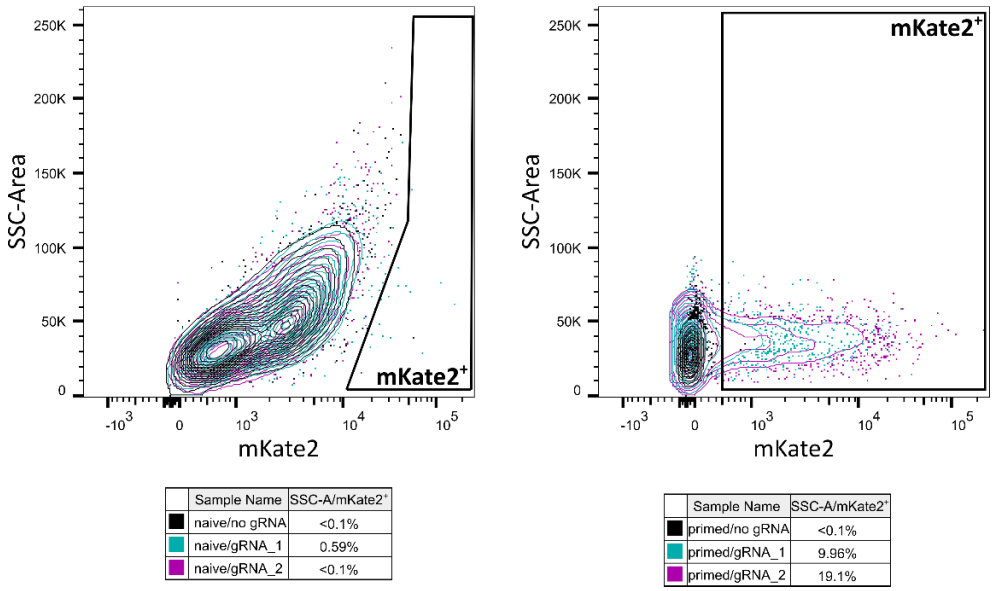


Figure 5.8 Flow cytometry contour plots showing efficiency of transfecting naïve and primed CRISPRi hPSCs with a pgRNA-CKB plasmid. X-axis shows mKate2 fluorescence. SSC – side scatter. Percentage of mKate2-positive cells is indicated in the tables. This experiment was performed 2 times in the naïve hPSCs and 7 times in the primed hPSCs, two representative examples are shown for each cell state. The gates were drawn based on mKate2 fluorescence of naïve WT CRISPRi hPSC sample (left) and primed WT CRISPRi (right). Naïve WT CRISPRi hPSCs were observed to have a notably high level of autofluorescence, the exact reason remained unknown.

5.2.2.1.5 CRISPRi cell line generation in primed hPSCs

Based on the results of the assays described above, I performed the initial transfection of gRNA-encoding plasmids into primed hPSCs. After transfecting the primed CRISPRi hPSCs with individual gRNA-encoding plasmids, blasticidin selection and flow-sorting were performed as described above. When the selected cell lines were expanded, they contained 92-98% mKate2$^+$ cells, which was sufficient for further experiments. Examples of mKate2+ cell population profile before and after cell line selection are shown in Figure 5.9.

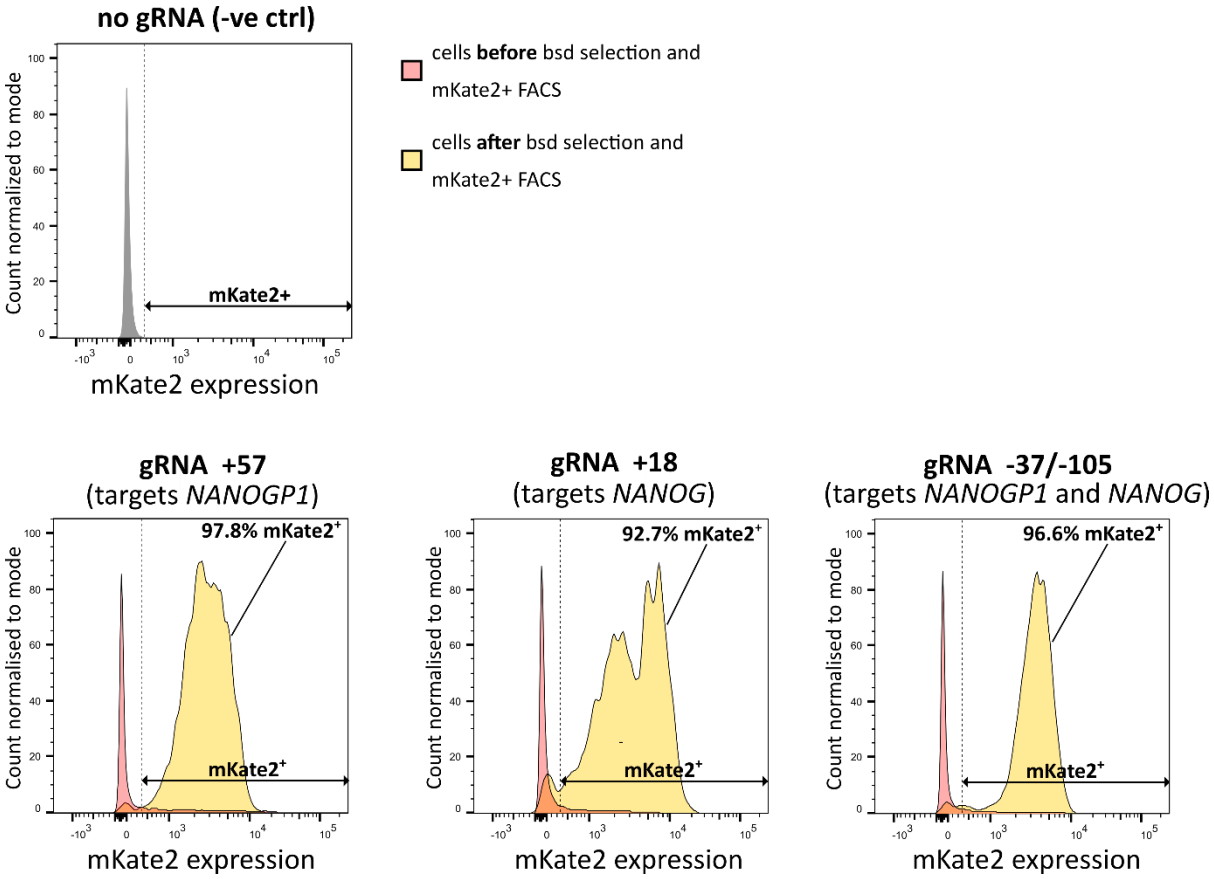


Figure 5.9 Flow cytometry histograms showing mKate2 reporter expression in the selected and non-selected CRISPRi lines. gRNA +N, where N is distance from the gRNA to its target transcription start site. This experiment was performed for each transfected cell line, three plots are shown as an example. Bsd – blasticidin.

As a next step, I aimed to test the responsiveness of the transfected cells to doxycycline. First, I tested which doxycycline concentration is optimal for primed CRISPRi hPSCs. Primed CRISPRi WT hPSCs were treated with 0.25 µM, 0.5 µM and 1 µM doxycycline for 24 h, and mCherry reporter expression was analysed by flow cytometry. All of the tested concentrations induced the reporter expression in 99.9% of cells (Figure 5.10). Doxycycline at 1 µM was used in all of the subsequent assays that involved primed CRISPRi hPSCs.

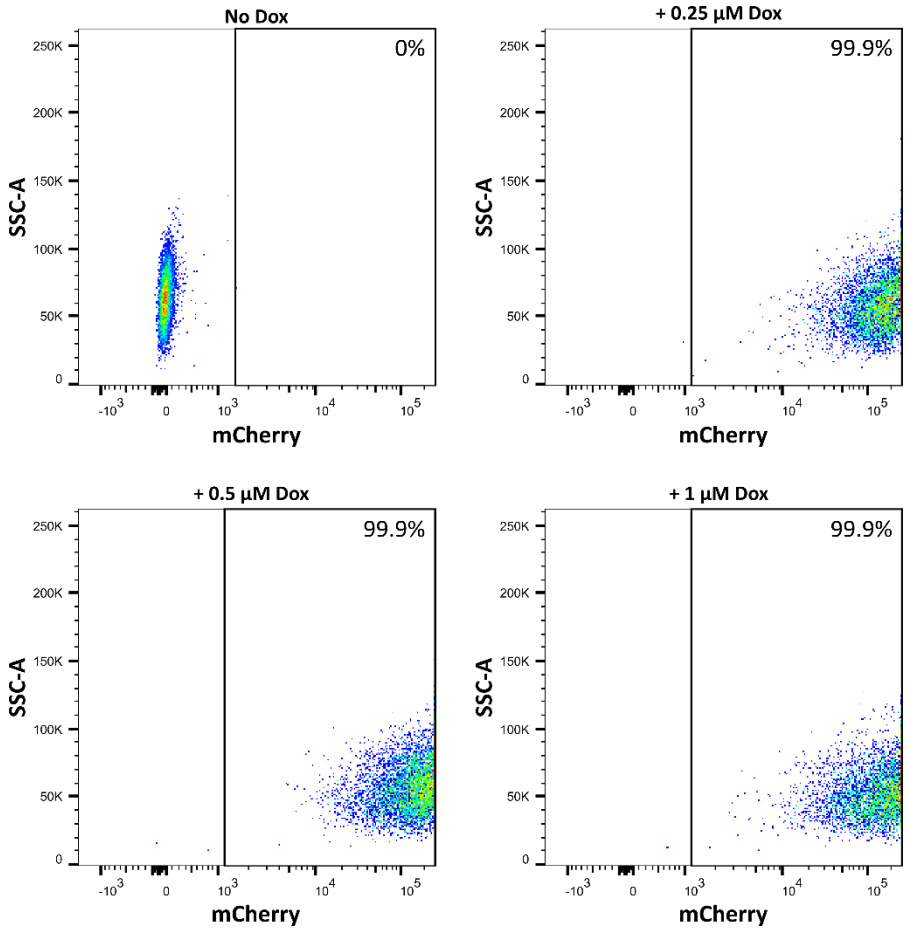


Figure 5.10 Flow cytometry scatterplots showing efficiency of doxycycline (Dox) induction in primed CRISPRi hPSCs. X-axis shows mCherry fluorescence. SSC-A – side scatter area. Percent of mCherry-positive cells is indicated in the top right corner of each scatterplot. This experiment was performed 3 times, one representative example is shown here.

As a next step, selected and expanded CRISPRi cells that expressed the gRNA plasmid were induced with 1 µM doxycycline for 24 h. It was not possible to completely distinguish the mKate2+ from the mKate2+/mCherry+ populations due to the two fluorophores being detected in the same channel on the flow cytometer. Nevertheless, it was clear that the majority of cells in the selected populations responded efficiency to doxycycline by inducing their mCherry expression (Figure 5.10).

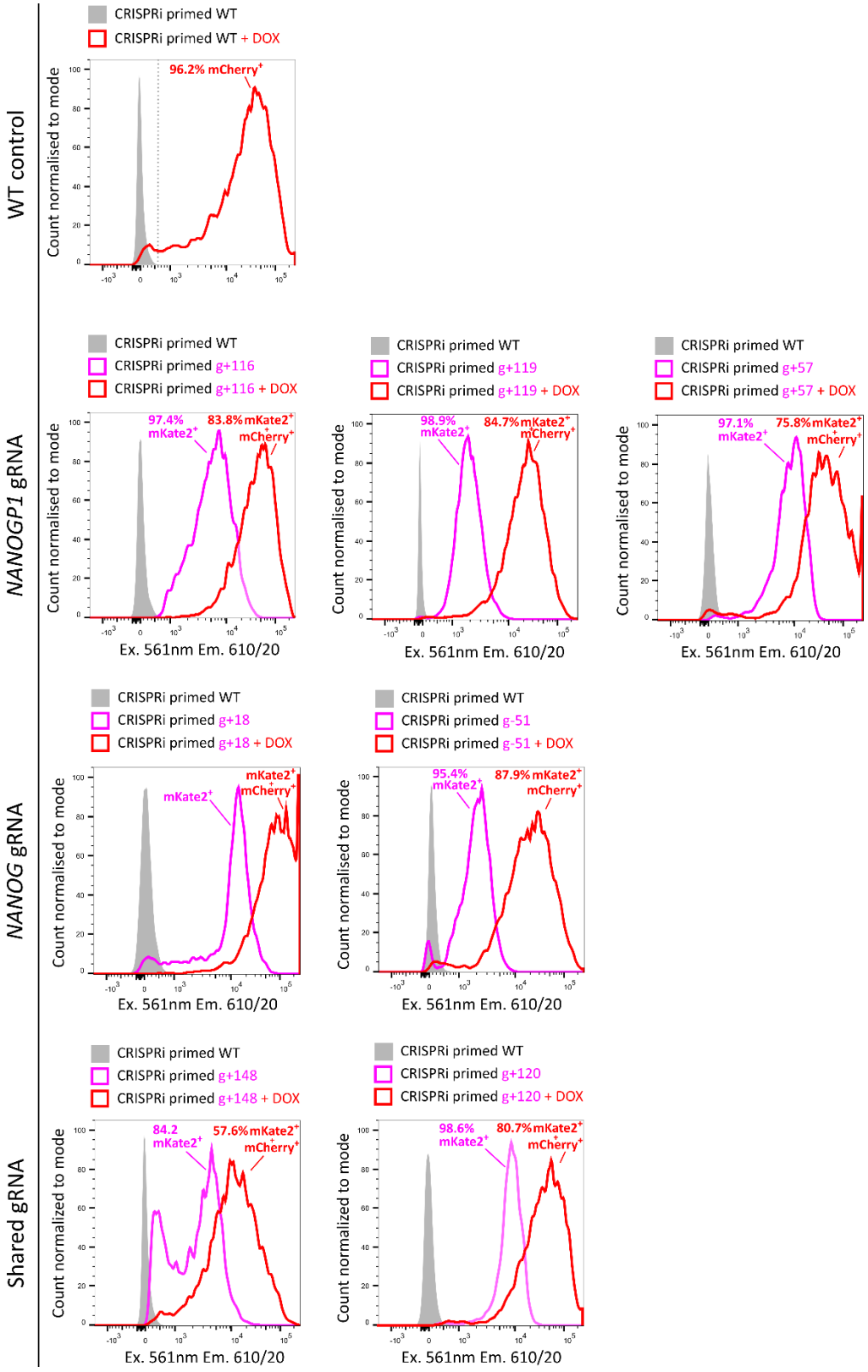


Figure 5.11 Flow cytometry histograms showing mKate2 and mCherry reporter expression in the selected CRISPRi lines. g+N, where g is 'gRNA' and N is distance from the gRNA to its target transcription start site. This experiment was performed once for each transfected cell line. Ex. 561nm – excitation wavelength. Em. 610/20 – emission spectrum. DOX – doxycycline.

One of the *NANOGP1*-specific lines, which contained the gRNA +57, did not expand at the same rate as the other six lines, even after repeated transfection of the plasmid into the WT cell line, suggesting a possible issue with the vector itself. Therefore, that line was disbanded and the remaining six lines were tested further.

As a next step, I aimed to test the efficiency of each individual gRNA. The CRISPRi hPSC lines were treated with doxycycline for four days, after which the expression levels of *NANOG* and *NANOGP1* were assayed by RT-qPCR using the gene-specific primers that I previously optimised.

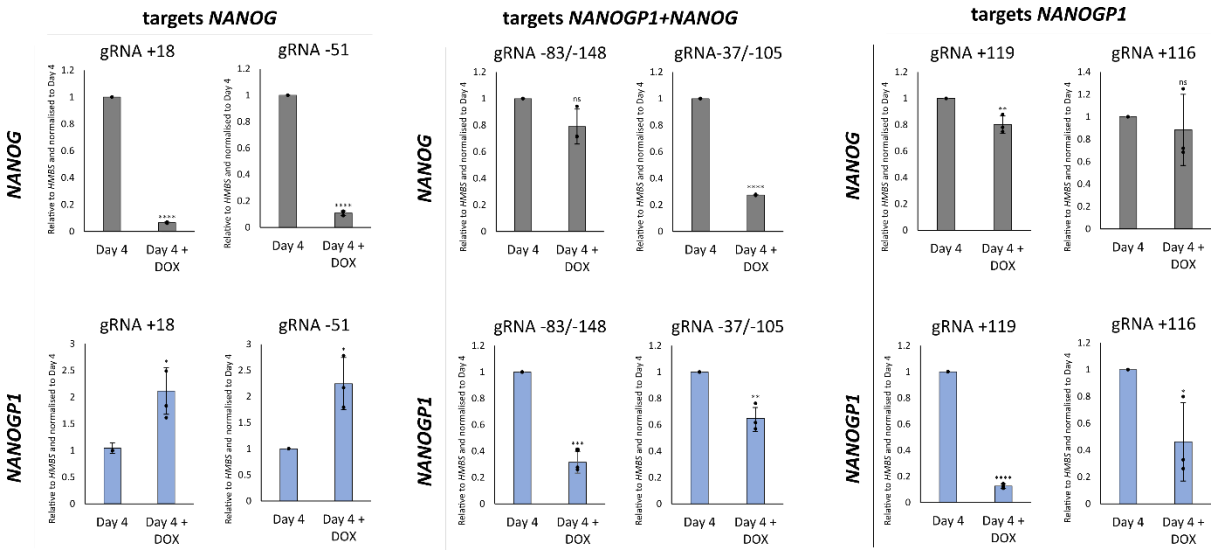


Figure 5.12 Bar charts showing the efficiency of expression downregulation in *NANOG*, *NANOGP1* and *NANOGP1/NANOG* primed CRISPRi hPSCs in mTeSR Plus culture medium. RT-qPCR values are relative to *HMBS* expression. Individual replicates (n=2/n=3) and mean ± SD are shown. T-test for each +/- DOX pair was performed ($p < 0.05$ (*), $p < 0.005$ (**), $p < 0.0005$ (***), $p < 0.00005$ (****)). gRNA +N, where and N is distance from the gRNA to its target transcription start site.

Both of the *NANOG*-targeting lines were very efficient, and the *NANOG* transcription level was downregulated by >90% compared to the non-induced control. Interestingly, in both cases, the levels of *NANOGP1* changed in the opposite direction, and its expression levels were ~2-fold higher than the non-induced controls. CRISPRi gRNA +18 resulted in stronger *NANOG* expression knockdown compared to gRNA -51; the former was therefore chosen to be used in the subsequent assays. The efficiency of the *NANOG* expression downregulation at the protein level was also assessed and confirmed by Western blotting (Figure 5.13). Moreover, downregulating *NANOG* in primed hPSCs led to a change in cell morphology and subsequent cell differentiation (Figure 5.13), in line with previously published data (Vallier et al., 2009).

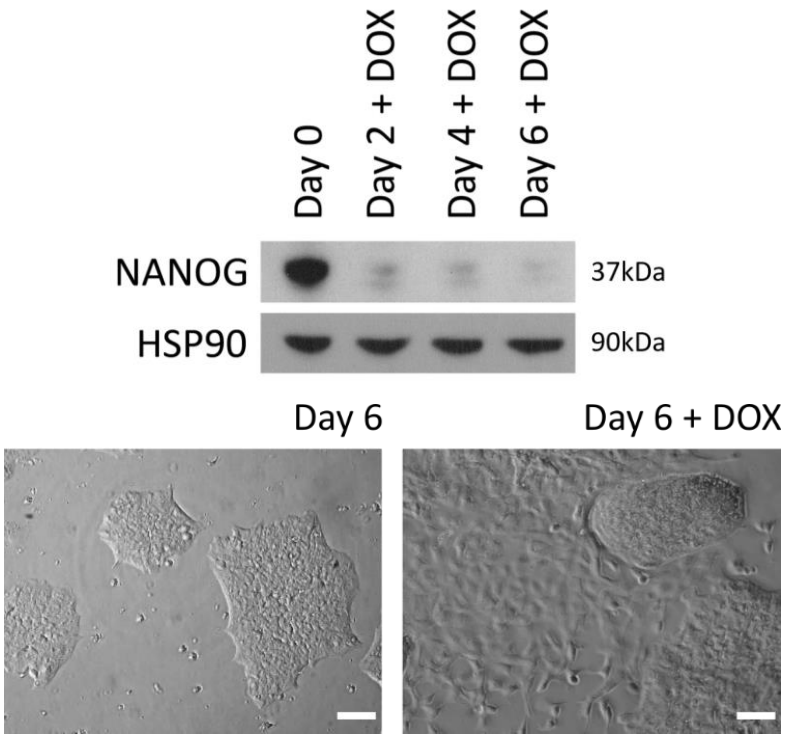


Figure 5.13 Western blotting image (top) and microscope images (bottom) showing the efficiency of the NANOG expression downregulation in primed gRNA+18 CRISPRi hPSCs. NANOG-specific band is at 37kDa. bands. HSP90 – loading control. DOX – doxycycline. Scale, 100 µm.

In the shared-target lines, the target expression was downregulated, however, none of the gRNAs had the same efficiency against both *NANOG* and *NANOGP1* when compared to the *NANOG*-specific CRISPRi hPSCs. It is likely to be challenging to establish a double knock-down line in the future, since *NANOGP1* appears to get upregulated during a good *NANOG* expression knock-down.

The two *NANOGP1*-specific cell lines produced the same result, but with two different efficiencies. The efficiency of gene expression downregulation and consistency between the replicates in CRISPRi gRNA +116 hPSCs was lower than those of CRISPRi gRNA +119. Indeed, the latter exhibited a significant reduction in *NANOGP1* expression level by Day 4, ~90%, and was comparable with the efficiency of the *NANOG*-specific lines. Of note, CRISPRi gRNA +119 exhibited some downregulation of *NANOG* levels as well, ~20% compared to the non-induced control. However, after inducing the cell line for 6 days, no change of cell morphology/differentiation phenotype was observed (Figure 5.14), suggesting that the *NANOG* downregulation was not sufficient to induce differentiation. At that stage, it was not clear whether the reduction was a consequence of *NANOGP1* downregulation or an off-target effect, and whether the same would be observed in naïve hPSCs. However, as the reduction was significantly lower compared to that of *NANOGP1* (~20% vs ~90%), this line was chosen to be tested further in the naïve context and, if successful, to study the consequences of *NANOGP1* expression downregulation.

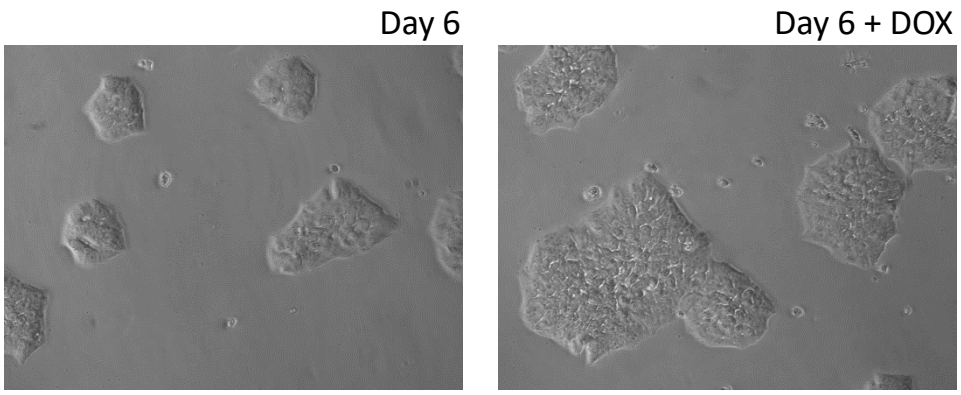


Figure 5.14 Microscope images showing the cell morphology phenotype of the induced and non-induced primed gRNA+119 CRISPRi hPSCs. DOX – doxycycline. Scale, 100 µm.

In summary, two CRISPRi cell lines were taken forward, one targeting *NANOG* and one targeting *NANOGP1*. No double *NANOG*/*NANOGP1* knock-down cell lines were found suitable for further experiments due to the inefficient expression downregulation.

5.2.2.1.6 CRISPRi: primed-to-naïve cell reprogramming and system validation

Both primed cell lines were then reprogrammed into the naïve state. Here, two different protocols were used: chemical resetting (VPA) and 5i/L/A (Guo et al., 2017; Theunissen et al., 2016). After more than 5 attempts to reprogramme the transfected and WT CRISPRi lines, the chemical reprogramming method was found to be inefficient and did not produce viable dome-shaped colonies. We think this is something specific to the CRISPRi line because the protocol works well for us when reprogramming other hPSC lines. In contrast, the 5i/L/A method was significantly more efficient and was used to generate *NANOG*- and *NANOGP1*-specific CRISPRi lines (Figure 5.15). After this, the reprogrammed lines were treated with 7.5 µg/ml blasticidin to eliminate any potential plasmid silencing that occurred during the reprogramming.

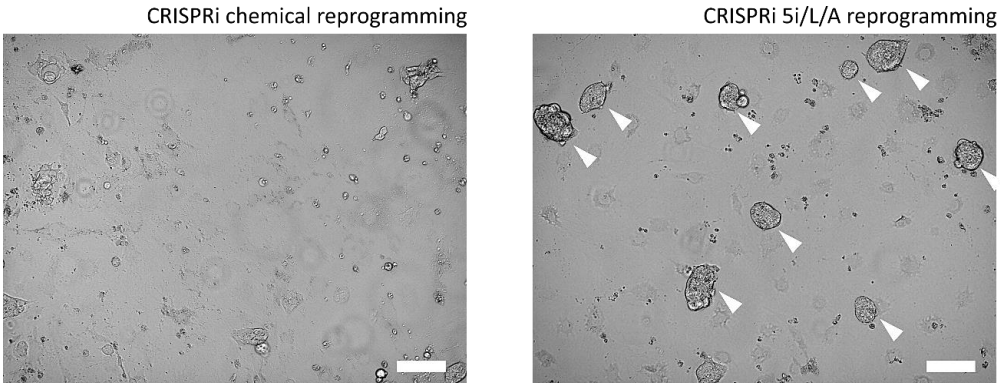


Figure 5.15 Microscope images illustrating the efficiency of the two primed-to-naïve hPSC reprogramming methods in CRISPRi hPSCs. Pluripotent colonies are indicated by white arrowheads. Scale, 100 µm.

To verify that the naive cell lines did not undergo *AAVS1* locus silencing and remain responsive to doxycycline treatment, the cells were induced by 1 nM, 10 nM, 100 nM, 500 nM and 1 µM Doxycycline for 24 h. In this experiment, it was important to ensure that as many cells as possible are capable to express dCas9-KRAB, this way ensuring a stable and consistent gene expression knock-down. Interestingly, even though the latter three concentrations were significantly more efficient than 1 nM and 10 nM, none were able to induce mCherry reporter expression in 99.9% of cells, as was observed earlier when the same assay performed in primed hPSCs (Figure 5.16, Figure 5.10).

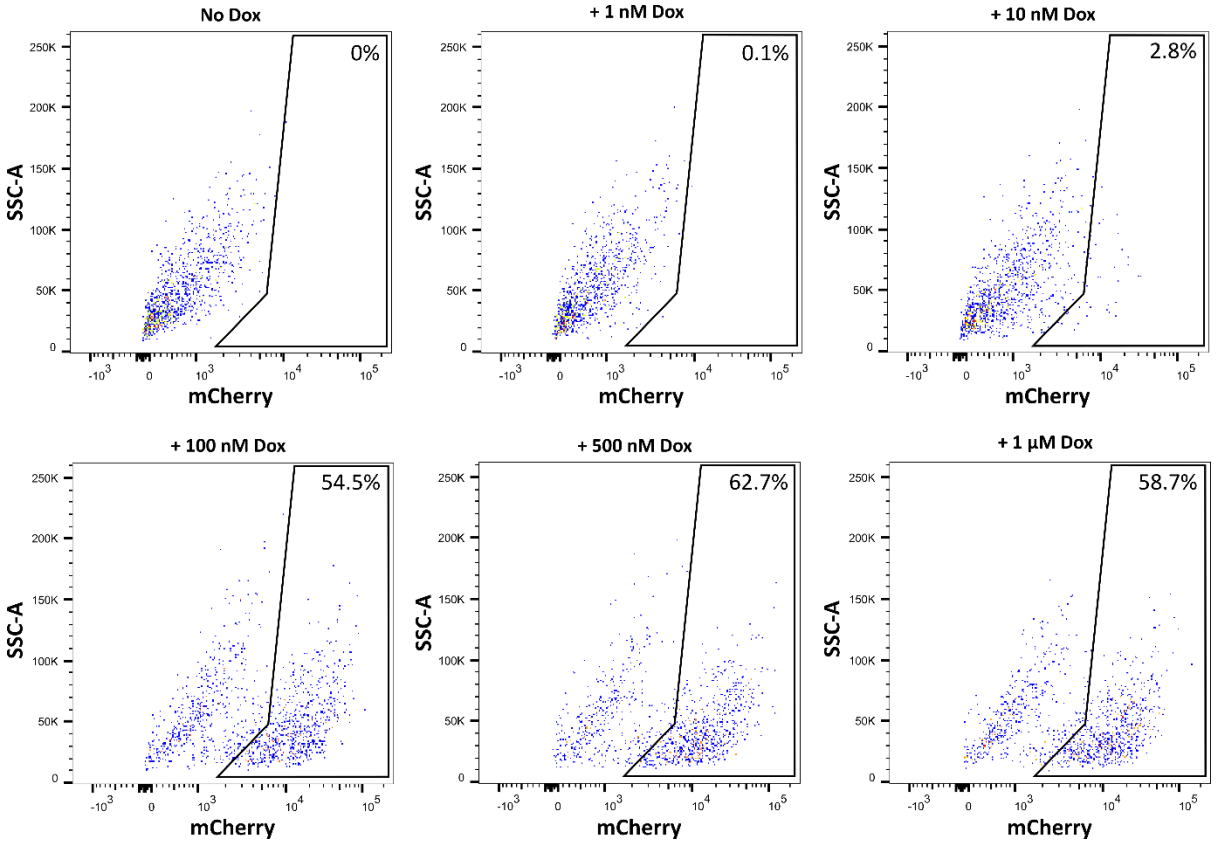


Figure 5.16 Flow cytometry scatterplots showing efficiency of doxycycline (Dox) induction in naive CRISPRi hPSCs. X-axis shows mCherry fluorescence. SSC-A – side scatter area. Percent of mCherry-positive cells is indicated in the top right corner of each graph. This experiment was performed 1 time for all concentrations (shown here) and >3 times for the 1 µM Dox.

In order to overcome this, mCherry positive cells were separated from the mCherry negative population by flow sorting after brief treatment with doxycycline, expanded and analysed by flow cytometry again (this was performed twice). However, this did not fully eliminate the problem, and a subpopulation of naïve CRISPRi hPSCs re-appeared that did not respond to the induction. Comparison of naïve and primed CRISPRi hPSC reporter induction is shown in Figure 5.17.

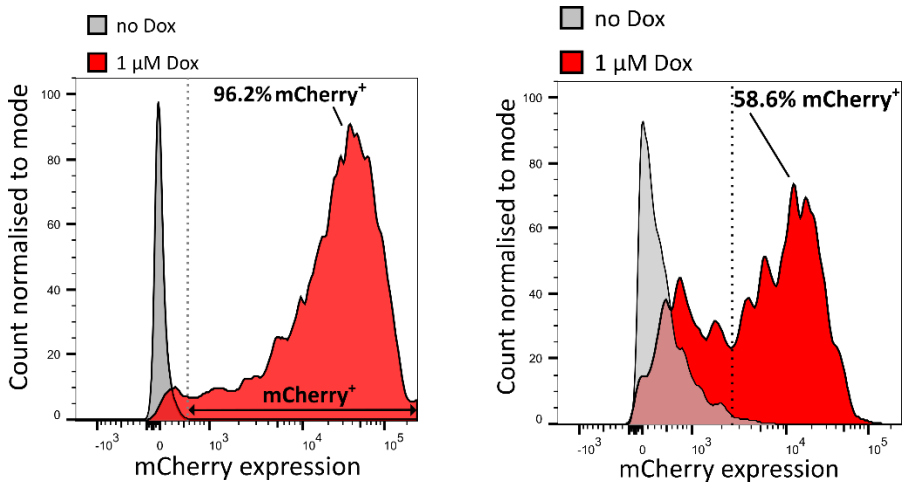


Figure 5.17 Flow cytometry histograms showing the difference between the mCherry reporter expression in naïve (right) and primed (left) isogenic CRISPRi hPSCs. Dox – doxycycline.

Another approach to eliminate this non-responsiveness was to treat the naïve cells with G418, in case the *AAVS1* locus did undergo silencing. The naive CRISPRi hPSCs were treated with the G418 at 3 different concentrations: 150 ng/µl, 200 ng/µl and 300 ng/µl in two different experiments. In one experiment, the cells were grown on feeders and were treated with the antibiotic for 12 days, followed by a two-day doxycycline induction. The second experiment had a longer timeline (20 days of G418 treatment and four days of doxycycline induction), as well as feeder-free naïve cell culture conditions. Using the prolonged antibiotic treatment and a feeder-free culture allowed me to obtain a cell line where >75% expressed mCherry after the Doxycycline treatment, which was consistent with the amount of mCherry+ colonies observed under a fluorescent microscope. (Figure 5.18, Figure 5.19).

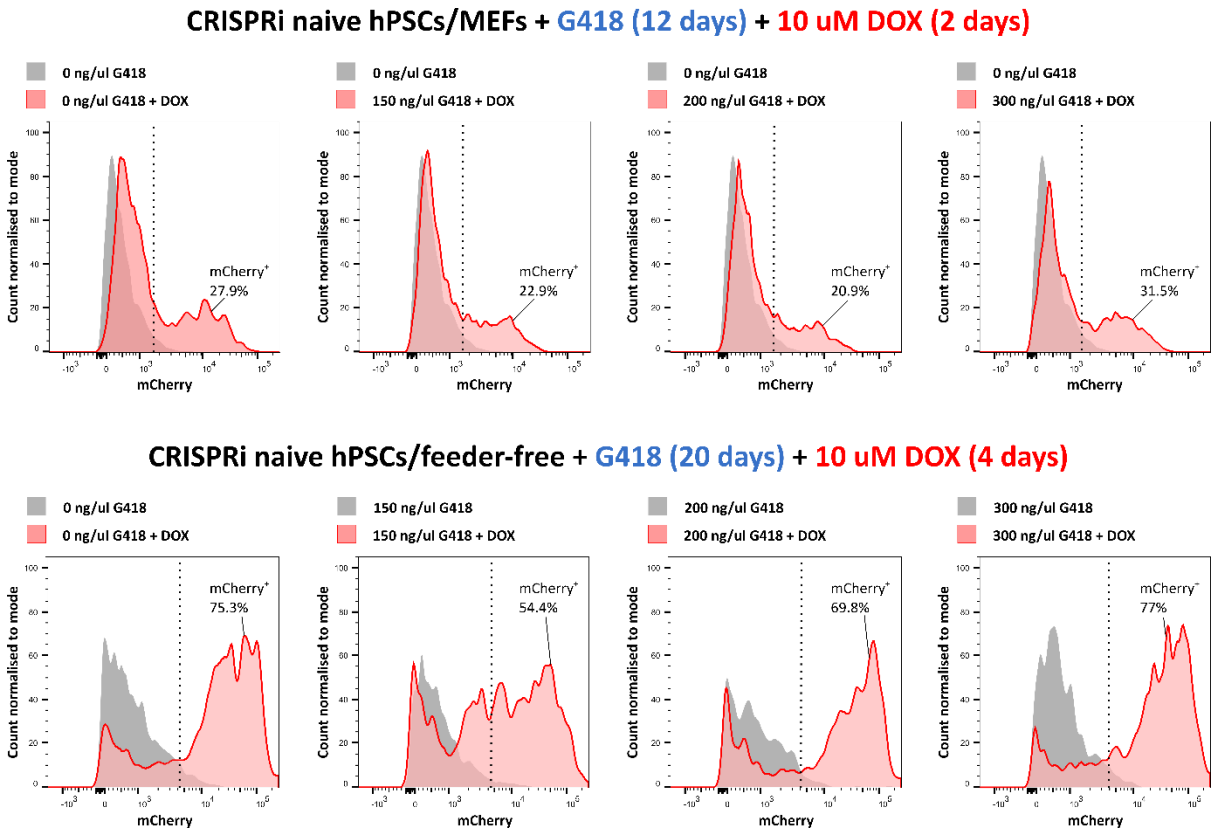


Figure 5.18 Flow cytometry histograms showing the G418 treatment of naïve CRISPRi hPSC in naïve culture medium t2iLGo. DOX – doxycycline. CRISPRi WT (no gRNA) hPSCs were used for optimising the approach.

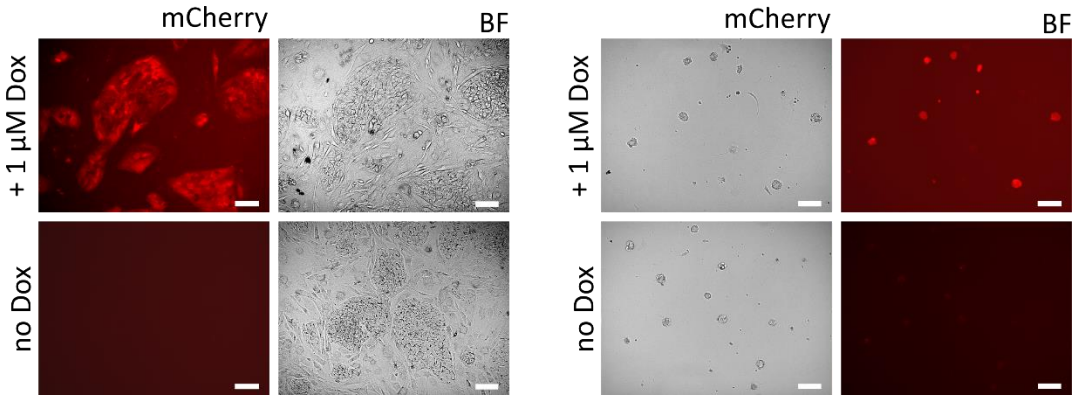


Figure 5.19 Fluorescence and bright filed (BF) microscope images illustrating mCherry reporter expression in the primed and naïve CRISPRi hPSCs. Scale, 100 µm.

After establishing cell lines with the sufficient number of doxycycline-responsive cells, I aimed to test the efficiency of gene expression knock-down in the reprogrammed lines, i.e., to check whether the plasmids did not get silenced during the reprogramming. To test this, *NANOG*- and *NANOGP1*-specific lines were induced with doxycycline for four days in t2iLGo naïve medium. This caused efficient and specific gene expression knockdown, whereby *NANOG* transcripts were reduced by ~80% and *NANOGP1* levels by ~90% (Figure 5.20). Moreover, the levels of NANOG protein were also strongly

reduced after two days of doxycycline treatment (Figure 5.21). Importantly, no significant *NANOG* expression downregulation was observed in the *NANOGP1* CRISPRi (Figure 5.12). Similarly, *NANOG* CRISPRi did not significantly alter *NANOGP1* RNA levels.

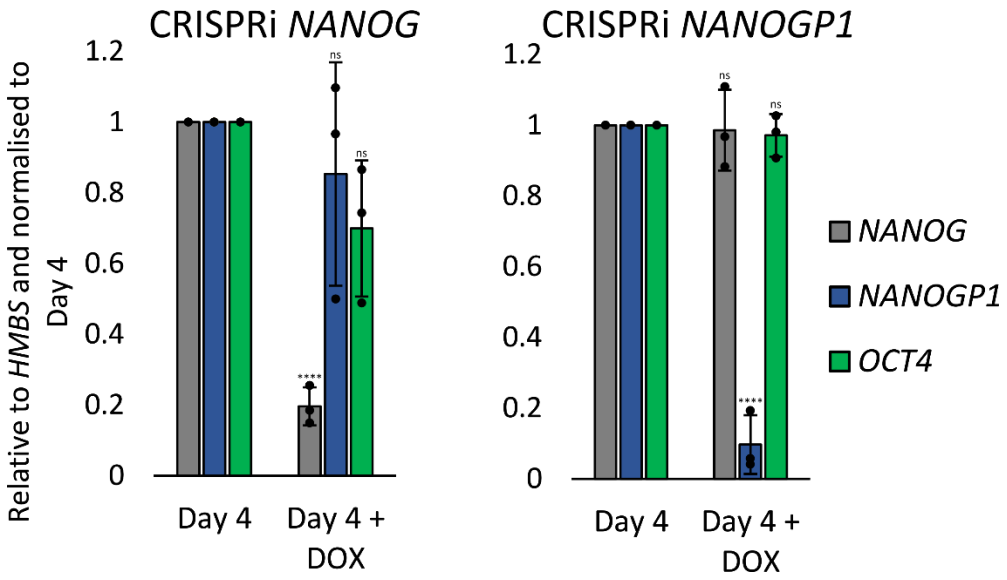


Figure 5.20 Bar charts showing the efficiency of expression downregulation in *NANOG* and *NANOGP1* naïve CRISPRi hPSCs in naïve culture medium t2iLGo. RT-qPCR values are relative to *HMBS* expression and normalised to Day 4 sample. Individual replicates (n=3) and mean ± SD are shown. T-test for each +/- DOX pair was performed (ns – nonsignificant, p < 0.00005 (****)). DOX – doxycycline.

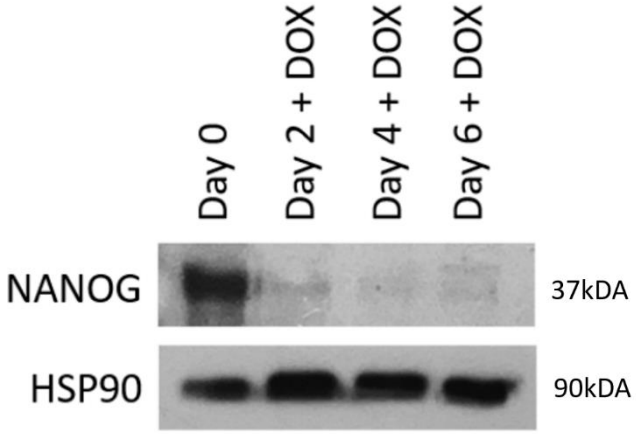


Figure 5.21 Western blotting image showing the efficiency of NANOG protein expression downregulation in *NANOG* naïve CRISPRi hPSCs in naïve culture medium t2iLGo. NANOG-specific band is at 37kDa. DOX – doxycycline. HSP90 – loading control.

5.2.2.1.7 CRISPRi *NANOG* and *NANOGP1* expression downregulation analysed by RNA-seq

After validating the CRISPRi cell lines, I set up a nine-day time course experiment, during which the *NANOG*- and *NANOGP1*-specific CRIPSRi cell lines were induced and harvested at defined

timepoints throughout the assay. By day 9 of doxycycline treatment, *NANOG* CRISPRi cell cultures were completely differentiated, with abundant flat, trophectoderm-like colonies, while, in contrast, *NANOGP1* CRISPRi colony morphology remained undifferentiated and domed-shaped (Figure 5.22).

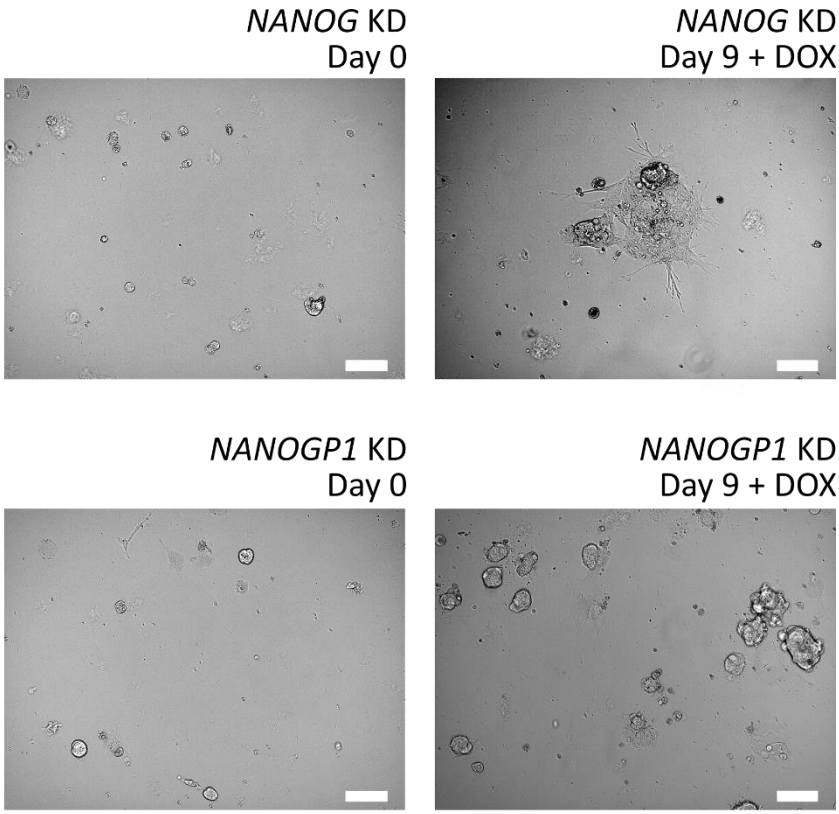


Figure 5.22 Microscope images showing cell morphology of *NANOG* and *NANOGP1* naïve CRISPRi hPSCs on Day 0 and Day 9 + DOX in naïve culture medium t2iLGo. DOX – doxycycline**.** KD – expression knockdown. Scale, 100 µm.

Defined time points were harvested in each experiment (Day 0, Day 2, Day 4, Day 6 and Day 9 in CRISPRi *NANOG* hPSCs, and Day 0, Day 2, Day 4, Day 9 CRISPRi *NANOGP1* hPSCs; n=3). The samples were analysed by RNA sequencing, obtaining 14-35 million reads per sample. Correlation scores were all >0.95 for each pairwise replicate of the same sample.

Transcriptional changes reflecting cell differentiation were clearly observed in the *NANOG* expression downregulation experiment, as shown in the PCA plot below (Figure 5.23). At the same time, all *NANOGP1* samples, both induced and non-induced, clustered in approximately the same area (Figure 5.23).

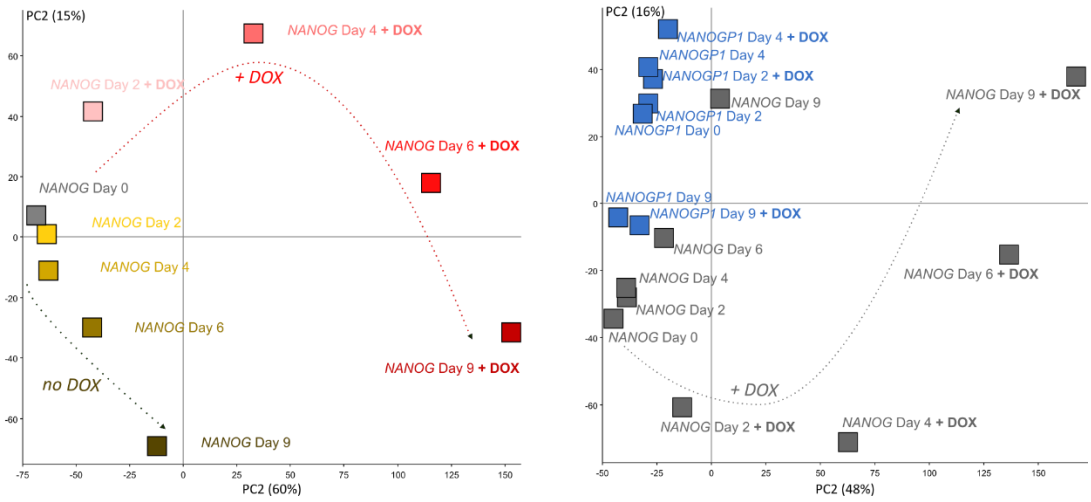


Figure 5.23 PCA plots showing transcriptional differences between *NANOG* and *NANOGP1* CRISPRi hPSCs analysed by RNA-seq. PC1 (x-axis) vs. PC2 (y-axis) are shown. Each data point is an average of n=3.

Over the nine-day time course, *NANOG* downregulation caused the naïve cells to differentiate, as evidenced by the reduction in the expression of naïve and core pluripotency factors including *KLF17*, *SOX2* and *POU5F1* (Figure 5.24). The changes in gene expression were apparent after two days of CRISPRi induction. Genes associated with the trophectoderm lineage, such as *GATA2, GATA3, CDX2, ESRRB* and *TACSTD2*, were upregulated in the knockdown samples starting from day 2 and continued to increase in their expression up to the day 9 endpoint. In contrast to CRISPRi *NANOG* hPSCs, induced CRISPRi *NANOGP1* hPSCs did not exhibit the same transcriptional changes, maintaining similar profiles throughout the time course. Interestingly, the uninduced CRISPRi *NANOG* hPSCs samples showed a slight increase in *GATA2*, *GATA3*, and *CDX2* expression during later stages of the time course. This had likely occurred due a prolonged cell culture during the time course, which was suboptimal in terms of the cell culture method, but necessary to assess the effect of *NANOG* downregulation within one passage.

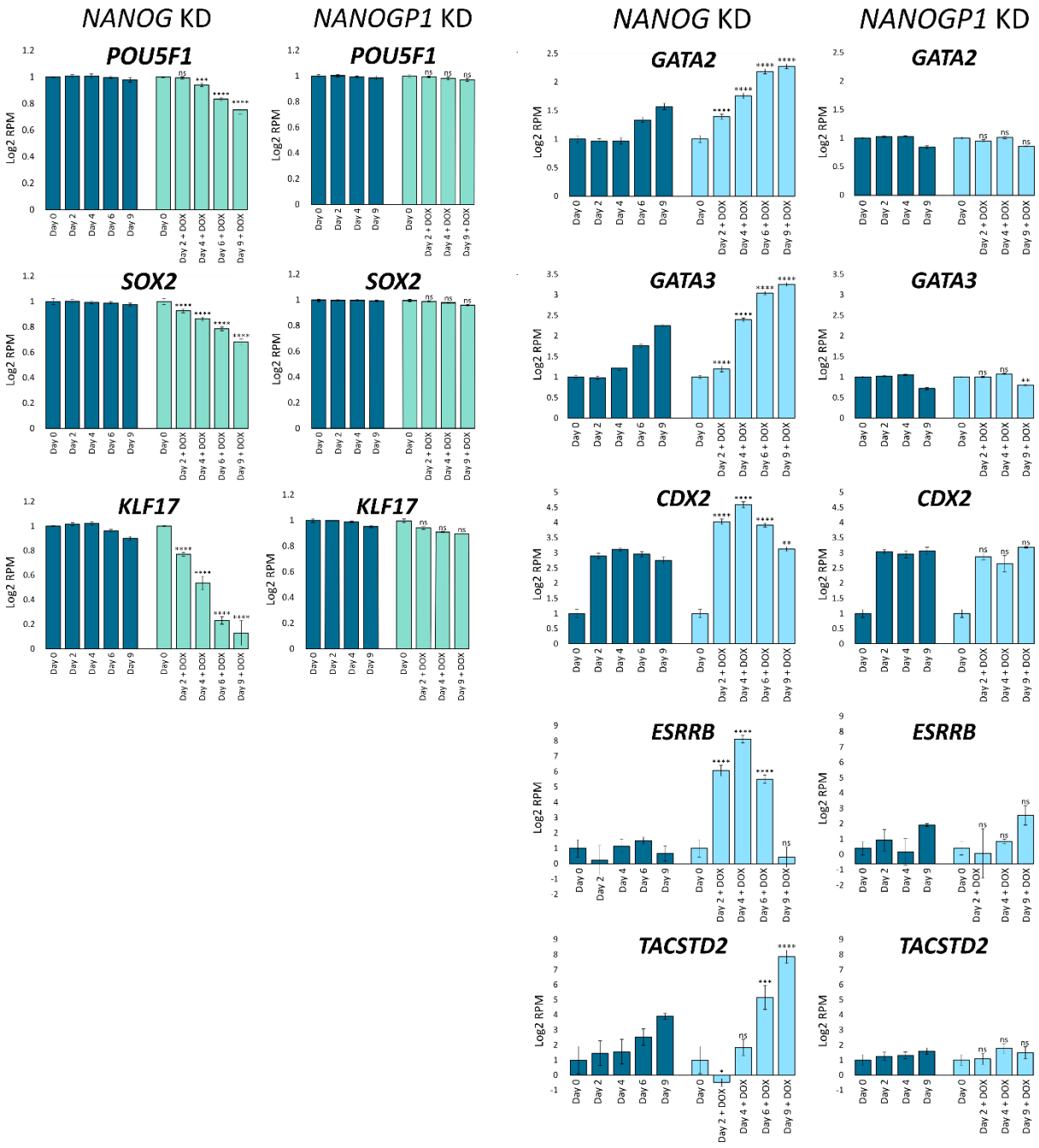


Figure 5.24 Bar charts showing RNA-seq expression values for pluripotency genes and trophectoderm lineage markers in NANOG and NANOGP1 naïve CRISPRi hPSCs. Gene expression is in Log2 RPM (reads per million) and is normalised to Day 0. Mean ±SD (n=3) is shown. One-way ANOVA with Tukey's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), $p < 0.0005$ (***), $p < 0.00005$ (****)). Within each graph, Day 2 was compared to Day 2 + DOX, Day 4 – to Day 4 + DOX, and so on. DOX – doxycycline. KD – expression knockdown.

Nevertheless, both *NANOG* KD and NANOGP1 KD had a substantial number of differentially expressed genes: a total of 8474 genes when comparing *NANOG* control versus knockdown samples at Day 9, and 1599 genes when comparing *NANOGP1* control versus knockdown samples at Day 9 (Figure 5.25).

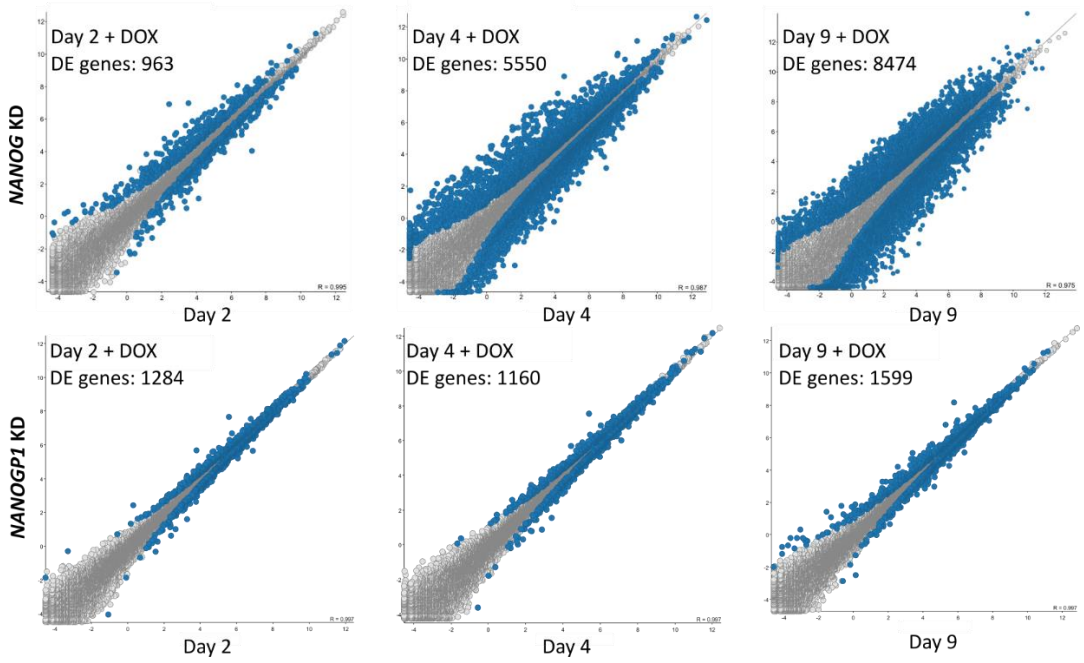


Figure 5.25 Scatterplots showing differentially expressed (DE) genes in NANOG and NANOGP1 naïve CRISPRi hPSCs. DE genes were filtered by statistical test (DeSeq2 in SeqMonk), $p<0.05$. DE genes are in blue, other genes are in grey. DE genes were identified in the RNA-seq analysis. DOX – doxycycline. KD – expression knockdown.

To further characterise the differentiation phenotype, the RNA-seq dataset was compared to single cell transcriptomes from cultured human embryos (Xiang et al., 2020). The analysis was performed with the help of Maria Rostovskaya (Babraham Institute). The results show that cells undergoing *NANOG* downregulation, starting from Day 4 of the induction, cluster very closely to the developing trophectoderm and cytotrophoblast, In contrast, the non-induced samples and earlier time points (Day 0 and Day 2 + DOX) cluster closer to the pre- and early post-implantation epiblast (Figure 5.26). This validates the identity of the differentiated cells that are induced following *NANOG* expression knockdown. In addition, the induced CRISPRi *NANOGP1* samples are similar to the non-induced samples, and cluster with the epiblast, demonstrating that the reduction in *NANOGP1* expression is not sufficient to disrupt the transcriptome of naïve pluripotent cells or cause trophectoderm differentiation. The results of the assay are shown in the PCA plot below (Figure 5.26). Interestingly, *NANOGP1* KD (+/- DOX) and *NANOG* KD (- DOX) did not plot in the same location when analysed separately. This is likely to some internal transcriptional changes between the two cell lines that were not characterised in the thesis. This would require future investigation. However, when these samples are plotted together they completely overlap, which led to a conclusion that these samples do not undergo differentiation towards the trophectoderm lineage.

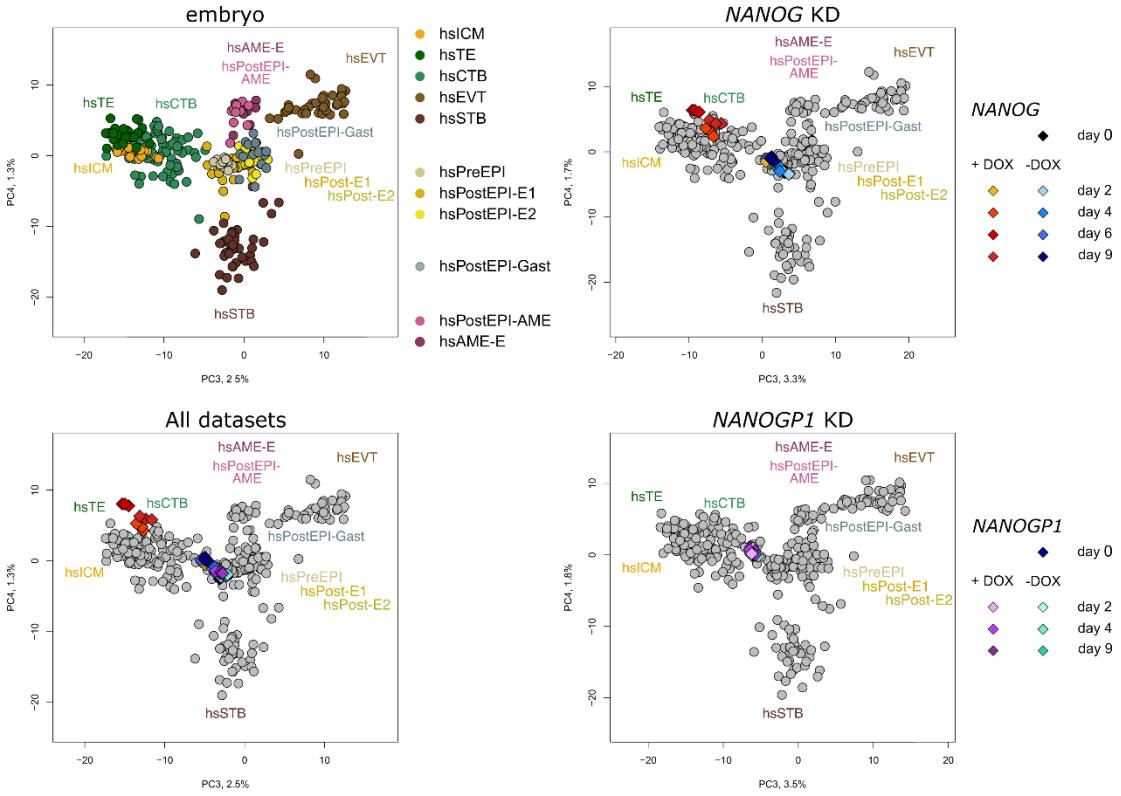


Figure 5.26 PCA plots, comparing *NANOG* and *NANOGP1* naïve CRISPRi hPSCs to human embryo transcriptome data. DOX – doxycycline. KD – expression knockdown. All datasets – Xiang et al., *NANOG* KD, *NANOGP1* KD. ICM – inner cell mass. TE – trophectoderm. CTB – cellular trophoblast. EVT – extravillous trophoblast. STB – syncytiotrophoblast. PreEPI – preimplantation epiblast. PostEPI – post-implantation epiblast. PostEPI-Gast – gastrulating stage. PostEPI-AME – post-implantation amniotic sac. AME – amniotic sac. Hs – human dataset. PC3 (x-axis) and PC4 (y-axis) are shown.

Lastly, I investigated whether *NANOGP1* might regulate the expression of the predicted NANOGP1 target genes that were identified in the earlier ChIP-seq experiments (Section 4.2.2). I found that the vast majority of NANOGP1-only gene targets were not differentially expressed following *NANOGP1* knockdown. Only three genes out of 72 were significantly altered: *PNPLA6* was upregulated, and *GABRD* and *DLGAP3* were downregulated (Figure 5.27). Thus, in these cell culture conditions, *NANOGP1* is not required to maintain appropriate expression levels of the identified target genes.

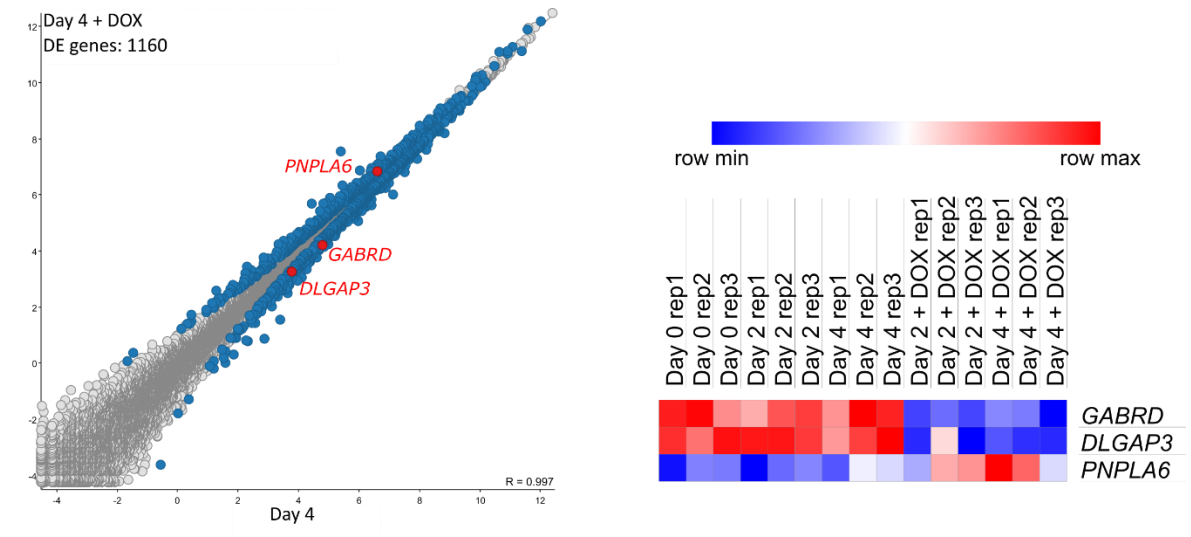


Figure 5.27 Scatterplot showing differentially expressed (DE) genes in naive *NANOGP1* CRISPRi hPSCs (left). Expression of three, differentially expressed, predicted NANOGP1 targets in the *NANOGP1* CRISPRi RNA-seq dataset is shown in the heatmap (right). DE genes were filtered by statistical test (DeSeq2 in SeqMonk), p<0.05. DE genes, identified in the RNA-seq analysis, are in blue, other genes are in grey. DOX – doxycycline. Predicted NANOGP1-only targets found among the DE genes are in red. Higher gene expression is in red, lower is in blue. Data is row-normalised. Rep1, rep2, rep3 – individual RNA-seq replicates.

Interestingly, out of 366 predicted shared targets, 43 were classified as differentially expressed genes in the *NANOGP1* KD dataset, and 91 genes in the *NANOG* KD dataset Figure 5.28. Seventeen of the differentially expressed predicted target genes were shared between the *NANOG* and *NANOGP1* expression knockdown datasets (Figure 5.28). No common gene expression pattern was found among these shared predicted targets: some targets were upregulated both in *NANOG* and *NANOGP1* KD, some were downregulated, and some exhibited opposite expression patterns.

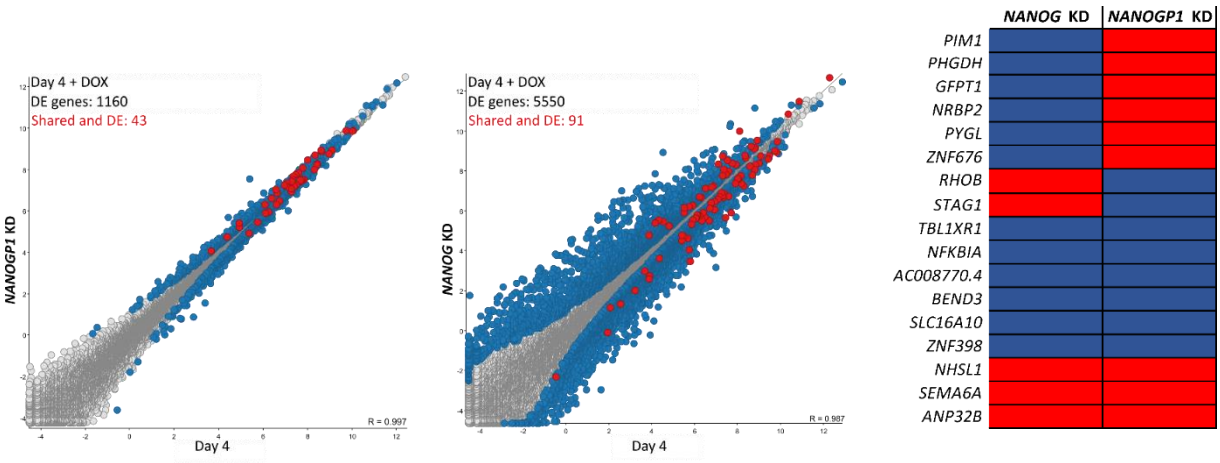


Figure 5.28 Scatterplots showing differentially expressed (DE) genes in *NANOGP1* and *NANOG* naïve CRISPRi hPSCs (left). Heatmap showing gene expression patterns of the DE NANOG/NANOGP1 targets found in both datasets (right). Scatterplots: All DE genes identified in the RNA-seq analysis are

in blue, other genes are in grey. DE genes were filtered by statistical test (DeSeq2 in SeqMonk), p<0.05. DE NANOG/NANOGP1 target genes are in red. DOX – doxycycline. KD – expression knockdown. Heatmap: blue – downregulated in + DOX, red – upregulated in + DOX (generated in Excel; only shows whether the gene was upregulated or downregulated)

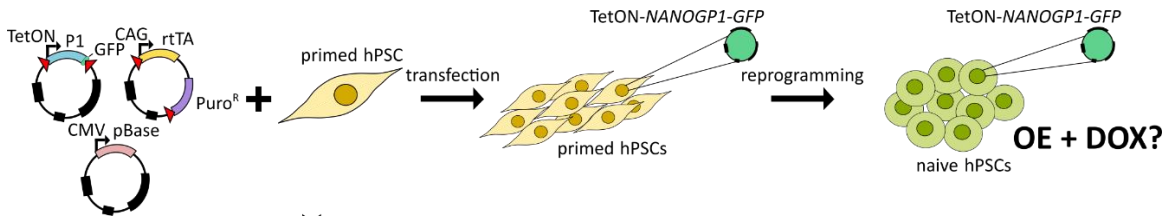
In summary, in this section I developed and validated functional *NANOG* and *NANOGP1* CRISPRi naïve hPSCs. For the first time, this thesis captures the transcriptional dynamics that are induced following the silencing of *NANOG* expression in naïve hPSCs. The findings also demonstrate that *NANOG*-depleted naive cells exit the pluripotent state and differentiate towards trophectoderm, and that the resulting cells have transcriptomes that cluster with developing trophectoderm cells in the embryo. Unlike for *NANOG*, the downregulation of *NANOGP1* expression in naïve hPSCs did not lead to cell differentiation, thereby, unlike *NANOG*, *NANOGP1* is not required for maintenance of naïve pluripotency. Over the time course following *NANOGP1* knockdown, the number differentially expressed genes increased, including several of the predicted unique and shared NANOGP1 target genes, although the expression changes were typically fairly small.

Overall, in this section I demonstrated that while downregulating expressions of *NANOG* in naïve hPSCs causes loss of pluripotency, this function is not conserved for *NANOGP1*.

5.2.2.2 Development, validation and application of *NANOGP1* overexpression methods in naïve hPSCs

Nanog has an autorepressive function in mouse (Navarro et al., 2012). In this section, by overexpressing *NANOG* and *NANOGP1* in the naïve medium t2iLGo I aimed to test whether they have the same autorepressive properties in human. Additionally, by overexpressing *NANOGP1* during formative capacitation (Rostovskaya et al., 2019), I aimed to check whether its downregulation in primed hPSCs is required and whether the overexpression would interfere with the cell fate transition. Finally, I tested whether *NANOGP1* could promote primed-to-naïve reprogramming when it is overexpressed together with *KLF2* in t2iL, similar to *NANOG* (Takashima et al., 2014).

Summary of the experiments performed in this section can be found in Figure 5.29.



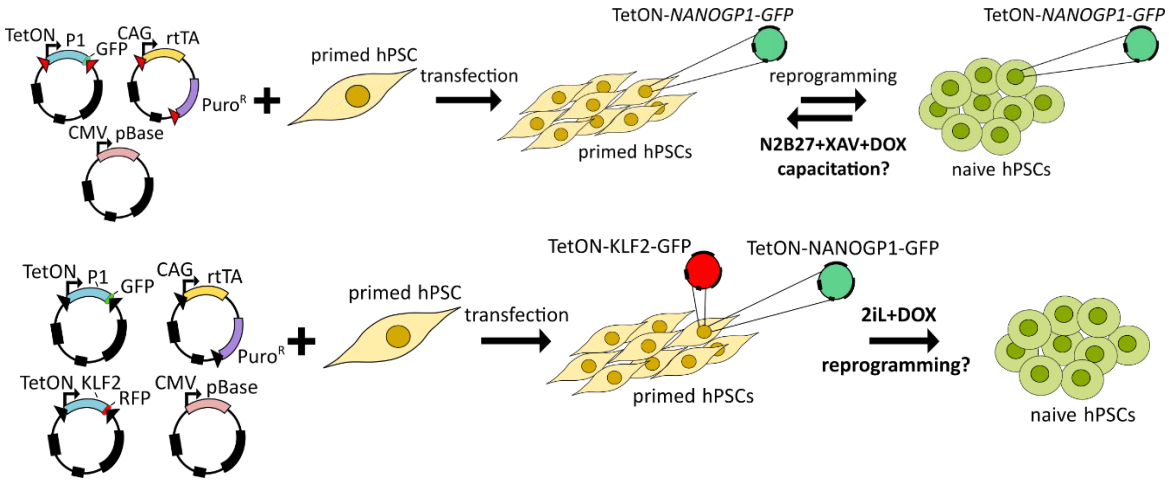


Figure 5.29 Diagram showing the summary of *NANOGP1* overexpression assays in the naive and primed hPSCs. See details below.

5.2.2.2.1 TetON-rtTa induction system: cell line generation

Gene overexpression constructs used in this section were doxycycline-inducible, and used a TetOn-rtTa induction system (Figure 5.30), similar to the one utilised by the CRISPRi lines described in the previous sections.

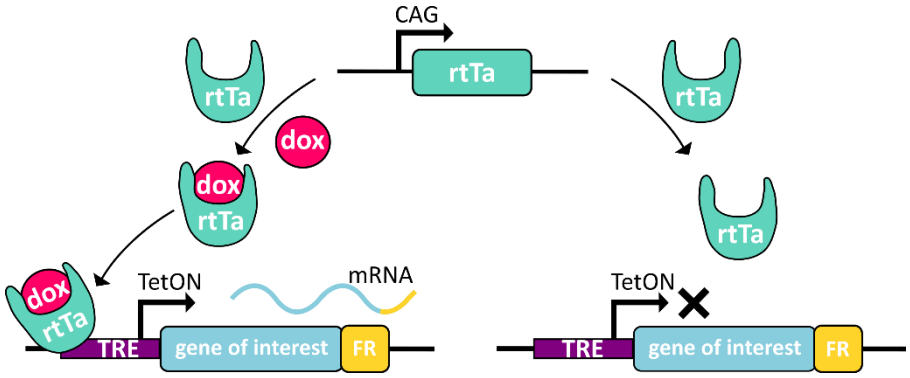


Figure 5.30 Diagram illustrating the mechanism of TetON induced gene overexpression rtTa – transactivator. CAG, TRE/TetON – promoters. dox – doxycycline. FR – fluorescent reporter.

To create a *NANOGP1* overexpressing cell line, primed H9 hPSCs were transfected with three plasmids: i) TetOn-*NANOGP1*-GFP, ii) pCAG-rtTA-puro^R and iii) a transposase-encoding pCMV-pBase. Here, I created overexpression lines for all three *NANOGP1* isoforms, labelled as 1-1, 1-2 and 1-3. After transfection and puromycin-resistance selection, the *NANOGP1-1*, *NANOGP1-2* and *NANOGP1*-3 inducible overexpression cell lines were treated briefly with 1 µM doxycycline and flow sorted for GFP expression. The resultant cells were expanded in the absence doxycycline, then treated with 1 µM doxycycline for 24 h and analysed by flow cytometry to examine transgene induction. More than 96% of the cells in the population were GFP+, demonstrating robust and stable transgene overexpression

(Figure 5.31). After analysing these cell lines by Western Blotting, I concluded that all cell lines were able to generate a stable protein, with all three isoforms (Figure 5.32).

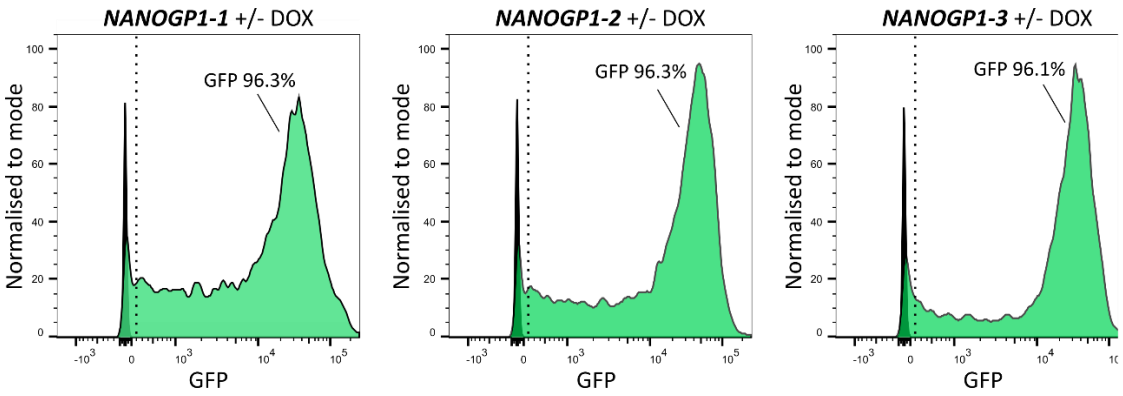


Figure 5.31 Flow cytometry histograms showing GFP reporter expression in the selected and sorted TetON-NANOGP1-GFP primed hPSCs. This experiment was performed twice for each line, one representative example is shown. Percent of GFP-positive cells is indicated. 'Non-induced' peak is in black, 'induced' peak is in green.

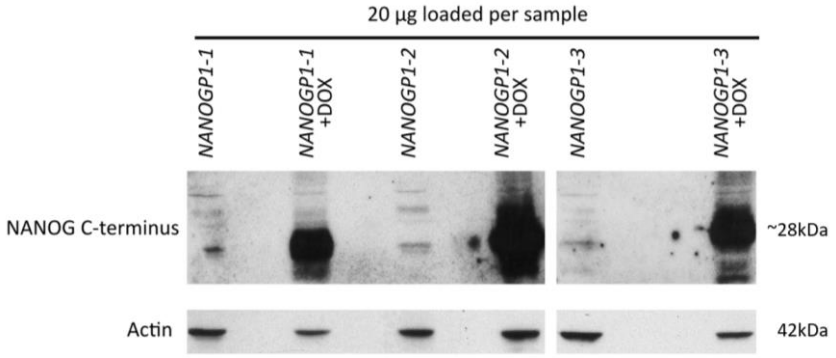


Figure 5.32 Western blotting image showing the efficiency of NANOGP1 protein overexpression in TetON primed hPSCs in primed culture medium mTeSR. NANOGP1-specific bands are at 28kDa. DOX – doxycycline. Actin – loading control.

Similar to CRISPRi, the overexpression system was to be used in the context of naïve hPSCs and therefore I needed to reprogramme the primed cells. As a control, a *NANOG*-overexpressing cell line was also generated and reprogrammed, using the same method as described above.

To simplify the experiment, only one *NANOGP1* overexpressing line was reprogrammed (isoform 1). The other cell lines were frozen and stored for future experiments (Section 5.2.2.5). This cell line, as well as TetOn-*NANOG*-GFP were successfully reprogrammed using chemical reprogramming (Guo et al., 2017)(Figure 5.33).

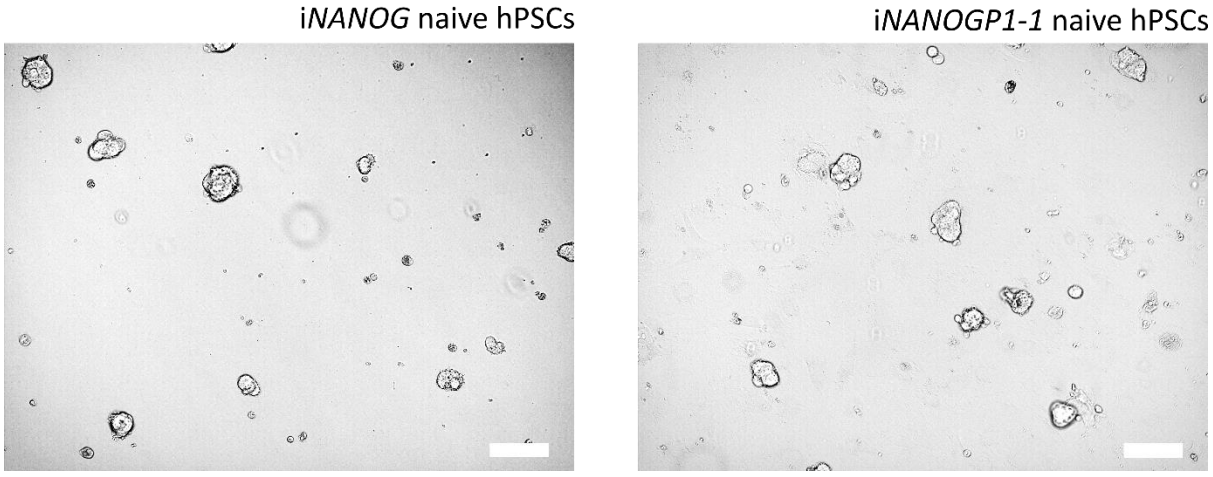


Figure 5.33 Microscope images showing naïve inducible overexpression NANOG and NANOGP1-1 hPSCs in t2iLGo. Scale, 100 µm.

To establish suitable and uniform protein overexpression levels, both lines were treated with 1 µM doxycycline for 48 h and flow sorted into three populations each, based on the level of GFP expression as a proxy for transgene levels: GFP low, GFP medium and GFP high (Figure 5.34). GFP low and GFP high cell populations were frozen and stored, while GFP medium cells were used in further experiments.

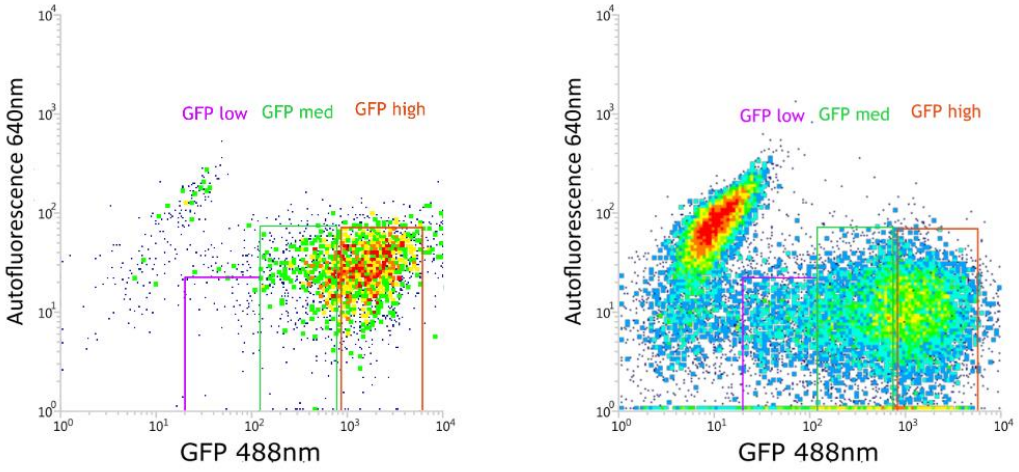


Figure 5.34 Flow cytometry scatterplots showing the cell sorting experiment of TetON-*NANOG*-GFP (left) and TetON-*NANOGP1-1*-GFP (right) naïve hPSC lines.

After expanding the GFP-medium *NANOG* overexpression (OE) and *NANOGP1* OE cells lines in the absence of doxycycline, the cells were induced for 72 h and the protein expression levels were assayed by Western Blotting (Figure 5.35). The overexpressed protein level was deemed suitable, therefore, these cells were used in the further experiments

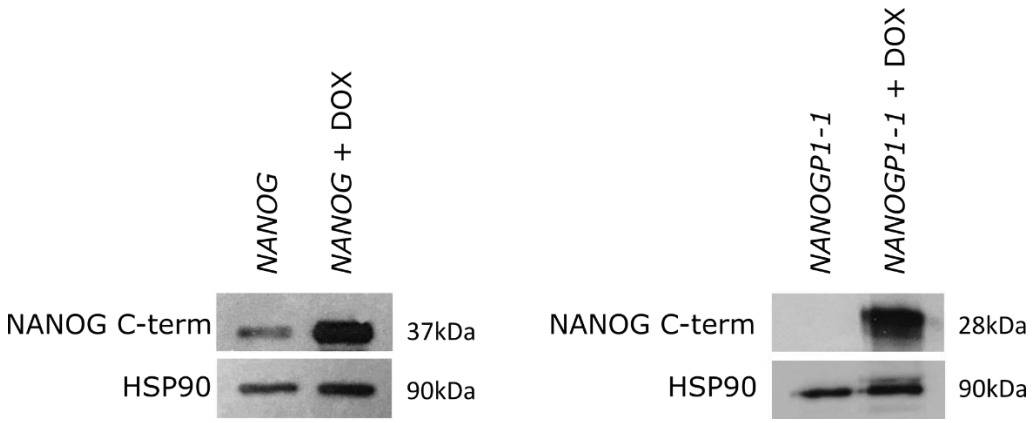


Figure 5.35 Western blotting images showing the efficiency of NANOG and NANOGP1 protein overexpression by the naïve GFP^{med} TetON hPSCs in t2iLGo medium. NANOG-specific band is at 37kDa. NANOGP1-specific band is at 28kDa. DOX – doxycycline. HSP90 – loading control.

5.2.2.3 Does NANOGP1 have an autorepressive and/or dominant negative function in the naïve hPSCs?

To test whether *NANOG* and *NANOGP1* have a conserved autorepressive function in human naïve pluripotency, doxycycline-inducible TetON-*NANOG*-GFP and TetON-NANOGP1-GFP hPSCs were used (see above). An additional aim of this experiment was to test whether *NANOGP1* has any dominant negative control over *NANOG*. I hypothesised that the dominant negative effect would be plausible if overexpressing *NANOGP1* in the naïve culture conditions leads to a similar effect as downregulating *NANOG* in the naive culture conditions (see CRISPRi in Section 5.2.2).

Naïve chemically-reset H9 hPSCs, overexpressing *NANOG* or *NANOGP1* (two separate cell lines), were induced for 18 h and 72 h in t2iLGo naïve media condition. The efficiency of overexpression was confirmed by highly elevated GFP levels in the induced samples at both time points.

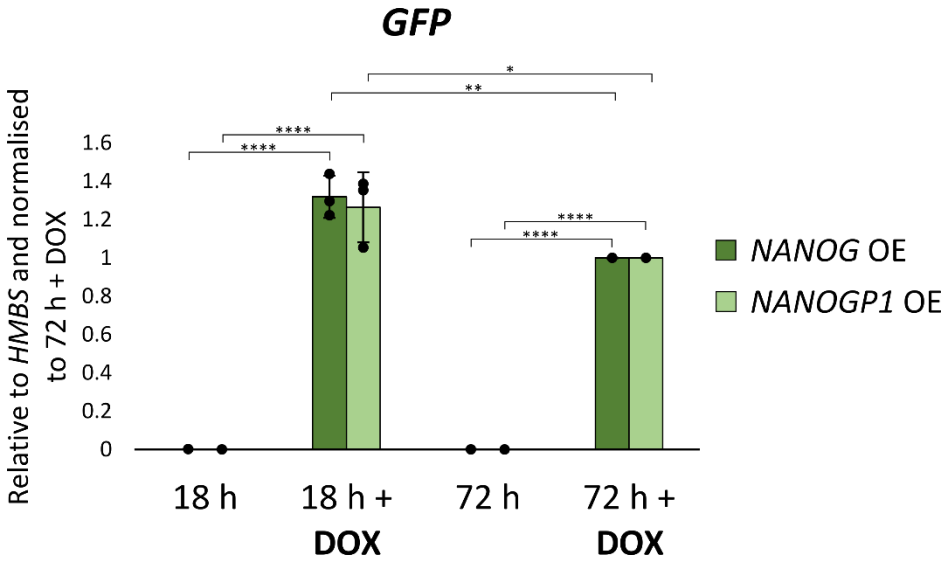


Figure 5.36 Bar charts showing *GFP* expression in *NANOG* and *NANOGP1* naïve TetOn hPSCs in naïve culture medium t2iLGo. RT-qPCR values are relative to *HMBS* expression and normalised to 72 h + DOX sample. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Tukey's multiple comparisons test was performed ($p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

Overexpressing *NANOG* and *NANOGP1-1* led to the downregulation of both endogenous *NANOG* and *NANOGP1* genes already by 18 h, and this effect was maintained at the 72 h time point (Figure 5.37). These results establish that *NANOGP1* has an autorepressive effect on its own expression, as well as on *NANOG*. It is interesting to highlight that *NANOG* also represses *NANOGP1*, showing that the repressive effect if mutual.

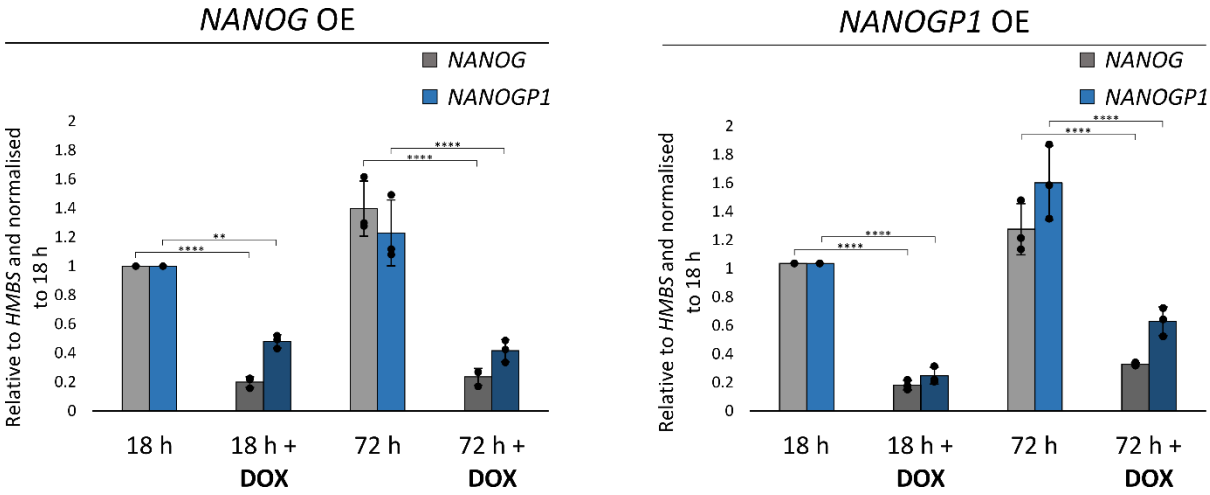


Figure 5.37 Bar charts showing *NANOG* and *NANOGP1* endogenous expression in *NANOG* and *NANOGP1* naïve TetOn hPSCs in naïve culture medium t2iLGo. RT-qPCR primers were designed to bind endogenous RNA only. RT-qPCR values are relative to *HMBS* expression and normalised to 18 h sample. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Tukey's multiple comparisons test was performed ($p < 0.005$ (**), $p < 0.00005$ (****)). DOX – doxycycline.

Other transcriptional changes were observed among pluripotency genes: the naïve factor *KLF4* and the core pluripotency factor *OCT4* were downregulated following the overexpression of *NANOGP1* and *NANOG*. In contrast, the expression levels of two other naive pluripotency factors, *DPPA3* and *KLF17*, instead increased in the induced samples. The cause of these transcriptional changes would require further investigation.

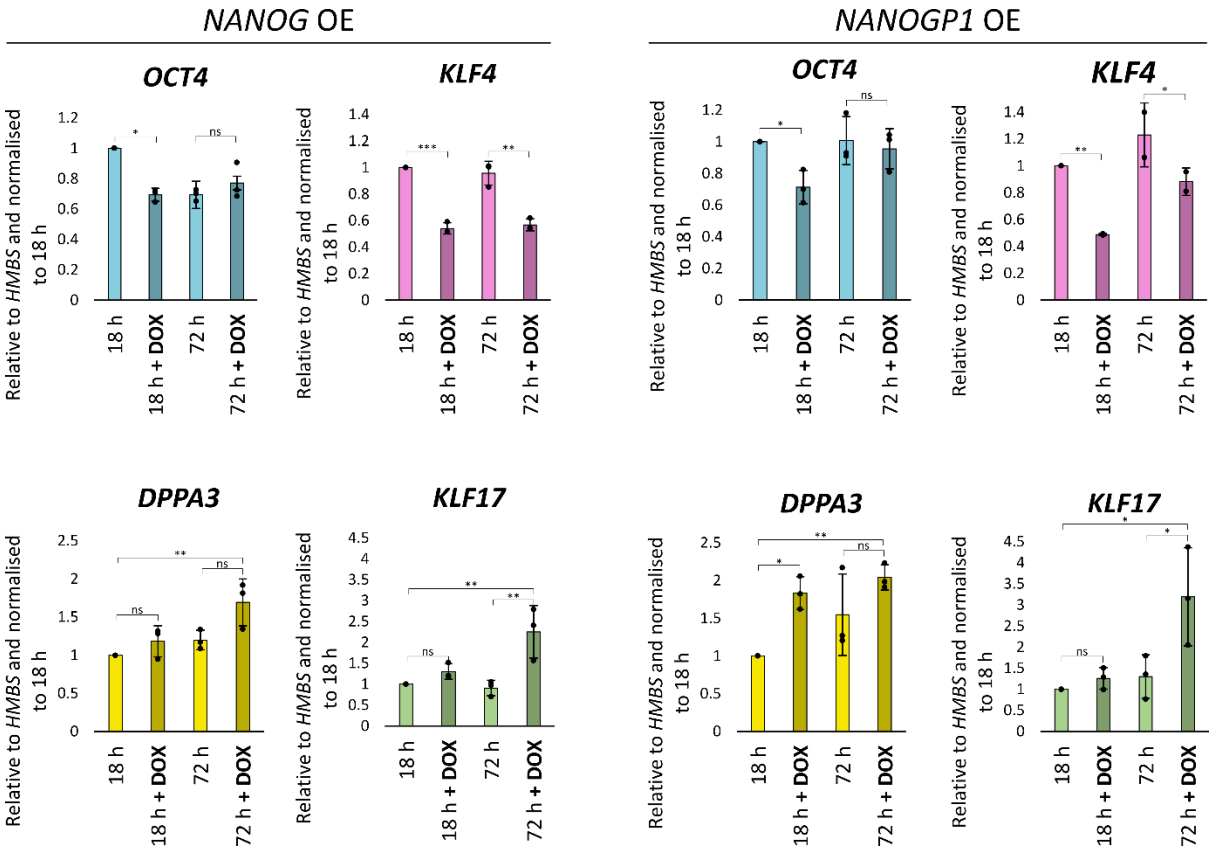


Figure 5.38 Bar charts showing pluripotency gene expression in *NANOG* and *NANOGP1* naïve TetOn hPSCs in naïve culture medium t2iLGo. RT-qPCR values are relative to *HMBS* expression and normalised to 18 h sample. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Tukey's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

Plausible dominant negative effect of *NANOGP1* was addressed in this section, but some aspects remained not fully understood. If *NANOGP1* had a complete dominant negative control over *NANOG*, I would have expected the *NANOGP1* OE cells to undergo the same changes as those that happen during the knockdown of *NANOG*. My earlier experiments using CRISPRi *NANOG* KD hPSCs showed that one of the hallmarks of loss of *NANOG* function in naïve hPSCs is the downregulation of *KLF17*, which occurs within 2-4 days following *NANOG* knockdown (Figure 5.24). In *NANOGP1* OE, *KLF17* is instead, upregulated after 72 h of overexpression, which is the opposite to what would have been expected. Also, during *NANOG* downregulation, the expression of *OCT4* does not exhibit any significant change, whereas in *NANOGP1* OE the expression is noticeably lower compared to the non-

induced control after 18 h of overexpression. Therefore, since these results do not match what was predicted to happen, I concluded that the dominant negative effect is likely not present. However, this experiment had certain limitations that do not allow a definite conclusion. The OE of *NANOGP1* was shorter than the *NANOG* KD time course (3 d vs 9 d), and it is possible that other changes would have been visible if OE was more prolonged. Additionally, while on Day 4 *NANOG* KD did not have a significant effect on the expression level of *NANOGP1*, in *NANOG* OE and *NANOGP1* OE the two duplicates show a very strong suppression of each other's and their own transcription. Therefore, it might be challenging to explore one specific function during such a prominent *NANOG*/*NANOGP1* downregulation phenotype. Based on the data available, I conclude that the dominant negative effect over the *NANOG* function likely does not exist, however, further experiments would be required to confirm this.

In summary, this section demonstrated that both *NANOG* and *NANOGP1* have an autorepressive function in naive hPSCs, previously shown in mouse ESCs. This is the first evidence proving that *NANOGP1* has a conserved *NANOG*-like property in naive hPSCs, supporting my hypothesis of their overlapping roles. Moreover, the data demonstrated that *NANOG* and *NANOGP1* regulate each other, which additionally emphasised their similarity. Other transcriptional changes observed in the overexpression experiments would require further investigation.

5.2.2.4 Is the downregulation of NANOGP1 required for hPSC capacitation?

NANOGP1 expression levels are elevated in naïve compared to primed hPSCs, as shown in Section 3.2. In keeping with this, I found that *NANOGP1* expression significantly decreases during naïve-to-primed state transition (termed capacitation) (Rostovskaya et al., 2019), while the opposite is observed during primed-to-naïve reprogramming (Collier et al., 2017) (Figure 5.39). This expression pattern is reminiscent of a naïve pluripotency transcription factor, such as *KLF17*, whereas *NANOG* expression is more constant and does not change to the same extent between primed and naïve hPSCs during these cell fate transition experiments (Figure 5.39).

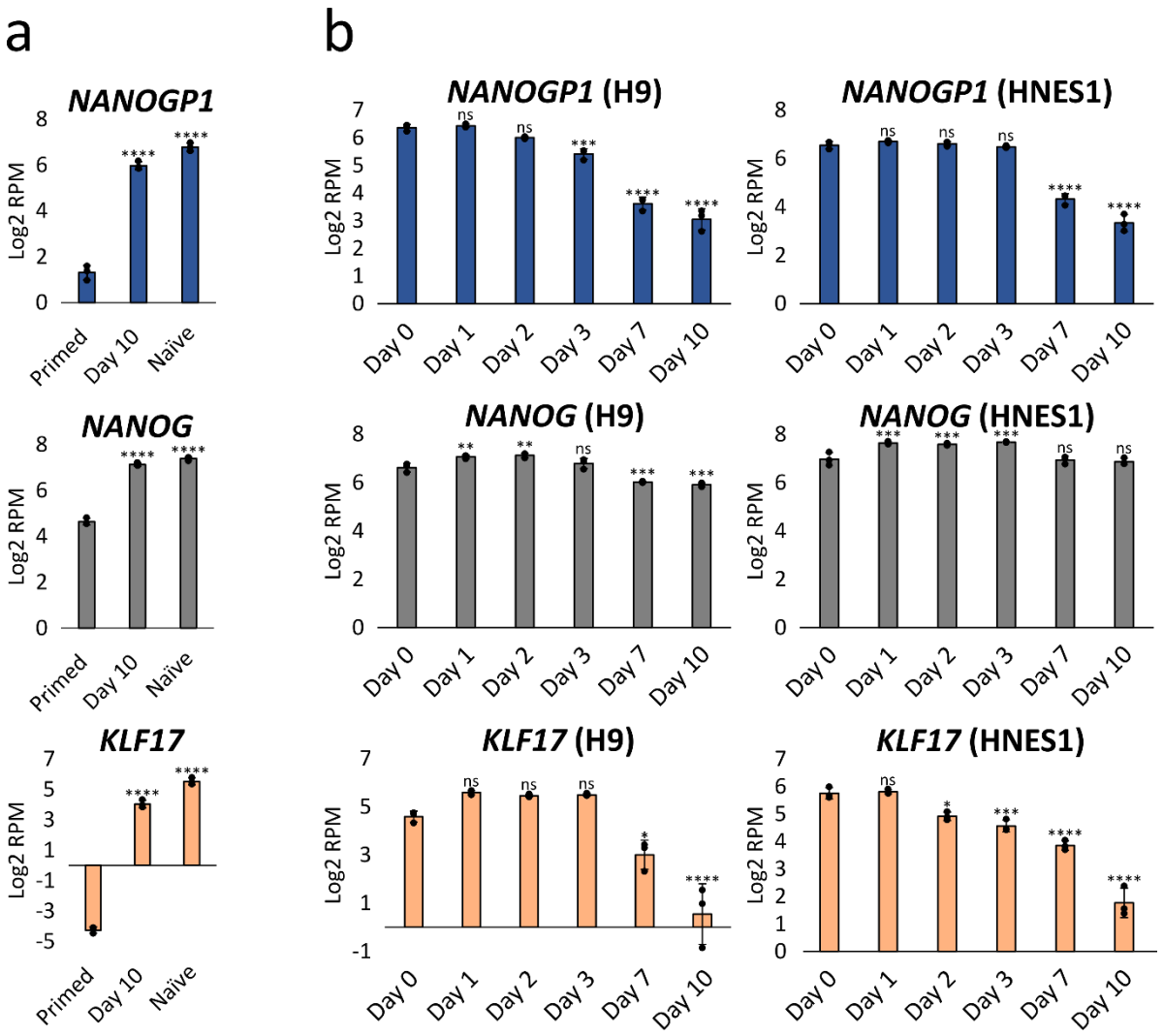


Figure 5.39 Bar charts showing RNA-seq expression values for *NANOG*, *NANOGP1* and *KLF17* in primed-to-naïve reprogramming (a; Collier et al., 2017) and capacitation (b; Rostovskaya et al., 2020) experiments. Gene expression is in Log2 RPM (reads per million) and is normalised to Day 0. Mean ±SD (n=3) is shown. One-way ANOVA with Dunnett's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), $p < 0.0005$ (***), $p < 0.00005$ (****)). 'Primed' and 'Day 0' were used as controls. H9 and HNES1 – naïve hPSC lines.

Here I tested whether preventing the normal downregulation of *NANOGP1* during naïve-to-primed capacitation could impede the transition in cell state. To achieve this, I maintained *NANOGP1* OE naïve hPSCs in the presence of doxycycline to sustain high levels of *NANOGP1* during the established capacitation protocol. This protocol uses N2B27 base medium supplemented with XAV939 which normally promotes formative transition by supressing Wnt signalling in the course of 14 days (Rostovskaya et al., 2019). During the 14-day capacitation time course, *NANOGP1* overexpression was

stable and was also maintained in all cells, as demonstrated by the GFP RT-qPCR and flow cytometry analysis (**Figure 5.40**).

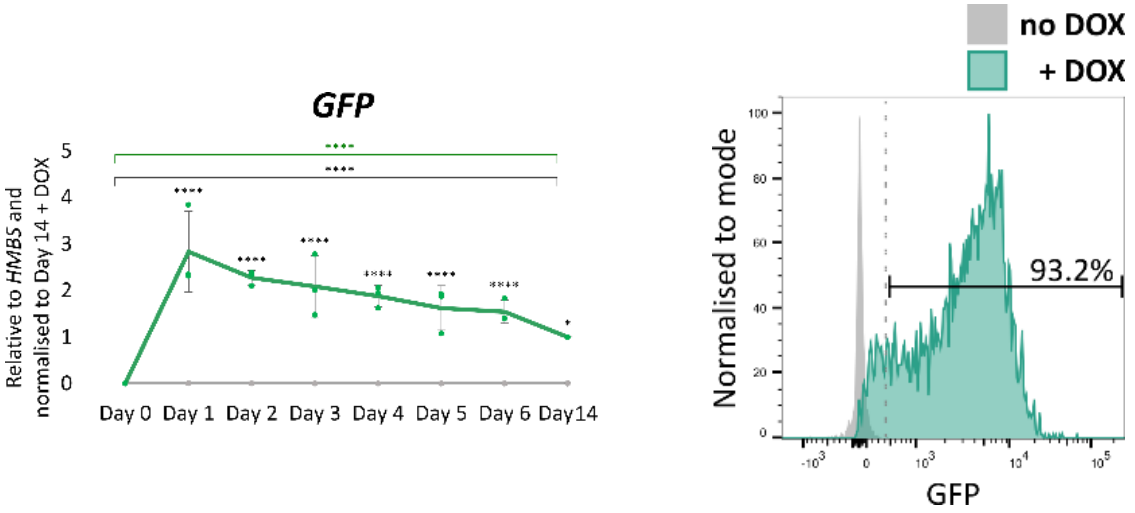


Figure 5.40 Line graphs (left) and flow cytometry histogram (right) showing *GFP* expression during the capacitation time course. DOX – doxycycline. **Line graph:** RT-qPCR values are relative to *HMBS* expression and normalised to Day 14. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)). **Histogram:** GFP expression was measured on Day 6, n=1.

Efficiency of the capacitation was analysed by flow cytometry. Four cell surface markers were used to assess presence of the naïve (CD77+SUSD2+) and primed (CD24+SSEA4+) hPSC populations. Here I show flow cytometry analysis of the capacitation experiment on Day 1, Day 2 and Day 6 (Figure 5.42). By Day 6, presence of the primed CD24+SSEA4+ double-positive hPSC population was obvious in both *NANOGP1* OE and the non-induced control, indicating that the capacitation was successful in both cases and was not blocked by the elevated levels of *NANOGP1*. However, the process was not the same between the *NANOGP1* OE and the control sample. A rapid decrease in the proportion of CD77+SUSD2+ double-positive naïve cells was observed in the induced overexpression sample already by Day 2; this was in contrast to the non-induced control. Also, by Day 6, in addition to the primed CD24+SSEA4+ double-positive population, *NANOGP1* OE hPSCs also had a separate subpopulation, SSEA4-CD24- and SSEA4+CD24-, presumably representing differentiating cells (Figure 5.41). Morphological signs of differentiation in the induced hPSCs were already observed by Day 2: *NANOGP1* OE hPSCs became flat, and the shape of the colonies was irregular (Figure 5.42). In contrast, the non-induced control preserved the naïve-like morphology. Collectively these data show that the increase of *NANOGP1* transcription led to a rapid exit from naïve pluripotency already by Day 2 and even thought *NANOGP1* OE hPSCs were able to progress in the capacitation, some cells were differentiating instead.

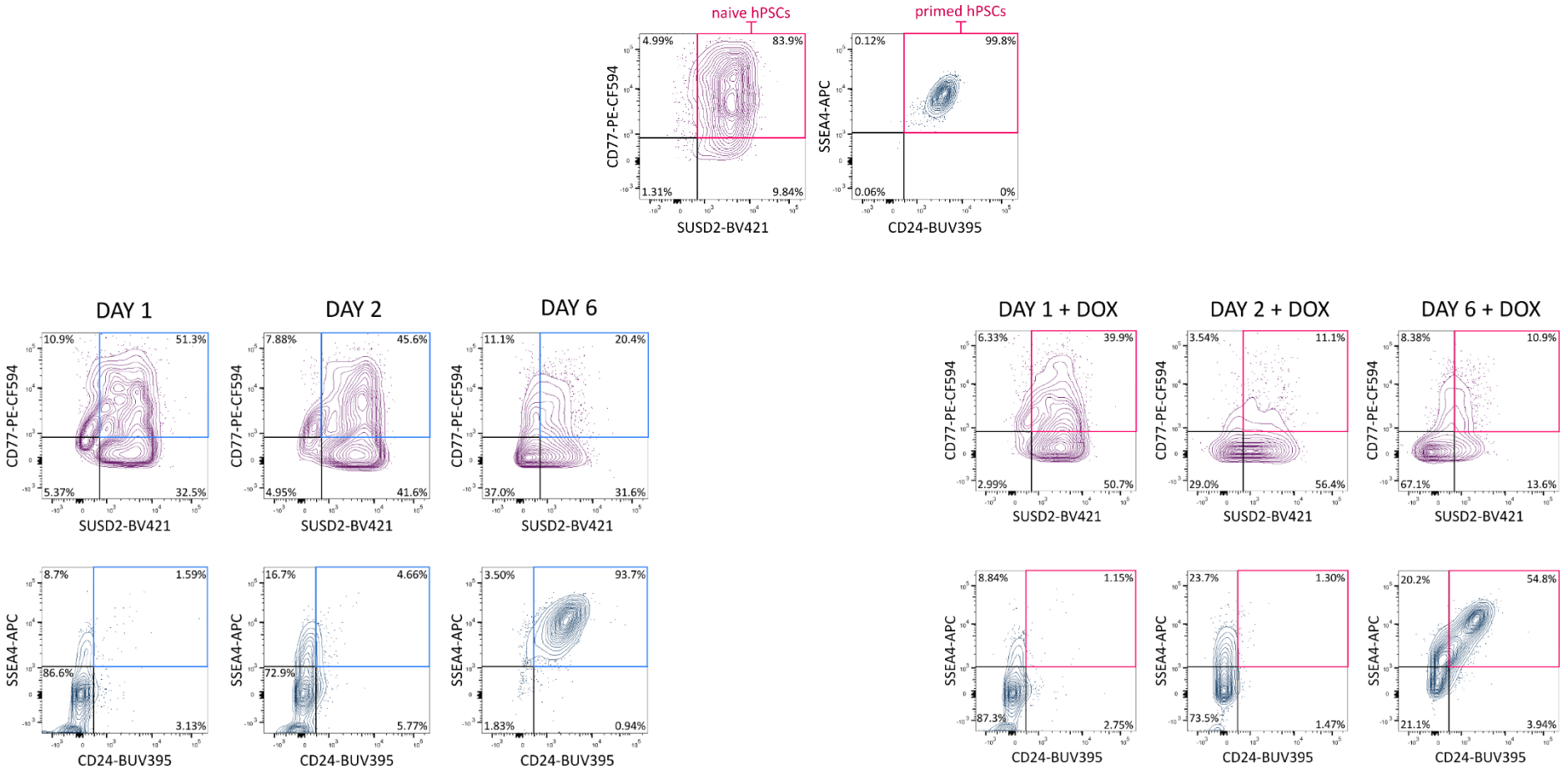


Figure 5.41 Flow cytometry contour plots showing the primed and naïve cell surface marker expression in the capacitation experiment on Day 1, Day 2and Day 6 (bottom), compared to the naïve and primed controls (top). Naïve markers: CD77, SUSD2; primed markers: CD24, SSEA4. Successful capacitation is indicated by loss of the naïve markers, coupled with expression of the primed markers. N=1.

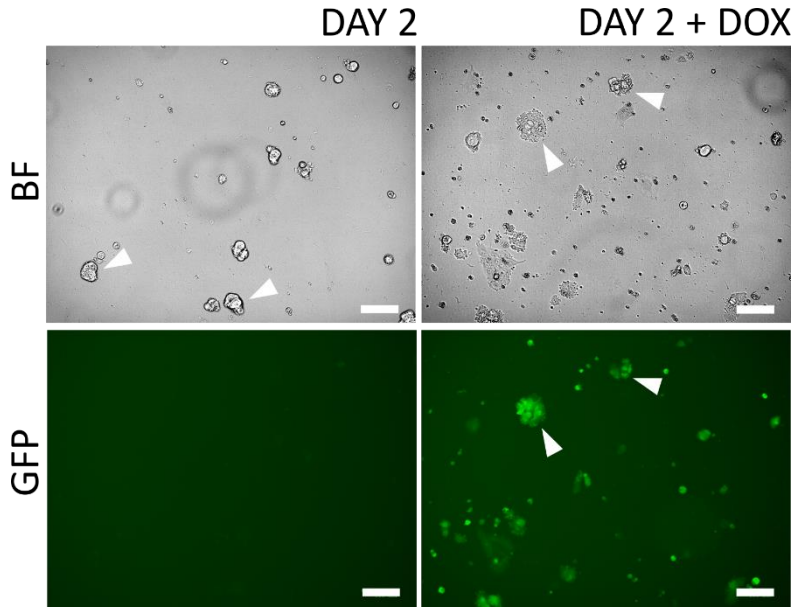


Figure 5.42 Fluorescence and bright filed (BF) microscope images illustrating GFP reporter expression and the cell morphology phenotype in the capacitation experiment. DOX – doxycycline. Scale, 100 µm. Arrowheads indicate the difference in morphology between the induced and non-induced ell lines, pluripotent-looking in Day 2 and flat and differentiation in Day 2 + DOX.

In addition to the presence of differentiating cells, between Day 2 and Day 6, the induced hPSC line exhibited an increased percentage of dying cells, reaching ~25% of the overall population on Day 4 (quantified using Countessa Cell Counter and trypan blue dye). This was in contrast with the control hPSCs, in which the percentage of dying cells was ~6% during the experiment (Figure 5.43).

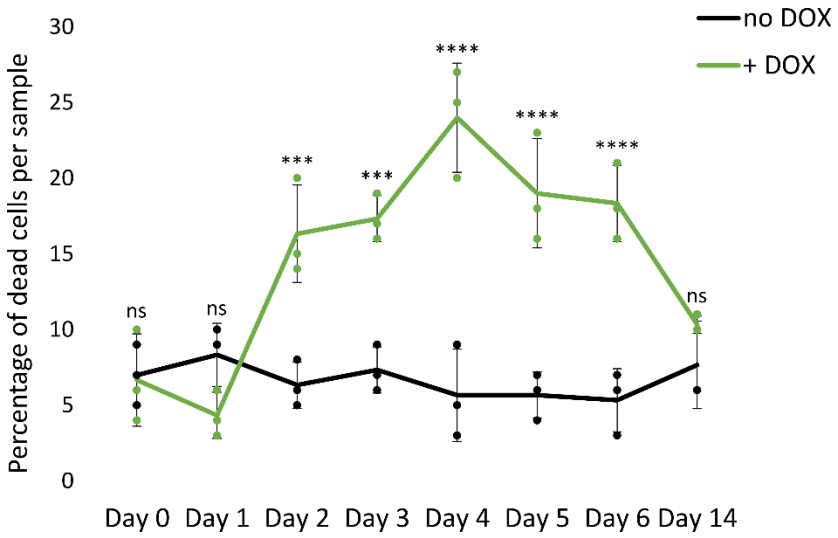


Figure 5.43 Line graph showing percentage of dead cells per sample in the capacitation experiment. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, 0.0005 (***), $p < 0.00005$ (****)) DOX – doxycycline.

Supporting the flow cytometry data, RT-qPCR analysis showed that both the induced and non-induced cell lines downregulated naïve marker (*TFCP2L1*, *KLF4*) expression, indicating exit from the naive pluripotency. Interestingly, *TFCP2L1* had a higher overall expression level in the induced cell line, in contrast to the non-induced one. Also, by the end of the time course, naïve marker *DPPA3* was downregulated by the non-induced line but not in the induced hPSCs (Figure 5.44).

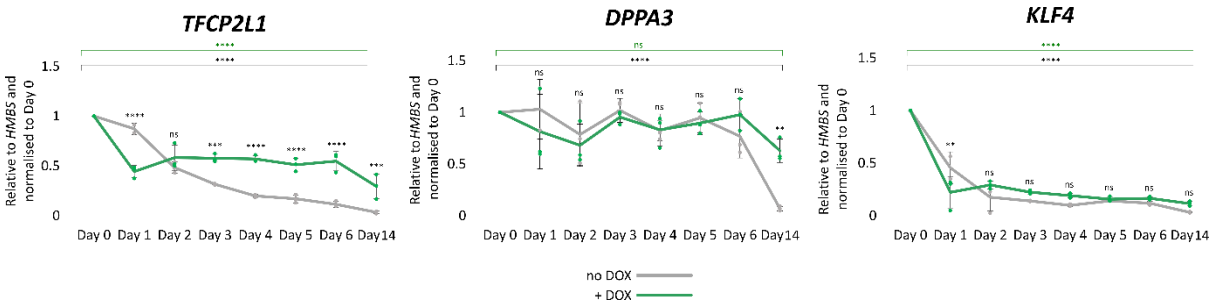


Figure 5.44 Line graphs showing expression of naïve markers in the capacitation experiment. RT-qPCR values are relative to *HMBS* expression and normalised to Day 0. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

Less efficient transition towards the primed state in the induced hPSCs was reflected in the expression of the two primed markers, *DUSP6* and *OTX2*. By the end of the experiment, on Day 14, cells in both conditions successfully upregulated the expression of *DUSP6*. However, the increase in transcript levels was more prominent and occurred earlier in the non-induced control. Interestingly, *OTX2* expression was not upregulated in the induced cell line at all, in contrast to the non-induced control cells. (Figure 5.45).

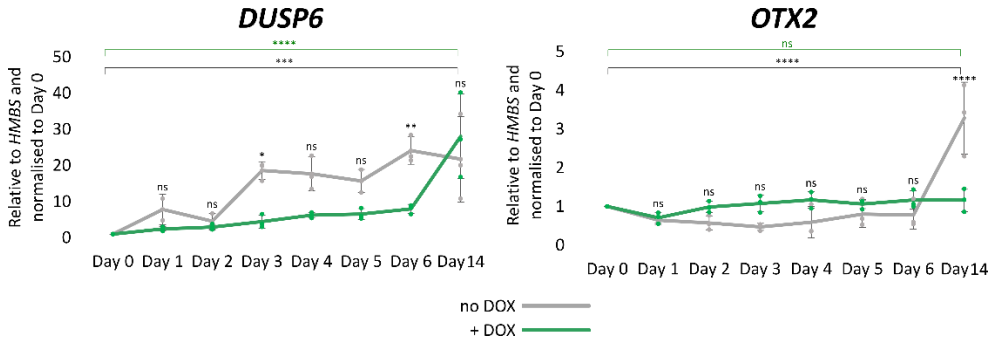


Figure 5.45 Line graphs showing expression of primed markers in the capacitation experiment. RT-qPCR values are relative to *HMBS* expression and normalised to Day 0. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

I hypothesised that the phenotype observed in *NANOGP1* OE hPSCs was linked to the rapid downregulation of *NANOG* caused by the autorepressive activity of *NANOGP1*, similar to what has been described previously in the chapter (Section 5.2.2.1). A rapid decrease in the *NANOG* levels (Figure 5.46) likely explains why *NANOGP1* OE cells started differentiating already by Day 2.

Additionally, it indicated that *NANOGP1* is capable to supress *OTX2*, a known target of *NANOG* repression (Su et al., 2018), showing that this property is also conserved by *NANOGP1*. It is also interesting that it was possible to generate primed hPSCs without upregulation of *OTX2*, indicating that this primed pluripotency factor is not required for the transition into the primed state.

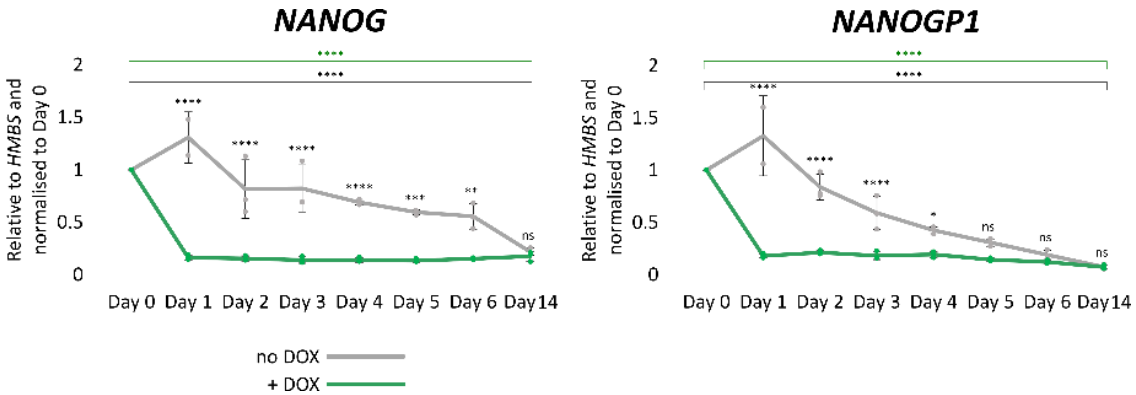


Figure 5.46 Line graphs showing expression of endogenous *NANOG* and *NANOGP1* in the capacitation experiment. RT-qPCR values are relative to *HMBS* expression and normalised to Day 0. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

To explore which possible differentiation routes *NANOGP1* OE were following, I analysed expression of three non-pluripotency lineage associated markers: *T-BRA* (mesoderm), *OLIG3* (neural lineage) and *GATA6* (endoderm). Based on the RT-qPCR analysis*, NANOGP1* overexpression led to the transient induction of *GATA6* between Day 1 and Day 4, whereas the mesodermal markers *T-BRA* and *OLIG3,* were not upregulated Figure 5.47. Whether the observed phenotype was related to the differentiation towards endoderm would need to be investigated more in depth in the future.

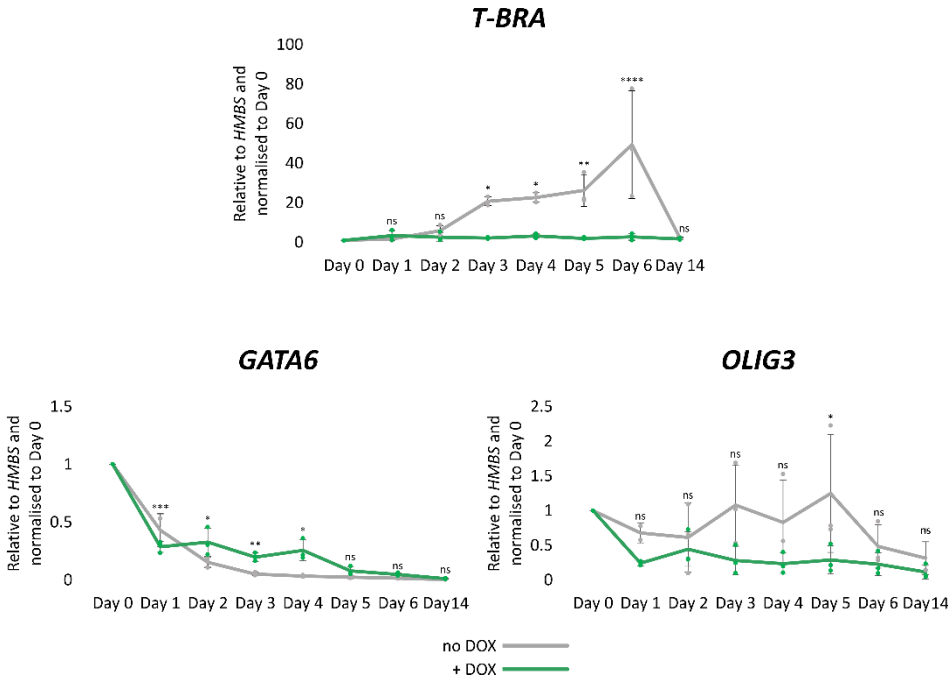


Figure 5.47 Line graphs showing expression of differentiation markers in the capacitation experiment. RT-qPCR values are relative to *HMBS* expression and normalised to Day 0. Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Sidak's multiple comparisons test was performed (ns – nonsignificant, $p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); DOX – doxycycline.

By the end of the capacitation experiment, CD77+SUSD2+ naïve cells could not be detected in either of the induced or non-induced hPSCs (Figure 5.48). Additionally, by Day 14, both conditions also had a substantial SSEA4+CD24+ population of primed cells. Almost all cells in the non-induced sample were primed-like (98.2% SSEA4+CD24+), compared to 88.6% in the induced hPSC line Figure 5.48. Notably, the *NANOGP1* OE population was not as homogenous as the non-induced control, likely demonstrating residual differentiation effects. However, morphologically, both lines appeared similar and primed-like (Figure 5.48).

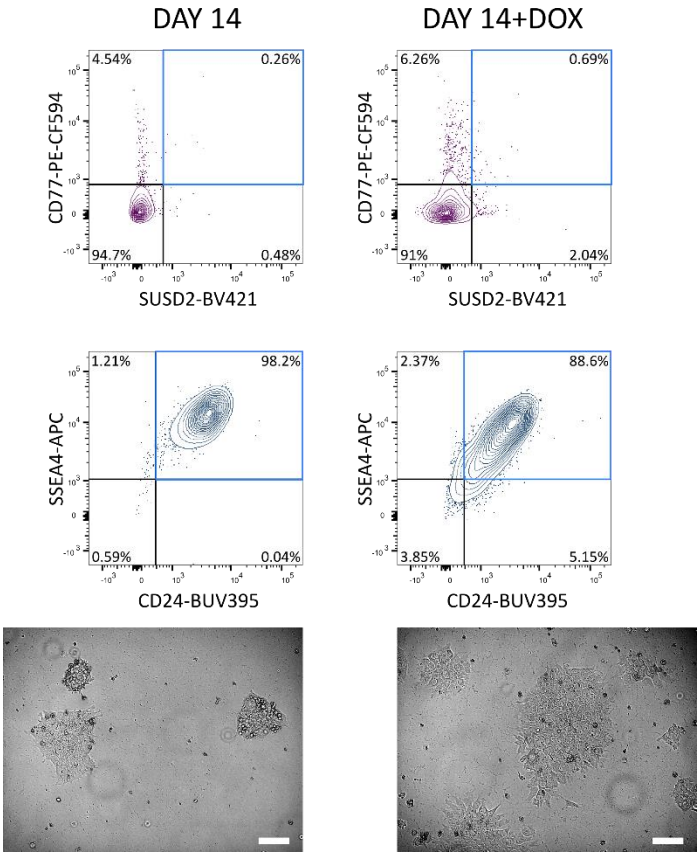


Figure 5.48 Flow cytometry contour plots showing the primed and naïve cell surface marker expression in the capacitation experiment on Day 14 (top) and microscope images showing cell morphology of the capacitation experiment Day 14 (bottom). Scale, 100 µm. The flow cytometry experiment was performed once.

Overall, maintaining high levels of *NANOGP1* disrupts the ability of hPSCs to undergo appropriate capacitation. Nevertheless, *NANOGP1* OE hPSCs still exited the naïve state, meaning that *NANOGP1* is not operating to maintain naïve pluripotency. Instead, *NANOGP1* OE hPSCs less efficiently transitioned to the primed state, and a large proportion of cells were lost to cell death or cell differentiation, presumably, due to the autorepressive activity of *NANOGP1* to downregulate *NANOG*.

5.2.2.5 Does *NANOGP1* overexpression promote primed-to-naïve reprogramming?

In primed hPSCs, *NANOG* is capable of inducing cell reprogramming when overexpressed with *KLF2* in 2iLIF culture medium (Takashima et al., 2014). To investigate, whether *NANOGP1* is also capable of promoting primed-to-naïve reprogramming, it was overexpressed together with human *KLF2* in primed hPSCs using a tetracycline-inducible (TetON) system in 2iLIF medium for 12 days (Takashima et al., 2014). Notably, on its own, 2iLIF is not sufficient to induce reprogramming, and the overexpression of the two transgenes is necessary (Takashima et al., 2014).

Here, I tested all three *NANOGP1* isoforms, each in combination with *KLF2*, and used *NANOG+KLF2* and *KLF2*-only TetOn systems as positive and negative controls, respectively. *NANOG*/*P1*

constructs were linked to GFP and *KLF2* – to RFP. TetOn cell lines were generated in the same way as outlined in Section 5.2.2.2 and Section 2.1.7.

During the reprogramming, fluorescent reporters GFP and RFP were used as a proxy for transgene levels, and therefore it was possible to ensure that they were all overexpressed at the same level. To ensure that the starting cell populations had similar *NANOG*/*NANOGP1* and/or *KLF2* overexpression levels, prior to the reprogramming, all of the TetON primed hPSC lines were induced with doxycycline for 48 h. After that, they were flow sorted by the expression of GFP and RFP (Figure 5.49). Non-induced controls were also validated on the flow cytometer to confirm the absence of TetOn 'leakage'. GFP+RFP+ double-positive cells in *NANOG*/*NANOGP1*+*KLF2* samples, and an equivalent number of RFP+ cells in the *KLF2*-only cell sample, were seeded in mTeSR+ Y-27632 medium and maintained for two days. The same number of non-induced cells was plated for a negative control (importantly, these negative control cells have never been induced).

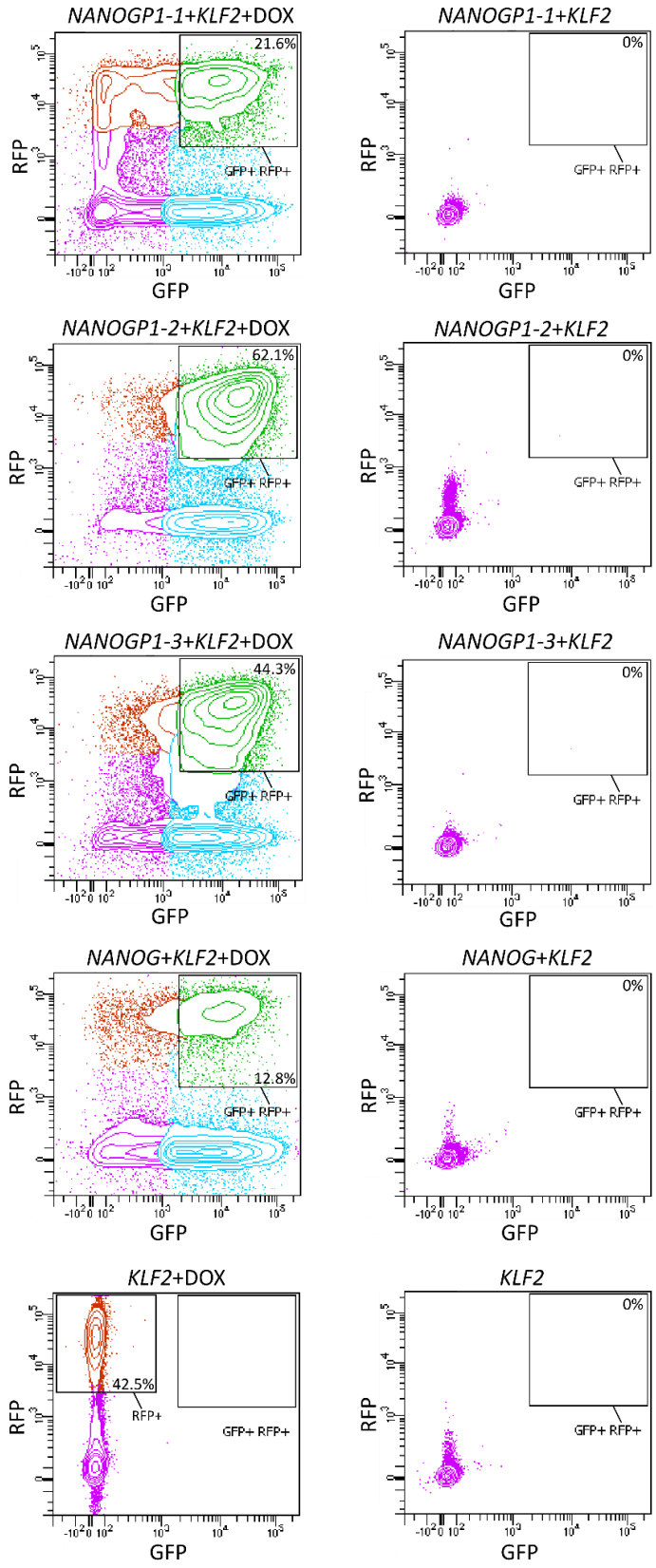


Figure 5.49 Flow cytometry contour plots showing RFP and GFP expression prior to the *NANOGP1+KLF2* reprogramming experiment. Percentages of GFP+RFP+ and RFP+ populations are indicated.

After two days, the cell media was replaced with 2iLIF+/-DOX. Two more days later, all induced cell colonies were GFP+RFP+ or RFP+, as expected, while no reporter expression was detected in the non-induced lines (Figure 5.50).

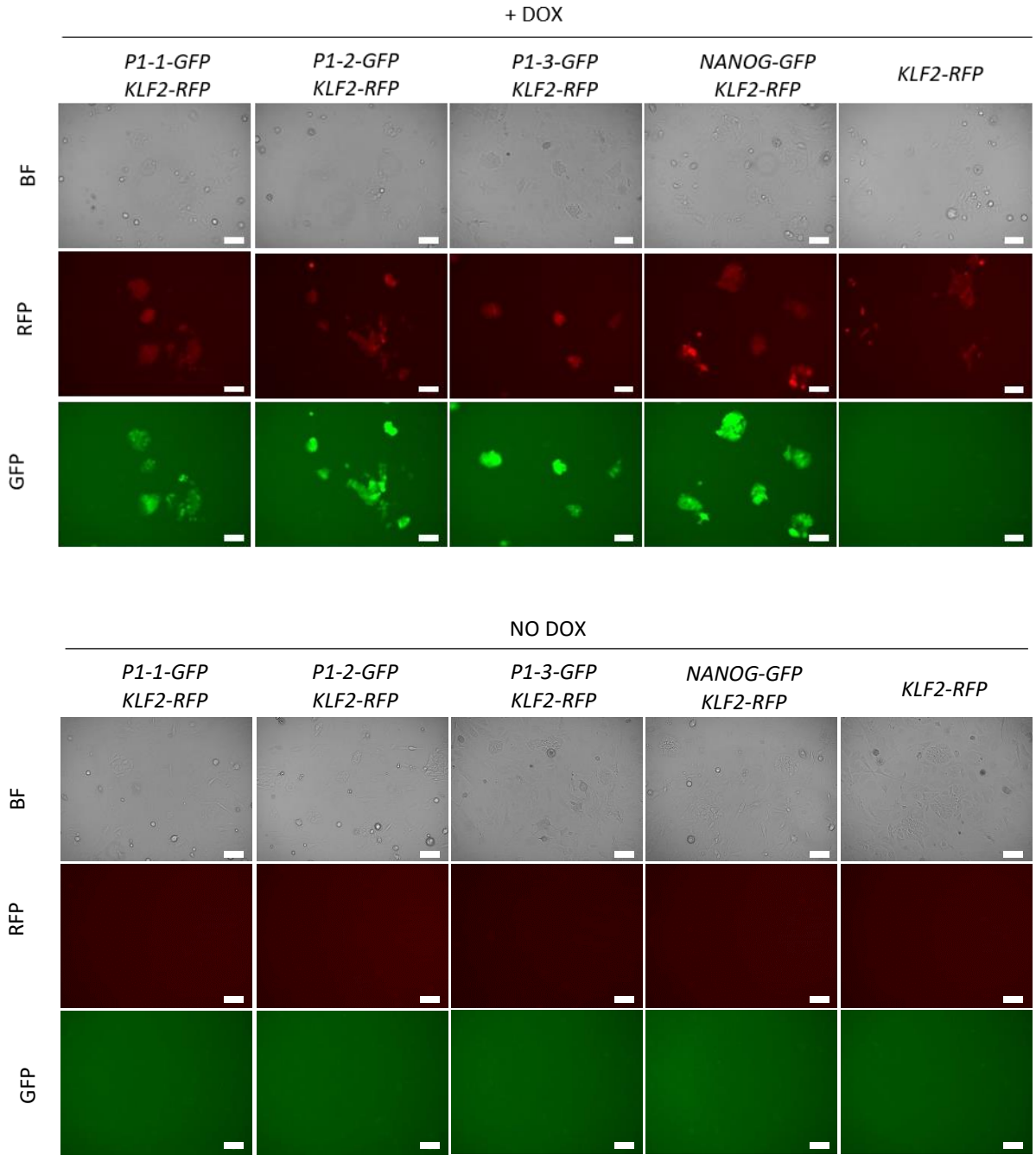


Figure 5.50 Bright field (BF) and fluorescence images showing reporter expression on Day 2 of the *NANOGP1+KLF2* reprogramming experiment. DOX – doxycycline. Scale, 100 µm.

By Day 7, *NANOG+KLF2* and *NANOGP1+KLF2* overexpressing colonies were increasing in size and were acquiring a naive-like domed morphology. In contrast, the *KLF2*-only cell line maintained primed hPSC-like morphology. The non-induced cell lines had flat morphology that resembled primed

hPSCs. All these morphological changes can be seen in Figure 5.51. After Day 7 all the non-induced negative control lines were abandoned as they did not show signs of reprogramming, and only the induced lines were propagated further. By Day 12 of reprogramming, numerous domed colonies with naïve hPSC morphology were observed in the *NANOGP1+KLF2* cultures, while the *KLF2* only cell line contained only differentiated cells (Figure 5.52).

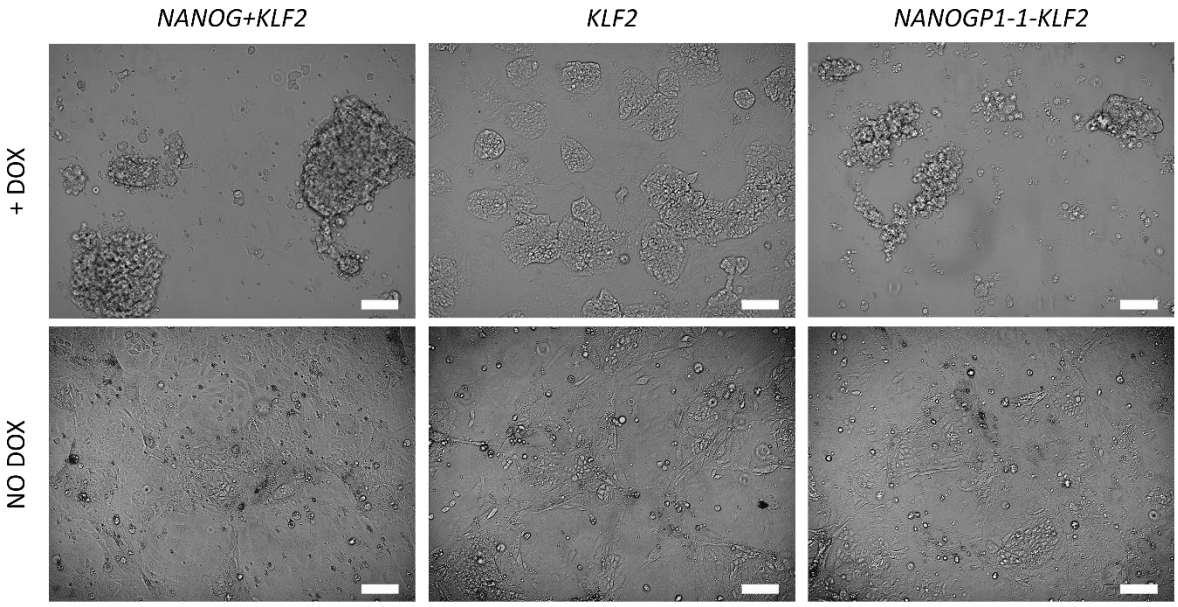


Figure 5.51 Bright field (BF) images showing cell morphology on Day 7 of the *NANOGP1+KLF2* reprogramming experiment. *NANOGP1-1* isoform only is shown for simplicity. DOX – doxycycline. Scale, 100 µm.

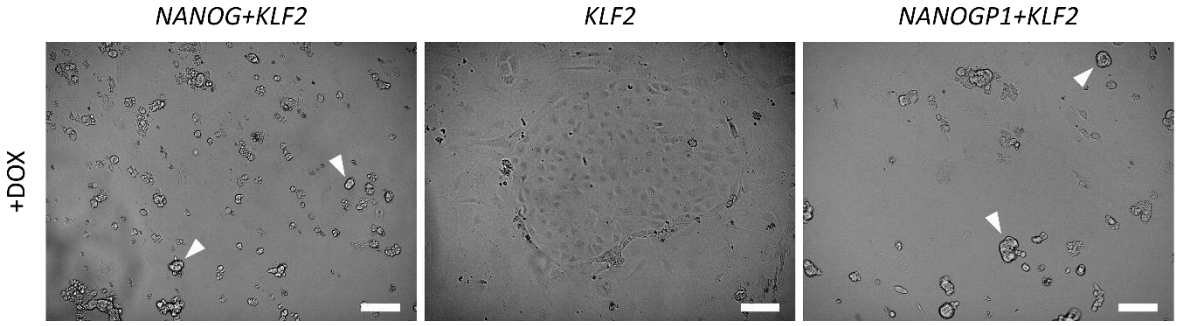


Figure 5.52 Bright field (BF) images showing cell morphology on Day 12 of the *NANOGP1+KLF2* reprogramming experiment. *NANOGP1-1* isoform only is shown for simplicity. DOX – doxycycline. Pluripotent colonies are indicated by white arrowheads. Scale, 100 µm.

The reprogrammed colonies were positive for alkaline phosphatase activity, and the number of positive colonies was similar for cultures overexpressing either *NANOGP1* or *NANOG*. *KLF2* only control did not have any pluripotent alkaline positive colonies (Figure 5.53).

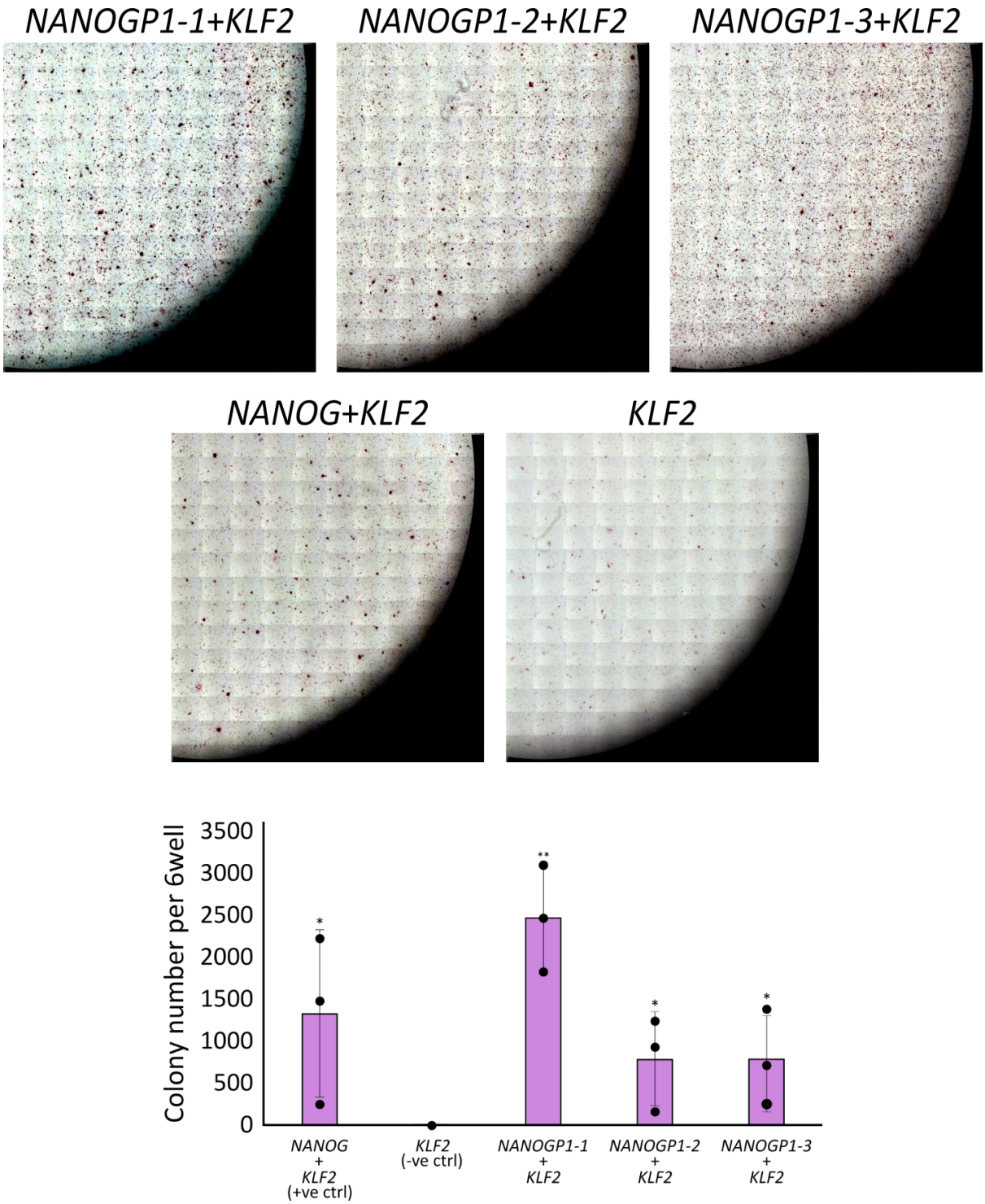


Figure 5.53 Alkaline phosphatase assay of the *NANOGP1+KLF2* reprogramming experiment: microscope images (top), bar chart (bottom). Individual replicates (n=3) and mean ± SD are shown. T-test was performed ($p < 0.05$ (*), $p < 0.005$ (**)).

The *NANOG+KLF2* and *NANOGP1+KLF2* induced cells upregulated naïve pluripotency markers, including *DPPA3* and *TFCP2L1*, and maintained high *OCT4* expression. All three *NANOGP1* isoforms showed similar effects. These changes were comparable to the positive control cells overexpressing *NANOG+KLF2*, thereby confirming naïve cell identity of the *NANOGP1* overexpressing cells (Figure 5.54).

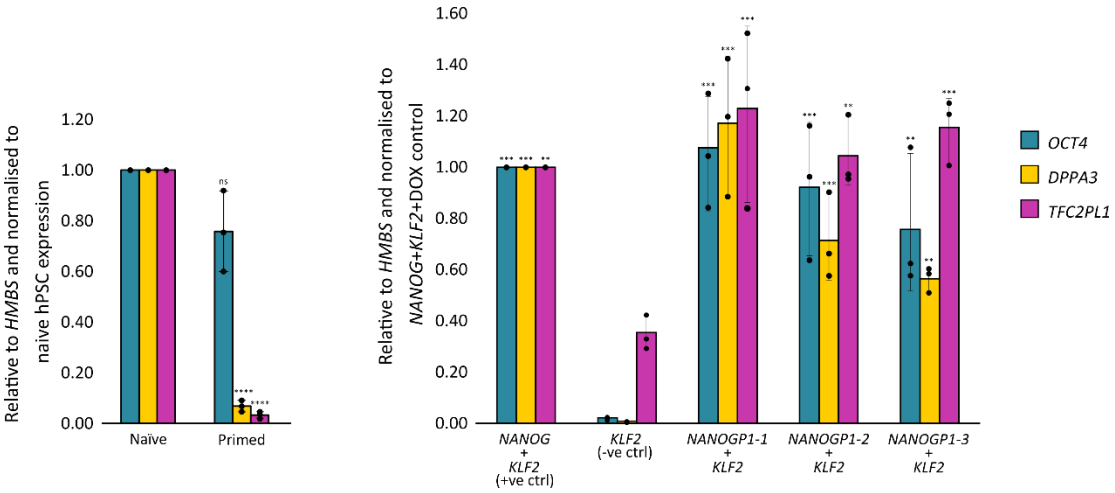


Figure 5.54 Bar charts showing pluripotency gene expression in *NANOGP1+KLF2* reprogramming experiment, Day 12 (right). Gene expression in established naïve and primed cell lines shown as controls (left). RT-qPCR values are relative to *HMBS* expression and normalised to NANOG+KLF2 sample (right) and Primed sample (left). Individual replicates (n=3) and mean ± SD are shown. One-way ANOVA with Dunnett's multiple comparisons test was performed ($p < 0.05$ (*), $p < 0.005$ (**), 0.0005 (***), $p < 0.00005$ (****)); right) and t-test (ns – nonsignificant, $p < 0.00005$ (****); left).

Flow cytometry analysis using stringent markers of naïve pluripotency (CD24 negative; CD75 positive; SUSD2 positive) validated successful cell state conversion in the *NANOGP1*-overexpressing cells. The *NANOGP1-1+KLF2* cell line had the largest proportion of reprogrammed naïve cells out of all tested samples (Figure 5.55, Figure 5.56).

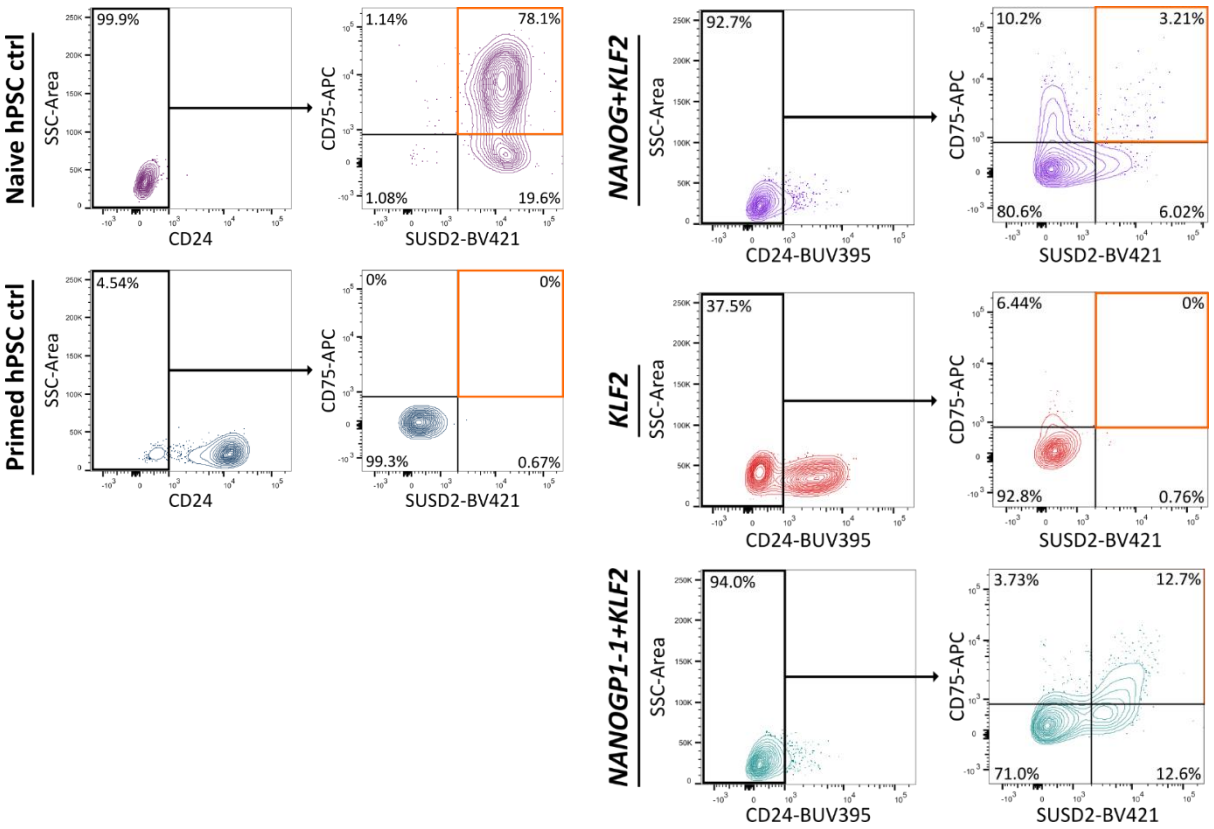


Figure 5.55 Flow cytometry contour plots showing the primed and naïve cell surface marker expression in the *NANOGP1+KLF2* reprogramming experiment on Day 12, compared to the naïve and primed controls. CD24 – primed hPSC marker. CD75 and SUSD2 – naïve hPSC markers. Contour plots demonstrate percentage of CD75+SUSD2+ population within CD24- population. N=2

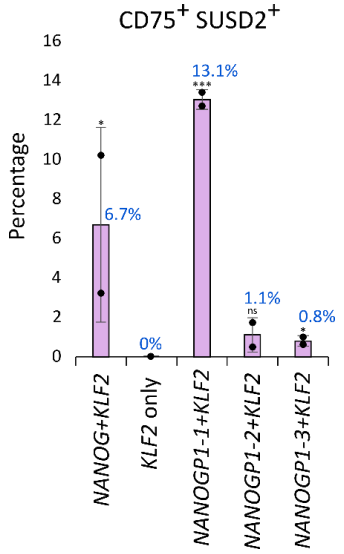


Figure 5.56 Bar chart showing the percentage of the naïve population (by cell surface marker) identified in the *NANOGP1+KLF2* reprogramming experiment, Day 12. Mean ±SD (N=2) is shown. T-test was performed (ns – nonsignificant, $p < 0.05$ (*),0.0005 (***)); data were compared to '*KLF2* only'.

Importantly, in all of the assays described above, the overexpression of *KLF2* alone did not induce reprogramming, which confirms the critical contribution of *NANOGP1* in establishing naïve pluripotency. Moreover, the change in pluripotent state was stable because the *NANOGP1-1*-induced reprogrammed cells not only retained but improved their cell-surface marker phenotype when cultured for seven passages without doxycycline in t2iLGo medium Figure 5.57.

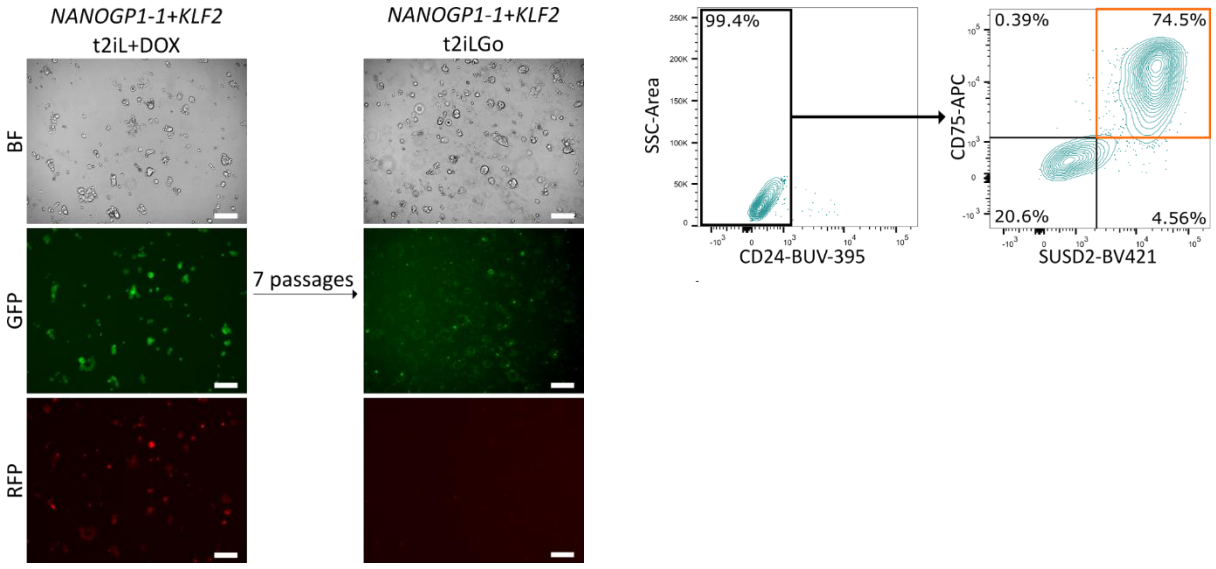


Figure 5.57 Fluorescence and bright field (BF) microscope images showing cell morphology during the adjustment to t2iLGo naïve culture medium (left). Flow cytometry contour plots showing the naïve cell surface marker expression in the *NANOGP1-1+KLF2* reprogrammed line in t2iLGo (right). Scale, 100 µm. N=1.

Overall, these results indicate that, like *NANOG*, *NANOGP1* is capable of reprogramming hPSCs into a naïve state, thereby demonstrating functional conservation in igniting the naïve pluripotency network.

5.3 Discussion

In this chapter I discovered that *NANOGP1* has functional properties in hPSCs, which was investigated by means of gene expression downregulation and overexpression. As a result, some functional properties were found to be conserved with *NANOG*, such as autorepression and the ability to reprogramme primed hPSCs into the naïve state. Interestingly *NANOG* KD caused a two-fold upregulation of NANOGP1 expression in primed hPSCs, but not in naïve hPSCs. This observation suggests that NANOG/NANOGP1 autorepression, described in Section 5.2.2.3, is also active in the primed state, but the mechanism is different from the naïve state due to two different knockdown phenotypes. Some functions of *NANOG* were found not to be shared with *NANOGP1*, such as the requirement for the maintenance of the naïve pluripotent state. Additionally, *NANOGP1* expression

alone could not provide functional redundancy for *NANOG*, and was therefore not sufficient to maintain naïve hPSCs in its absence.

For the first time, this study described generation of an inducible loss of function system in naïve hPSCs. Prior to this, similar experiments have been performed, however, they were mostly aiming to generate gene knockout systems, such as (Chen et al., 2015a). Using inducible loss of function system has a series of advantages, such as, for instance, the ability to transiently induce reporter cell lines for FACS selection. Additionally, with an inducible knockdown it is possible to titrate the system by adjusting the level of knockdown, as well as to choose length of the time course and whether cells could be re-used/rescued afterwards. Possible drawbacks of using such systems in naïve hPSCs could be unpredicted silencing of a plasmid and/or AAVS1 locus.

Analysis of the *NANOG* knockdown dataset, described in this chapter, presents a careful and detailed investigation of a loss of a naïve pluripotency factor. In the future, it would be interesting to integrate the RNA-seq data obtained here with the NANOG ChIP-seq and the chromatin interaction maps (Chovanec et al., 2021) to build up a more complete picture of how loss of *NANOG* disrupts gene regulatory networks.

This chapter also showed that loss of *NANOG* in the naïve hPSCs leads to transitioning towards the trophoblast fate, which was confirmed by comparison of the data to the embryo scRNA-seq data set (Xiang et al., 2020). This raises an existing question: is downregulation of *NANOG* what defines whether the cell in human embryo would become a TE cell - or not?

Unfortunately, I was unable to develop efficient shared CRISPRi gRNA hPSC lines, which prevented me from testing whether the shared expression knockdown would have a stronger/altered effect to that of *NANOG* KD. Downregulating both *NANOG* and *NANOGP1* at the same time could open new interesting research avenues and highlight shared functions, currently masked by the high *NANOG* expression level. In the future, multi-gRNA constructs could be used instead of singular gRNAs, as it could allow to increase the gRNA targeting efficiency.

To conclude this part, the most likely reason *NANOGP1* is not required for naïve pluripotency maintenance is that its expression levels are significantly lower than those of *NANOG*, making it redundant, even if it could have the same effect as NANOG. This 'masking' effect could potentially be resolved by performing *NANOGP1* KD in *NANOG* Het (heterozygous) hPSCs, where *NANOG* levels are lower but still sufficient to maintain pluripotency.

In the *NANOGP1* overexpression assays, I discovered that *NANOGP1* autorepression function is conserved and, additionally, so is the autorepression function of *NANOG* in the naïve hPSCs. Moreover, *NANOG* and *NANOGP1* have the ability to transcriptionally repress each other. The

similarity of *NANOG* and *NANOGP1* overexpression phenotypes suggested that *NANOGP1* was contributing to the function of *NANOG*. Interestingly, *OTX2* was strongly suppressed when *NANOGP1* was upregulated, showing that this repressive function of *NANOG* (Acampora et al., 2017; Su et al., 2018) is also likely shared by *NANOGP1.* A limitation of this experiment was that the overexpression levels of *NANOGP1* were rather high, and in the future, it might be beneficial to titrate the level of overexpression down, to the endogenous level of *NANOG* expression, for instance. Also, the flow cytometry analysis of the capacitation assay was only performed once and would need to be replicated in the future.

Another *NANOG*-like function discovered here was that *NANOGP1* could reprogramme primed hPSCs to the naïve state. This conserved role was likely mediated via its conserved homeodomain (Theunissen et al., 2011b). Does this suggest then, if the homeodomain is conserved, that in *NANOGP1* OE, *NANOGP1* can bind and activate the same gene targets as *NANOG*? Has it developed an ability to bind other targets? In my opinion, this could be an interesting topic to research in the future.

In summary, this chapter revealed several conserved *NANOGP1* functions, as well as showed that it is not required for the pluripotency maintenance. Collectively, the results allowed me to conclude that *NANOGP1* is a conserved functional duplicate, and it likely appears not to be required for the pluripotency maintenance due to a lower expression level, compared to that of *NANOG*.

6 Summary and conclusions

6.1 *NANOG/NANOGP1* duplication: summary of the main findings, study limitations and potential functions

This thesis presents the characterisation and functional study of *NANOGP1*, a tandem duplicate of the key pluripotency gene *NANOG*. The findings show that the *NANOGP1* RNA expression pattern is different from that of *NANOG*, both in hPSCs and human embryos. Putative regulatory regions upstream of the *NANOG* and *NANOGP1* TSSs were also found to be formed by duplication, and differential SOX2 binding in those regions was identified as one putative reason for the distinct expression patterns of the two genes; the mechanism behind the differential expression requires further investigation. The work presented in thesis established for the first time that endogenous *NANOGP1* is translated into a stable protein, thereby challenging previous research that concluded that NANOGP1 is unable to form a protein (Booth and Holland, 2004). Moreover, NANOGP1 was found to share chromatin binding sites with NANOG, and, surprisingly, to have a small number of unique binding regions, which overlap with the binding profile of REST, a repressor of neural fate in the non-neuronal cell types. This finding raises the exciting possibility that NANOGP1 could have a NANOG-independent role in preventing neural differentiation. In functional assays, *NANOGP1* could reprogramme primed hPSCs into the naïve state, as well as reduce its own expression levels presumably through an autorepressive feedback mechanism, which collectively demonstrated functional conservation with *NANOG*. *NANOGP1* overexpression could also lower the RNA levels of *NANOG* and vice versa, suggesting that the autorepressive mechanism was affecting both genes, due to the high similarity of their functional domains and putative regulatory regions. At the same time, *NANOGP1* expression was found to be redundant in the maintenance of naïve pluripotency, potentially due to it being 'masked' by the higher levels of *NANOG*. Interestingly, the overall phenotype for *NANOGP1* resembles that of *KLF17*, recently described by Lea and colleagues (Lea et al., 2021): both factors are highly expressed in the naïve hPSCs, are capable to induce primed-to-naïve reprogramming, but lack a knock-down phenotype, suggesting compensation by other factors. In summary, this thesis provides evidence that NANOGP1 is a functional protein in human pluripotency, and its properties are partially but not fully conserved with its ancestor.

The conservation of *NANOG*-associated functions in *NANOGP1* could be explained relatively easily, since the predicted homeodomain and transactivation domain sequences of the two proteins are almost identical. Mutations confirmed in the N-terminal domain, however, remain to be explored in the future and, in my opinion, open a new exciting research route. The reason for this is the potential function of the N-terminus of human NANOG remains poorly understood. In mouse, both the N-terminus and C-terminus have a transactivation ability, while human NANOG lost this function

in the N-terminal (Chang et al., 2009; Do et al., 2009; Oh et al., 2005). It had been suggested (Chang et al., 2009), that the NANOG N-terminus could act as a 'transrepressor', however this has not been confirmed in hPSCs yet. Therefore, the large N-terminal deletion and an array of substitutions in the remainder of the N-terminus of NANOGP1 is likely affecting the function that it is not even fully understood in NANOG. This way, by investigating *NANOGP1* further, more evidence could become available on the human-specific properties of *NANOG*, and vice versa. In my opinion, this ability to simultaneously investigate two evolutionarily-connected proteins is highly advantageous not only from their individual perspective, but also could be useful in understanding how/if they could function together and whether this cooperation would alter their individual properties. For instance, mutating NANOGP1 N-terminus could have consequences for its ubiquitination and protein turnover (Section 3.3.2), and if NANOG and NANOGP1 are capable of dimerising in naïve cells in the same way they do it in insect cells (Section 4.2.3), this could affect the stability of the NANOG:NANOGP1 vs. NANOG:NANOG dimers. A similar effect had been shown by Charrier et al., 2012 and Dennis & Eichler, 2016, which described that the truncated duplicate SRGAP2C can dimerise with its functional ancestor, leading to debilitating the heterodimer and, hence, the ancestral function. If this also occurs with NANOG and NANOGP1, this could be a potential mechanism enabling NANOGP1 to regulate NANOG function. It is worth noting that it is unlikely that NANOGP1 would completely disrupt the ancestral function of NANOG and instead affect it to a certain extent, as NANOGP1 does not appear to be dominant negative over NANOG. This potential role, however, does not mean that *NANOGP1* itself could lack other functions, as it was capable of reprogramming primed hPSCs and contributing to the *NANOG* function when overexpressed. A potential study to investigate the stability of NANOGP1 dimers and NANOG:NANOGP1 heterodimers could involve assessing proteins half-lives in a cycloheximide chase assay, as described in Kao et al., 2015.

Another reason to investigate *NANOGP1* and not only or primarily *NANOG* is that the development of the human epiblast appears to be different from other species, and the presence of human-specific mechanisms are currently fairly poorly understood. Indeed, inhibiting TGF-β/ACTIVIN/NODAL signalling leads to the human blastocyst losing its epiblast completely, which does not occur in other species (Blakeley et al., 2015; Boroviak et al., 2015). This emphasises that any human-specific features require further attention as they could shed light onto human early development. This current thesis, in my view, contributed to our understanding of human-specific processes. Not only it described the presence of a partially conserved pluripotency factor that could alter the function of its ancestor *NANOG*, but it also showed that other pluripotency factors, *OCT4* and *DPPA3*, have highly expressed pseudogenes. These novel components of naïve pluripotency have the potential to integrate into the regulatory network at the protein or RNA level, and in this way adding

complexity and versatility to the developmental programme of the human embryo. Therefore, investigating pseudogenes and other overlooked duplicates is highly beneficial for learning about human embryo development, as it could help to separate the knowledge about its biological processes from assumptions made while studying other model species, helping to fill the knowledge gap about early human embryo development.

6.2 Pseudogenes and the need to re-define their functional potential in early human development

This study demonstrated that a pseudogene, which had originally been classified as unfunctional, has the potential to contribute to early human pluripotency. The main reason for it being overlooked in the past was its predicted structure, significantly shortened in comparison with the ancestral copy *NANOG* (Booth and Holland, 2004). However, as this thesis and other similar studies show (Section 1.4.3), 'truncation' does equal 'absence of function'. This means that, perhaps, too frequently, truncated copies of duplicated genes are disregarded while they still bear the potential of functional contribution. Based on the data presented in this study, it is possible that *NANOGP1* is not, in fact, a pseudogene (a defective copy of a functional ancestor), but may instead be a paralogue gene copy of *NANOG*, similar to the highly diverged *NANOGNB*, discussed previously. This also made me question - what genomic structure could then be called a 'true pseudogene'? Is it really possible to conclude that a certain sequence had definitely lost all of its potential and will never gain any sort of functionality? In my opinion, since scientists cannot fully predict evolution and mostly assess its progress in retrospect, we cannot reject what has not occurred yet. In regard to pseudogenes and other duplicated copies, it is especially correct, since there is evidence that pseudogenes are involved in evolutionarily processes but how exactly is not fully understood. Section 6.2 briefly describes what we do and do not know about functional pseudogenes, allowing assessment of the novel finding of this thesis regarding potential *NANOGP1* functionality in a larger perspective.

How do pseudogenes of pluripotency factors appear? A recent study showed that genes that are highly expressed prior to duplication have higher chances to be preserved for a longer evolutionary period and a wider phylogenetic range, as shown in yeast (Mattenberger et al., 2017). This could explain why among the highest expressed pseudogenes in naïve hPSCs and presumably early epiblast cells I find copies of highly expressed pluripotency factors, as well as of ribosomal complex components and genes, responsible for the mitochondrial function (Section 3.2.8).

The majority of duplications in the human genome are segmental duplications, which were shown to drive evolution of Great Apes and humans in particular (Marques-Bonet et al., 2009a; Marques-Bonet et al., 2009b). *NANOG*, however, was formed by tandem duplication, an older

evolutionarily mechanism. Strikingly, in the case of human *NANOG*, a tandem duplication had occurred and was conserved at least twice – once, forming *NANOGP1*, and once, at a substantially earlier point, forming *NANOGNB*, which had diverged to such an extent that was only recognised as a duplicate of *NANOG* in 2017 (Dunwell and Holland, 2017). Independent *NANOG* duplications were also reported in birds (Cañón et al., 2006), guinea pigs and some fish species (Scerbo et al., 2014). In all these cases, *NANOG* duplicates bear high similarity to the original ancestral sequence. Does it mean that the *NANOG* region is somehow predisposed to duplication? In mouse, it was shown that *Nanog* retrotransposition occurs at a high frequency, which led to formation of several highly conserved retrogenes that are expressed in ESCs cells (Robertson et al., 2006). Interestingly, in human, the chromosomal area where *NANOG* is located, ch12p13, also contains *DPPA3*, *OCT4P3* and another pluripotency factor *GDF3*, and collectively is called a 'hotspot for teratocarcinoma' for the high rate of chromosomal abnormalities that appear there, eventually leading to cancer (Clark et al., 2004; de Jong et al., 1990; Murty et al., 1990; Pain et al., 2005). Moreover, chromosome 12p was also found to be one of the most common amplification hotspots in hPSCs, which tend to accumulate large genomic duplications during the hPSC culture (ISCI, 2011). Does that mean that amplification of the *NANOG*-containing region, commonly linked to cancer and genomic aberrations, could also be developmentally beneficial, based on the evolutionarily evidence?

If it is the case, then understanding what function pseudogenes normally have might help understanding the duplication of pluripotency pseudogenes as well. Typically, pseudogenes are expected to experience rapid evolution, gaining random mutations faster than their functional ancestors (Blake et al., 1992; Ophir and Graur, 1997). However, recently it had been demonstrated that if a genetic element is processed, it does not mean that it is non-functional. Processed pseudogenes can continue to exist as RNA involved in siRNA regulation of their parental copies, which was demonstrated in both mammals and Drosophila (Sasidharan and Gerstein, 2008). Some processed pseudogenes even developed to become microRNA in primates (Devor, 2006)

An additional copy of a gene is thought to contribute to gene dosage by increasing the amount of the parental copy. However, often, instead of a simple doubling of dosage, tandem duplicates develop novel expression patterns if the duplication involves exon and regulatory element shuffling and formation of chimeric structures, as shown in Drosophila (Rogers et al., 2017). Also, even if a new tandem copy contributes to the parental gene dosage, often the amount of expression is unpredictable. Instead of doubling, changes of up to 5-fold and higher expression differences have been reported, linking it to the 'position effect' (Loehlin and Carroll, 2016). Finally, it is also argued that in addition to regulating gene dosage, duplicated genes could contribute to cell robustness by compensating for each other's potential loss-of-function mutations (Brookfield, 1997; Diss et al., 2014;

Pickett and Meeks-Wagner, 1995). However, again, in some cases higher number of duplicates is associated with a higher chance of developing a disease and not the compensation of the mutation(Chen et al., 2013; Ihmels et al., 2007) making the relationship between the duplicates and their potential functionality/redundancy less clear (Dandage and Landry, 2019; Lavi, 2015).

Pseudogenisation has always been following evolutionary complexity. For instance, aquatic and semiaquatic species always had high rates of pseudogene formation due to evolving alternative ways to sense surrounding environment. For instance, platypus can sense its surroundings using a complex combination of electro- and mechanoreception (Niimura and Nei, 2007). Another highly pseudogenised gene family is olfactory receptor genes (ORs), encoding detection of odorants, which is crucial for the survival of most mammals (Niimura and Nei, 2007). Primates are particularly notable as they have a very high percentage of pseudogenes among ORs, compared to other species (Niimura and Nei, 2007). For instance, humans have ~400 functional and ~400 pseudogenised OR copies (Glusman et al., 2001; Niimura and Nei, 2003; Zozulya et al., 2001), while out of 1,400 OR genes in mice only a quarter are pseudogenes (Niimura and Nei, 2005; Young, 2002; Zhang and Firestein, 2002). And while it was originally accepted that the number of functional ORs correlates with adaptation and importance of the sense of smell (Niimura and Nei, 2005; Niimura and Nei, 2007) it is not always the case. For example, dogs are supposed to have a great sense of smell, yet they do not possess the largest number of ORs. Similarly, cows, not particularly known for their sense of smell, have around 1000 functional ORs and approximately another 1000 of OR pseudogenes. Finally, it was shown that the removal of 80% of the rat olfactory bulb glomerular layer does not cause any significant effect of the function (Shepherd, 2004). Therefore, whether its pseudogenes, functional ORs, or their combination in various amounts is responsible for the quality of the function is not fully clear. Finally, Mahmudi and colleagues had shown that both the OR family and the Zinc Finger family (the second largest family in humans, which also formed via duplication (Nowick et al., 2010) have such old pseudogenes that it is highly likely that they are functional (Mahmudi et al., 2015). Strikingly, the oldest pseudogene identified in this study was ~182 million years old and was formed at the split of the human and platypus branch (Mahmudi et al., 2015). Another expected pseudogene behaviour that not always appears to be correct has already been discussed previously and is its truncation. Gene duplicates *SRGAP2C* and *ARHGAP11B* (see Section 1.4.3) are both truncated, however, both evidently contributed to evolution of human neocortex (Charrier et al., 2012; Dennis and Eichler, 2016; Florio et al., 2015).

Collectively this shows that pseudogenes have yet to fully reveal their qualitative contribution to species-specific crucial functions, and that the number of the copies, level of expression, final pseudogene structure and level of truncation cannot be simply translated into one specific qualitative

outcome, and likely have to be investigated separately in each individual case to understand their contribution. Luckily, there are a few examples where a gene copy number affected various biological properties of an organism that could be investigated, such as a reduction in susceptibility to HIV and AIDS in humans and macaques (Degenhardt et al., 2009; Gonzalez et al., 2005), evolution of mammalian milk and lactation (Lemay et al., 2009) and existence of venomous monotremes (Whittington et al., 2008).

Based on the above, could we speculate that *NANOG* pseudogenes are collectively more important than we anticipate? It is possible that the expression of processed *NANOGP4* and *NANOGP8* has functional contributions at the siRNA level? Could the same be the case for human *OCT4,* which has several processed duplicates? Of note, not all naïve pluripotency genes that were found to have pseudogenes exhibit high expression in naïve hPSCs, as identified in this study. For instance, genes such as *DPPA5, IL6ST, KLF4, KLF17* and *KHDC1* are established naïve markers (Messmer et al., 2019) and they do have pseudogene copies, yet, the pseudogenes are expressed only at very low levels in naïve hPSCs. As *OCT4* and *NANOG* are two of the most crucial human pluripotency genes, it is plausible that they require the additional regulatory and dosage compensation 'support' their pseudogenes could potentially offer. Thus, they have highly expressed pseudogenes, whereas other factors do not.

Finally, based on the evidence above, even small mutations could cause significant change to the duplicate function. *NANOGP1,* that initially appeared as 'highly similar and conserved' remains that, but it is clear now that all the mutations it has could have significant consequences not only to its own function, but also to the function of *NANOG*, if they functionally interact.

From the early mouse *Nanog* studies, it is known that the amount of NANOG protein is tightly related to whether the embryonic stem cell is likely to differentiate or to stay pluripotent (Chambers et al., 2007). Additionally, the dynamic of *Nanog* fluctuation is different between individual colonies, where some pluripotent cells, surprisingly, stay *Nanog*-negative for several generations, while some are negative only for a short period of time and tend to fluctuate between negative/low/high states more often (Hastreiter and Schroeder, 2016). These two types of *Nanog*-negative cells were also predicted to respond to differentiation cues differently and to therefore have divergent differentiating potential (Hastreiter and Schroeder, 2016), which adds another layer of complexity to the role of *Nanog* in maintaining balance between pluripotency and differentiation. Therefore, I hypothesise that in humans, where *NANOG* expression is present for longer and in more cell types, it might require additional help to maintain its correct 'pluripotent' level. Potentially, *NANOGP1*, *NANOGP8* and/or *NANOGP4* are involved in this regulation. Possibly, this regulation is more crucial in the naïve state, and therefore, *NANOGP1* gets mostly downregulated in the primed hPSCs and in Day 14 epiblast.

Finally, it is also possible that the *NANOG*-like activity is not less important at later stages, but instead *must* get downregulated at a later developmental point to remove strong repression on genes, such as *OTX2* (Su et al., 2018) that need to become activated. Based on the variety of qualitative changes caused by pseudogenes and other duplicates, it is also entirely possible, that *NANOGP1* has developed another role we are not familiar with now and cannot predict yet. Collectively, I concluded that *NANOGP1* could in fact be a paralogue duplicate and not a pseudogene as previously suggested.

Overall, this thesis provides evidence to argue that pseudogenes should be studied in more detail and should not be overlooked as 'primarily dysfunctional'. In the case of *NANOGP1*, its ability to form a protein was disregarded mainly due to a misassumption about the start codon, which turned out to be incorrect, and NANOGP1 protein is formed in naïve hPSCs. Other potentially functional pseudogenes could have their functionality also concealed and less predictable. Therefore, what we conclude and think about pseudogenes must be constantly reviewed, since software, quality of the data and new ways of thinking will be constantly changing and improving, which will likely lead to finding more pseudogene functions, 'hidden' from the eye today.

Bibliography

Acampora, D., di Giovannantonio, L. G., Garofalo, A., Nigro, V., Omodei, D., Lombardi, A., Zhang, J., Chambers, I. and Simeone, A. (2017). Functional Antagonism between OTX2 and NANOG Specifies a Spectrum of Heterogeneous Identities in Embryonic Stem Cells. *Stem Cell Reports* **9**, 1642–1659.

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., et al. (2017). Ensembl 2017. *Nucleic Acids Research* **45**, D635–D642.

Ambady, S., Malcuit, C., Kashpur, O., Kole, D., Holmes, W. F., Hedblom, E., Page, R. L. and Dominko, T. (2010). Expression of NANOG and NANOGP8 in a variety of undifferentiated and differentiated human cells. *Int J Dev Biol* **54**, 1743–54.

Artus, J., Piliszek, A. and Hadjantonakis, A.-K. (2011). The primitive endoderm lineage of the mouse blastocyst: Sequential transcription factor activation and regulation of differentiation by Sox17. *Developmental Biology* **350**, 393–404.

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

Bailey, J. A. (2002). Recent Segmental Duplications in the Human Genome. *Science (1979)* **297**, 1003–1007.

Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C. and Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276-288.e8.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837.

Barson, G. and Griffiths, E. (2016). SeqTools: visual tools for manual analysis of sequence alignments. *BMC Research Notes* **9**, 39.

Basic Local Alignment Search Tool (BLAST) In *Bioinformatics and Functional Genomics*, pp. 100–138. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Beattie, G. M., Lopez, A. D., Bucay, N., Hinton, A., Firpo, M. T., King, C. C. and Hayek, A. (2005). Activin A Maintains Pluripotency of Human Embryonic Stem Cells in the Absence of Feeder Layers. *Stem Cells* **23**, 489–495.

Bedzhov, I. and Zernicka-Goetz, M. (2014). Self-Organizing Properties of Mouse Pluripotent Cells Initiate Morphogenesis upon Implantation. *Cell* **156**, 1032–1044.

Bendall, S. C., Stewart, M. H., Menendez, P., George, D., Vijayaragavan, K., Werbowetski-Ogilvie, T., Ramos-Mejia, V., Rouleau, A., Yang, J., Bossé, M., et al. (2007). IGF and FGF cooperatively establish the regulatory stem cell niche of pluripotent human cells in vitro. *Nature* **448**, 1015–1021.

Bertero, A., Madrigal, P., Galli, A., Hubner, N. C., Moreno, I., Burks, D., Brown, S., Pedersen, R. A., Gaffney, D., Mendjan, S., et al. (2015). Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. *Genes & Development* **29**, 702–717.

Betschinger, J., Nichols, J., Dietmann, S., Corrin, P. D., Paddison, P. J. and Smith, A. (2013). Exit from Pluripotency Is Gated by Intracellular Redistribution of the bHLH Transcription Factor Tfe3. *Cell* **153**, 335–347.

Bhan, A., Galas, D. J. and Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486–1493.

Blake, R. D., Hess, S. T. and Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution* **34**, 189–200.

Blakeley, P., Fogarty, N. M. E., del Valle, I., Wamaitha, S. E., Hu, T. X., Elder, K., Snell, P., Christie, L., Robson, P. and Niakan, K. K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., Oakley, D. H., et al. (2011). Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. *Cell* **144**, 439–452.

Booth, H. A. F. and Holland, P. W. H. (2004). Eleven daughters of NANOG. *Genomics* **84**, 229–38.

Boroviak, T. and Nichols, J. (2017). Primate embryogenesis predicts the hallmarks of human naïve pluripotency. *Development* **144**, 175–186.

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**, 366–382.

Boroviak, T., Stirparo, G. G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J. and Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development* **145**, dev167833.

Bouckenheimer, J., Fauque, P., Lecellier, C.-H., Bruno, C., Commes, T., Lemaître, J.-M., de Vos, J. and Assou, S. (2018). Differential long non-coding RNA expression profiles in human oocytes and cumulus cells. *Scientific Reports* **8**, 2202.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* **122**, 947–956.

Bredenkamp, N., Stirparo, G. G., Nichols, J., Smith, A. and Guo, G. (2019a). The Cell-Surface Marker Sushi Containing Domain 2 Facilitates Establishment of Human Naive Pluripotent Stem Cells. *Stem Cell Reports* **12**, 1212–1222.

Bredenkamp, N., Yang, J., Clarke, J., Stirparo, G. G., von Meyenn, F., Dietmann, S., Baker, D., Drummond, R., Ren, Y., Li, D., et al. (2019b). Wnt Inhibition Facilitates RNA-Mediated Reprogramming of Human Somatic Cells to Naive Pluripotency. *Stem Cell Reports* **13**, 1083–1098.

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

Brons, I. G. M., Smithers, L. E., Trotter, M. W. B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S. M., Howlett, S. K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–5.

Brookfield, J. F. Y. (1997). Genetic Redundancy.pp. 137–155.

Brumbaugh, J., Russell, J. D., Yu, P., Westphall, M. S., Coon, J. J. and Thomson, J. A. (2014). NANOG Is Multiply Phosphorylated and Directly Modified by ERK2 and CDK1 In Vitro. *Stem Cell Reports* **2**, 18–25.

Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T. and Wysocka, J. (2014). Reorganization of Enhancer Patterns in Transition from Naive to Primed Pluripotency. *Cell Stem Cell* **14**, 838–853.

Camp, E., Sánchez-Sánchez, A. v., García-España, A., DeSalle, R., Odqvist, L., Enrique O'Connor, J. and Mullor, J. L. (2009). Nanog Regulates Proliferation During Early Fish Development. *STEM CELLS* **27**, 2081–2091.

Cañón, S., Herranz, C. and Manzanares, M. (2006). Germ cell restricted expression of chick Nanog. *Developmental Dynamics* **235**, 2889–2894.

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**, 119–126.

Carter, M. G., Smagghe, B. J., Stewart, A. K., Rapley, J. A., Lynch, E., Bernier, K. J., Keating, K. W., Hatziioannou, V. M., Hartman, E. J. and Bamdad, C. C. (2016). A Primitive Growth Factor, NME7 $_{AB}$, Is Sufficient to Induce Stable Naïve State Human Pluripotency; Reprogramming in This Novel Growth Factor Confers Superior Differentiation. *STEM CELLS* **34**, 847–859.

Cauffman, G., Liebaers, I., van Steirteghem, A. and van de Velde, H. (2006). POU5F1 Isoforms Show Different Expression Patterns in Human Embryonic Stem Cells and Preimplantation Embryos. *Stem Cells* **24**, 2685–2691.

Cauffman, G., de Rycke, M., Sermon, K., Liebaers, I. and van de Velde, H. (2009). Markers that define stemness in ESC are unable to identify the totipotent cells in human preimplantation embryos. *Human Reproduction* **24**, 63–70.

Cavaliere, G. (2017). A 14-day limit for bioethics: the debate over human embryo research. *BMC Medical Ethics* **18**, 38.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S. and Smith, A. (2003). Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. *Cell* **113**, 643–655.

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

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**, 663–675.

Chang, C. P., Brocchieri, L., Shen, W. F., Largman, C. and Cleary, M. L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Molecular and Cellular Biology* **16**, 1734–1745.

Chang, D. F., Tsai, S. C., Wang, X. C., Xia, P., Senadheera, D. and Lutzko, C. (2009). Molecular Characterization of the Human NANOG Protein. *STEM CELLS* **27**, 812–821.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J. E., Lambert, N., de Marchena, J., Jin, W. L., Vanderhaeghen, P., Ghosh, A., Sassa, T., et al. (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923–935.

Chazaud, C., Yamanaka, Y., Pawson, T. and Rossant, J. (2006). Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway. *Developmental Cell* **10**, 615–624.

Chen, A. E., Egli, D., Niakan, K., Deng, J., Akutsu, H., Yamaki, M., Cowan, C., Fitz-Gerald, C., Zhang, K., Melton, D. A., et al. (2009). Optimal Timing of Inner Cell Mass Isolation Increases the Efficiency of Human Embryonic Stem Cell Derivation and Allows Generation of Sibling Cell Lines. *Cell Stem Cell* **4**, 103–106.

Chen, W.-H., Zhao, X.-M., van Noort, V. and Bork, P. (2013). Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. *PLoS Computational Biology* **9**, e1003073.

Chen, Y., Cao, J., Xiong, M., Petersen, A. J., Dong, Y., Tao, Y., Huang, C. T.-L., Du, Z. and Zhang, S.-C. (2015a). Engineering Human Stem Cell Lines with Inducible Gene Knockout using CRISPR/Cas9. *Cell Stem Cell* **17**, 233–244.

Chen, H., Aksoy, I., Gonnot, F., Osteil, P., Aubry, M., Hamela, C., Rognard, C., Hochard, A., Voisin, S., Fontaine, E., et al. (2015b). Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nature Communications* **6**, 7095.

Cheng, L., Albers, P., Berney, D. M., Feldman, D. R., Daugaard, G., Gilligan, T. and Looijenga, L. H. J. (2018). Testicular cancer. *Nature Reviews Disease Primers* **4**, 29.

Chia, N.-Y., Chan, Y.-S., Feng, B., Lu, X., Orlov, Y. L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.-S., et al. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316–320.

Chovanec, P., Collier, A. J., Krueger, C., Várnai, C., Semprich, C. I., Schoenfelder, S., Corcoran, A. E. and Rugg-Gunn, P. J. (2021). Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nature Communications* **12**, 2098.

Cinkornpumin, J. K., Kwon, S. Y., Guo, Y., Hossain, I., Sirois, J., Russett, C. S., Tseng, H.-W., Okae, H., Arima, T., Duchaine, T. F., et al. (2020). Naive Human Embryonic Stem Cells Can Give Rise to Cells with a Trophoblast-like Transcriptome and Methylome. *Stem Cell Reports* **15**, 198–213.

Clark, A. T., Rodriguez, R. T., Bodnar, M. S., Abeyta, M. J., Cedars, M. I., Turek, P. J., Firpo, M. T. and Reijo Pera, R. A. (2004). Human *STELLAR*, *NANOG*, and *GDF3* Genes Are Expressed in Pluripotent Cells and Map to Chromosome 12p13, a Hotspot for Teratocarcinoma. *STEM CELLS* **22**, 169–179.

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E., et al. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology* **37**, 224–226.

Codner, G. F., Mianné, J., Caulder, A., Loeffler, J., Fell, R., King, R., Allan, A. J., Mackenzie, M., Pike, F. J., McCabe, C. v., et al. (2018). Application of long single-stranded DNA donors in genome editing: generation and validation of mouse mutants. *BMC Biology* **16**, 70.

Collier, A. (2019). CHARACTERISING THE REPROGRAMMING DYNAMICS BETWEEN HUMAN PLURIPOTENT STATES (Doctoral thesis).

Collier, A. J. and Rugg-Gunn, P. J. (2018). Identifying Human Naïve Pluripotent Stem Cells − Evaluating State-Specific Reporter Lines and Cell-Surface Markers. *BioEssays* **40**, e1700239.

Collier, A. J., Panula, S. P., Schell, J. P., Chovanec, P., Plaza Reyes, A., Petropoulos, S., Corcoran, A. E., Walker, R., Douagi, I., Lanner, F., et al. (2017). Comprehensive Cell Surface Protein Profiling Identifies Specific Markers of Human Naive and Primed Pluripotent States. *Cell Stem Cell* **20**, 874-890.e7.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (1979)* **339**, 819–823.

Culty, M. (2009). Gonocytes, the forgotten cells of the germ cell lineage. *Birth Defects Research Part C: Embryo Today: Reviews* **87**, 1–26.

Dahéron, L., Opitz, S. L., Zaehres, H., Lensch, W. M., Andrews, P. W., Itskovitz-Eldor, J. and Daley, G. Q. (2004). LIF/STAT3 Signaling Fails to Maintain Self-Renewal of Human Embryonic Stem Cells. *Stem Cells* **22**, 770–778.

Dandage, R. and Landry, C. R. (2019). Paralog dependency indirectly affects the robustness of human cells. *Molecular Systems Biology* **15**, e8871.

Darr, H., Mayshar, Y. and Benvenisty, N. (2006). Overexpression of NANOG in human ES cells enables feeder-free growth while inducing primitive ectoderm features. *Development* **133**, 1193–201.

de Felici, M. (2013). Origin, Migration, and Proliferation of Human Primordial Germ Cells. In *Oogenesis*, pp. 19–37. London: Springer London.

de Jong, B., Oosterhuis, J. W., Castedo, S. M. M. J., Vos, A. and te Meerman, G. J. (1990). Pathogenesis of adult testicular germ cell tumors. *Cancer Genetics and Cytogenetics* **48**, 143–167.

de Paepe, C., Cauffman, G., Verloes, A., Sterckx, J., Devroey, P., Tournaye, H., Liebaers, I. and van de Velde, H. (2013). Human trophectoderm cells are not yet committed. *Human Reproduction* **28**, 740–749.

Degenhardt, J. D., de Candia, P., Chabot, A., Schwartz, S., Henderson, L., Ling, B., Hunter, M., Jiang, Z., Palermo, R. E., Katze, M., et al. (2009). Copy Number Variation of CCL3-like Genes

Affects Rate of Progression to Simian-AIDS in Rhesus Macaques (Macaca mulatta). *PLoS Genetics* **5**, e1000346.

Deglincerti, A., Croft, G. F., Pietila, L. N., Zernicka-Goetz, M., Siggia, E. D. and Brivanlou, A. H. (2016a). Self-organization of the in vitro attached human embryo. *Nature* **533**, 251–254.

Deglincerti, A., Etoc, F., Guerra, M. C., Martyn, I., Metzger, J., Ruzo, A., Simunovic, M., Yoney, A., Brivanlou, A. H., Siggia, E., et al. (2016b). Self-organization of human embryonic stem cells on micropatterns. *Nature Protocols* **11**, 2223–2232.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J. and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607.

Dennis, M. Y. and Eichler, E. E. (2016). Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics & Development* **41**, 44–52.

Devor, E. J. (2006). Primate MicroRNAs miR-220 and miR-492 Lie within Processed Pseudogenes. *Journal of Heredity* **97**, 186–190.

Diecke, S., Quiroga-Negreira, A., Redmer, T. and Besser, D. (2008). FGF2 Signaling in Mouse Embryonic Fibroblasts Is Crucial for Self-Renewal of Embryonic Stem Cells. *Cells Tissues Organs* **188**, 52–61.

Dietrich, J.-E. and Hiiragi, T. (2007). Stochastic patterning in the mouse pre-implantation embryo. *Development* **134**, 4219–4231.

Diss, G., Ascencio, D., DeLuna, A. and Landry, C. R. (2014). Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **322**, 488–499.

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

Do, H.-J., Lee, W.-Y., Lim, H.-Y., Oh, J.-H., Kim, D.-K., Kim, J.-H., Kim, T. and Kim, J.-H. (2009). Two potent transactivation domains in the C-terminal region of human NANOG mediate transcriptional activation in human embryonic carcinoma cells. *Journal of Cellular Biochemistry* **106**, 1079–1089.

Dobson, A. T., Raja, R., Abeyta, M. J., Taylor, T., Shen, S., Haqq, C. and Pera, R. A. R. (2004). The unique transcriptome through day 3 of human preimplantation development. *Human Molecular Genetics* **13**, 1461–1470.

Dong, C., Beltcheva, M., Gontarz, P., Zhang, B., Popli, P., Fischer, L. A., Khan, S. A., Park, K., Yoon, E.-J., Xing, X., et al. (2020). Derivation of trophoblast stem cells from naïve human pluripotent stem cells. *Elife* **9**, e52504.

Donnison, M., Beaton, A., Davey, H. W., Broadhurst, R., L'Huillier, P. and Pfeffer, P. L. (2005). Loss of the extraembryonic ectoderm in *Elf5* mutants leads to defects in embryonic patterning. *Development* **132**, 2299–2308.

Doss, M. X. and Sachinidis, A. (2019). Current Challenges of iPSC-Based Disease Modeling and Therapeutic Implications. *Cells* **8**, 403.

Du, Y., Xie, W., Zhang, F. and Liu, C. (2019). Chimeric Mouse Generation by ES Cell Blastocyst Microinjection and Uterine Transfer.pp. 99–114.

Duggal, G., Warrier, S., Ghimire, S., Broekaert, D., van der Jeught, M., Lierman, S., Deroo, T., Peelman, L., van Soom, A., Cornelissen, R., et al. (2015). Alternative Routes to Induce Naïve Pluripotency in Human Embryonic Stem Cells. *STEM CELLS* **33**, 2686–2698.

Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S. and Smith, A. G. (2014). Defining an essential transcription factor program for naïve pluripotency. *Science (1979)* **344**, 1156–1160.

Dunwell, T. L. and Holland, P. W. H. (2017). A sister of NANOG regulates genes expressed in pre-implantation human development. *Open Biology* **7**, 170027.

Dutta, D., Ray, S., Home, P., Larson, M., Wolfe, M. W. and Paul, S. (2011). Self-Renewal Versus Lineage Commitment of Embryonic Stem Cells: Protein Kinase C Signaling Shifts the Balance. *STEM CELLS* **29**, 618–628.

Eberle, I., Pless, B., Braun, M., Dingermann, T. and Marschalek, R. (2010). Transcriptional properties of human NANOG1 and NANOG2 in acute leukemic cells. *Nucleic Acids Research* **38**, 5384–5395.

Edson, M. A., Nagaraja, A. K. and Matzuk, M. M. (2009). The Mammalian Ovary from Genesis to Revelation. *Endocrine Reviews* **30**, 624–712.

Edwards, R. G., Steptoe, P. C. and Purdy, J. M. (1970). Fertilization and Cleavage in vitro of Preovulator Human Oocytes. *Nature* **227**, 1307–1309.

Edwards, R. G., Steptoe, P. C. and Purdy, J. M. (1980). Establishing full-term human pregnancies using cleaving embryos grown in vitro*. *BJOG: An International Journal of Obstetrics and Gynaecology* **87**, 737–756.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D. J., Dash, C., Bazett-Jones, D. P., le Grice, S., McKay, R. D. G., Buetow, K. H., et al. (2008). Global Transcription in Pluripotent Embryonic Stem Cells. *Cell Stem Cell* **2**, 437–447.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216.

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

Fairbanks, D. J. and Maughan, P. J. (2006). Evolution of the NANOG pseudogene family in the human and chimpanzee genomes. *BMC Evolutionary Biology* **6**, 12.

Fan, Y., Ma, Z.-L., Zhong, K., Zhang, P.-Y., Kang, X.-J., Zhang, Y.-Y., Zhu, H.-Y., Qiao, J., Li, M. and Yu, Y. (2021). Generation of human blastocyst-like structures from pluripotent stem cells. *Preprint, BioRxiv*.

Fares, M. A. (2014). The evolution of protein moonlighting: adaptive traps and promiscuity in the chaperonins. *Biochemical Society Transactions* **42**, 1709–1714.

Faulkner-Jones, A., Fyfe, C., Cornelissen, D.-J., Gardner, J., King, J., Courtney, A. and Shu, W. (2015). Bioprinting of human pluripotent stem cells and their directed differentiation into hepatocyte-like cells for the generation of mini-livers in 3D. *Biofabrication* **7**, 044102.

Felix, W. (1911). Die entwicklung der harn-und geschlechtsorgane. *Hirzel*.

Finn, C. A. and McLaren, A. (1967). A study of the early stages of implantation in mice. *Reproduction* **13**, 259–267.

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F. K., Peters, J., et al. (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science (1979)* **347**, 1465–1470.

Fogarty, N. M. E., McCarthy, A., Snijders, K. E., Powell, B. E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaitha, S. E., Kim, D., et al. (2017). Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* **550**, 67–73.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–45.

French, D. B., Sabanegh, E. S., Goldfarb, J. and Desai, N. (2010). Does severe teratozoospermia affect blastocyst formation, live birth rate, and other clinical outcome parameters in ICSI cycles? *Fertility and Sterility* **93**, 1097–1103.

Frum, T., Halbisen, M. A., Wang, C., Amiri, H., Robson, P. and Ralston, A. (2013). Oct4 Cell-Autonomously Promotes Primitive Endoderm Development in the Mouse Blastocyst. *Developmental Cell* **25**, 610–622.

Fuss, A. (1911). Uber extraregionare Geschlechtszellen bei einem menschlichen Embryo von 4 Wochen. *Anat Am* **39**, 407–409.

Fuss, A. (1912). Über die Geschlechtszellen des Menschen und der Säugetiere. *Archiv für mikroskopische Anatomie* **81.1**, 1–23.

Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 710.

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

Gagnon, J. A., Obbad, K. and Schier, A. F. (2017). Zebrafish *nanog* is primarily required in extraembryonic tissue. *Development* dev.147793.

Galonska, C., Ziller, M. J., Karnik, R. and Meissner, A. (2015). Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming. *Cell Stem Cell* **17**, 462–470.

Gardner, R. L. (1978). The Relationship Between Cell Lineage and Differentiation in the Early Mouse Embryo.pp. 205–241.

Gardner, D. K., Lane, M. and Schoolcraft, W. B. (2000). Culture and transfer of viable blastocysts: a feasible proposition for human IVF. *Hum Reprod* **15 Suppl 6**, 9–23.

Gehring, W. J., Affolter, M. and Bürglin, T. (1994). Homeodomain proteins. *Annual Review of Biochemistry* **63**, 487–526.

Gerami-Naini, B., Dovzhenko, O. v., Durning, M., Wegner, F. H., Thomson, J. A. and Golos, T. G. (2004). Trophoblast Differentiation in Embryoid Bodies Derived from Human Embryonic Stem Cells. *Endocrinology* **145**, 17–24.

Gerri, C., McCarthy, A., Alanis-Lobato, G., Demtschenko, A., Bruneau, A., Loubersac, S., Fogarty, N. M. E., Hampshire, D., Elder, K., Snell, P., et al. (2020). Initiation of a conserved trophectoderm program in human, cow and mouse embryos. *Nature* **587**, 443–447.

Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., Cooper, G. M., Reddy, T. E., Crawford, G. E. and Myers, R. M. (2013). Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Molecular Cell* **52**, 25–36.

Ghimire, S., van der Jeught, M., Neupane, J., Roost, M. S., Anckaert, J., Popovic, M., van Nieuwerburgh, F., Mestdagh, P., Vandesompele, J., Deforce, D., et al. (2018). Comparative analysis of naive, primed and ground state pluripotency in mouse embryonic stem cells originating from the same genetic background. *Scientific Reports* **8**, 5884.

Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., et al. (2013). Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. *Cell* **153**, 1149–1163.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451.

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

Ginsburg, M., Snow, M. H. and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. *Development* **110**, 521–528.

Gkountela, S., Zhang, K. X., Shafiq, T. A., Liao, W.-W., Hargan-Calvopiña, J., Chen, P.-Y. and Clark, A. T. (2015). DNA Demethylation Dynamics in the Human Prenatal Germline. *Cell* **161**, 1425–1436.

Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001). The Complete Human Olfactory Subgenome. *Genome Research* **11**, 685–702.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., et al. (2005). The Influence of *CCL3L1* Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science (1979)* **307**, 1434–1440.

Gould, G. W. and Holman, G. D. (1993). The glucose transporter family: structure, function and tissue-specific expression. *Biochemical Journal* **295**, 329–341.

Grabarek, J. B., Żyżyńska, K., Saiz, N., Piliszek, A., Frankenberg, S., Nichols, J., Hadjantonakis, A.-K. and Plusa, B. (2012). Differential plasticity of epiblast and primitive endoderm precursors within the ICM of the early mouse embryo. *Development* **139**, 129–139.

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., Martin, L., Ware, C. B., Blish, C. A., Chang, H. Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225.

Gruschus, J. M., Tsao, D. H. H., Wang, L.-H., Nirenberg, M. and Ferretti, J. A. (1997). Interactions of the vnd/NK-2 Homeodomain with DNA by Nuclear Magnetic Resonance Spectroscopy: Basis of Binding Specificity. *Biochemistry* **36**, 5372–5380.

Guhr, A., Kobold, S., Seltmann, S., Seiler Wulczyn, A. E. M., Kurtz, A. and Löser, P. (2018). Recent Trends in Research with Human Pluripotent Stem Cells: Impact of Research and Use of Cell Lines in Experimental Research and Clinical Trials. *Stem Cell Reports* **11**, 485–496.

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., et al. (2014). The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610.

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**, 437–446.

Guo, G., von Meyenn, F., Rostovskaya, M., Clarke, J., Dietmann, S., Baker, D., Sahakyan, A., Myers, S., Bertone, P., Reik, W., et al. (2017). Epigenetic resetting of human pluripotency. *Development* **144**, 2748–2763.

Guo, Q., Mintier, G., Ma-Edmonds, M., Storton, D., Wang, X., Xiao, X., Kienzle, B., Zhao, D. and Feder, J. N. (2018). 'Cold shock' increases the frequency of homology directed repair gene editing in induced pluripotent stem cells. *Scientific Reports* **8**, 2080.

Guo, G., Stirparo, G. G., Strawbridge, S. E., Spindlow, D., Yang, J., Clarke, J., Dattani, A., Yanagida, A., Li, M. A., Myers, S., et al. (2021). Human naive epiblast cells possess unrestricted lineage potential. *Cell Stem Cell* **28**, 1040–1056.

Hackett, J. A., Dietmann, S., Murakami, K., Down, T. A., Leitch, H. G. and Surani, M. A. (2013). Synergistic Mechanisms of DNA Demethylation during Transition to Ground-State Pluripotency. *Stem Cell Reports* **1**, 518–531.

Hahn, M. W., Demuth, J. P. and Han, S.-G. (2007). Accelerated Rate of Gene Gain and Loss in Primates. *Genetics* **177**, 1941–1949.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J. and Surani, M. A. (2002). Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of Development* **117**, 15–23.

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**, 9222–9227.

Hart, A. H., Hartley, L., Ibrahim, M. and Robb, L. (2004). Identification, cloning and expression analysis of the pluripotency promoting Nanog genes in mouse and human. *Dev Dyn* **230**, 187–98.

Hartley, J. L. (2000). DNA Cloning Using In Vitro Site-Specific Recombination. *Genome Research* **10**, 1788–1795.

Hartley, J. L. (2002). Use of the Gateway System for Protein Expression in Multiple Hosts. *Current Protocols in Protein Science* **30**.

Harvey, A. J., Armant, D. R., Bavister, B. D., Nichols, S. M. and Brenner, C. A. (2009). Inner Cell Mass Localization of NANOG Precedes OCT3/4 in Rhesus Monkey Blastocysts. *Stem Cells and Development* **18**, 1451–1459.

Hastreiter, S. and Schroeder, T. (2016). Nanog dynamics in single embryonic stem cells. *Cell Cycle* **15**, 770–771.

Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. and Saitou, M. (2011). Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–32.

Hayashi, Y., Caboni, L., Das, D., Yumoto, F., Clayton, T., Deller, M. C., Nguyen, P., Farr, C. L., Chiu, H.-J., Miller, M. D., et al. (2015). Structure-based discovery of NANOG variant with enhanced properties to promote self-renewal and reprogramming of pluripotent stem cells. *Proceedings of the National Academy of Sciences* **112**, 4666–4671.

Hemberger, M., Udayashankar, R., Tesar, P., Moore, H. and Burton, G. J. (2010). ELF5-enforced transcriptional networks define an epigenetically regulated trophoblast stem cell compartment in the human placenta. *Human Molecular Genetics* **19**, 2456–2467.

Hertig, A. T., Rock, J., Adams, E. C. and Mulligan, W. J. (1954). *On the preimplantation stages of the human ovum: a description of four normal and four abnormal specimens ranging from the second to the fifth day of development*.

Hertig, A. T., Rock, J. and Adams, E. C. (1956). A description of 34 human ova within the first 17 days of development. *American Journal of Anatomy* **98**, 435–493.

Hertig, A. T., Rock, J., Adams, E. C. and Menkin, M. C. (1959). Thirty-four fertilized human ova, good, bad and indifferent, recovered from 210 women of known fertility: a study of biologic wastage in early human pregnancy. *Pediatrics* **23**, 202–211.

Heuser, C. F. and Streeter, G. I. (1941a). *Contributions to Embryology.*

Heuser, C. H. and Streeter, G. L. (1941b). Development of the macaque embryo. *Contributions to Embryology* **29**, 15–55.

Hilscher, B., Hilscher, W., Bulthoff-Ohnolz, B., Kramer, U., Birke, A., Pelzer, H. and Gauss, G. (1974). Kinetics of gametogenesis. *Cell and Tissue Research* **154**, 443–470.

Ho, B., Olson, G., Figel, S., Gelman, I., Cance, W. G. and Golubovskaya, V. M. (2012). Nanog Increases Focal Adhesion Kinase (FAK) Promoter Activity and Expression and Directly Binds to FAK Protein to Be Phosphorylated. *Journal of Biological Chemistry* **287**, 18656–18673.

Hotta, A., Cheung, A. Y. L., Farra, N., Garcha, K., Chang, W. Y., Pasceri, P., Stanford, W. L. and Ellis, J. (2009). EOS lentiviral vector selection system for human induced pluripotent stem cells. *Nature Protocols* **4**, 1828–1844.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* **31**, 827–832.

Hu, B.-Y., Weick, J. P., Yu, J., Ma, L.-X., Zhang, X.-Q., Thomson, J. A. and Zhang, S.-C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proceedings of the National Academy of Sciences* **107**, 4335–4340.

Huang, S.-M. A., Mishina, Y. M., Liu, S., Cheung, A., Stegmeier, F., Michaud, G. A., Charlat, O., Wiellette, E., Zhang, Y., Wiessner, S., et al. (2009). Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* **461**, 614–620.

Huang, Y., Osorno, R., Tsakiridis, A. and Wilson, V. (2012). In Vivo Differentiation Potential of Epiblast Stem Cells Revealed by Chimeric Embryo Formation. *Cell Reports* **2**, 1571–1578.

Huang, K., Maruyama, T. and Fan, G. (2014). The Naive State of Human Pluripotent Stem Cells: A Synthesis of Stem Cell and Preimplantation Embryo Transcriptome Analyses. *Cell Stem Cell* **15**, 410–415.

Humphrey, R. K. (2004). Maintenance of Pluripotency in Human Embryonic Stem Cells Is STAT3 Independent. *Stem Cells* **22**, 522–530.

Hyslop, L., Stojkovic, M., Armstrong, L., Walter, T., Stojkovic, P., Przyborski, S., Herbert, M., Murdoch, A., Strachan, T. and Lako, M. (2005). Downregulation of NANOG Induces Differentiation of Human Embryonic Stem Cells to Extraembryonic Lineages. *Stem Cells* **23**, 1035–1043.

Ihmels, J., Collins, S. R., Schuldiner, M., Krogan, N. J. and Weissman, J. S. (2007). Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Molecular Systems Biology* **3**, 86.

ISCI (2011). Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nature Biotechnology* **29**, 1132–1144.

James, D., Levine, A. J., Besser, D. and Hemmati-Brivanlou, A. (2005). TGFβ/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* **132**, 1273–1282.

James, D., Noggle, S. A., Swigut, T. and Brivanlou, A. H. (2006). Contribution of human embryonic stem cells to mouse blastocysts. *Developmental Biology* **295**, 90–102.

Jauch, R., Ng, C. K. L., Saikatendu, K. S., Stevens, R. C. and Kolatkar, P. R. (2008). Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog. *Journal of Molecular Biology* **376**, 758–770.

Jez, M., Ambady, S., Kashpur, O., Grella, A., Malcuit, C., Vilner, L., Rozman, P. and Dominko, T. (2014). Expression and Differentiation between OCT4A and Its Pseudogenes in Human ESCs and Differentiated Adult Somatic Cells. *PLoS ONE* **9**, e89546.

Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., Weintraub, A. S., Hnisz, D., Pegoraro, G., Lee, T. I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, 262–275.

Johnsen, D. O., Johnson, D. K. and Whitney, R. A. (2012). *Nonhuman Primates in Biomedical Research*.

Johnson, R., Teh, C. H., Kunarso, G., Wong, K. Y., Srinivasan, G., Cooper, M. L., Volta, M., Chan, S. S., Lipovich, L., Pollard, S. M., et al. (2008). REST Regulates Distinct Transcriptional Networks in Embryonic and Neural Stem Cells. *PLoS Biology* **6**, e256.

Kagawa, H., Javali, A., Khoei, H. H., Sommer, T. M., Sestini, G., Novatchkova, M., Scholte op Reimer, Y., Castel, G., Bruneau, A., Maenhoudt, N., et al. (2021). Human blastoids model blastocyst development and implantation. *Nature* **601**, 600–605.

Kalkan, T., Olova, N., Roode, M., Mulas, C., Lee, H. J., Nett, I., Marks, H., Walker, R., Stunnenberg, H. G., Lilley, K. S., et al. (2017). Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221–1234.

Kaufman, M. H. (1992). *Atlas of mouse development*. Academic pres.

Kearns, N. A., Genga, R. M. J., Enuameh, M. S., Garber, M., Wolfe, S. A. and Maehr, R. (2014). Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development* **141**, 219–223.

Kerr, C. L., Hill, C. M., Blumenthal, P. D. and Gearhart, J. D. (2008). Expression of pluripotent stem cell markers in the human fetal ovary. *Human Reproduction* **23**, 589–599.

Khan, S. A., Park, K., Fischer, L. A., Dong, C., Lungjangwa, T., Jimenez, M., Casalena, D., Chew, B., Dietmann, S., Auld, D. S., et al. (2021). Probing the signaling requirements for naive human pluripotency by high-throughput chemical screening. *Cell Reports* **35**, 109233.

Khong, T. Y. and Robertson, W. B. (1987). Placenta creta and placenta praevia creta. *Placenta* **8**, 399–409.

Kidder, B. L., Hu, G. and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology* **12**, 918–922.

Kim, D., Langmead, B. and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360.

Kim, D. K., Seo, E. J., Choi, E. J., Lee, S. I., Kwon, Y. W., Jang, I. H., Kim, S.-C., Kim, K.-H., Suh, D.-S., Seong-Jang, K., et al. (2016). Crucial role of HMGA1 in the self-renewal and drug resistance of ovarian cancer stem cells. *Experimental & Molecular Medicine* **48**, e255–e255.

Kimber, S. J., Sneddon, S. F., Bloor, D. J., El-Bareg, A. M., Hawkhead, J. A., Metcalfe, A. D., Houghton, F. D., Leese, H. J., Rutherford, A., Lieberman, B. A., et al. (2008). Expression of genes involved in early cell fate decisions in human embryos and their regulation by growth factors. *REPRODUCTION* **135**, 635–647.

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., et al. (2014). The Transcriptional and Functional Properties of Mouse Epiblast Stem Cells Resemble the Anterior Primitive Streak. *Cell Stem Cell* **14**, 107–120.

Kondrashov, F. A. and Kondrashov, A. S. (2006). Role of selection in fixation of gene duplications. *Journal of Theoretical Biology* **239**, 141–151.

Kossack, N., Terwort, N., Wistuba, J., Ehmcke, J., Schlatt, S., Schöler, H., Kliesch, S. and Gromoll, J. (2013). A combined approach facilitates the reliable detection of human spermatogonia in vitro. *Human Reproduction* **28**, 3012–3025.

Kraehenbuehl, T. P., Langer, R. and Ferreira, L. S. (2011). Three-dimensional biomaterials for the study of human pluripotent stem cells. *Nature Methods* **8**, 731–736.

Kuijk, E. W., de Gier, J., Chuva de Sousa Lopes, S. M., Chambers, I., van Pelt, A. M. M., Colenbrander, B. and Roelen, B. A. J. (2010). A Distinct Expression Pattern in Mammalian Testes Indicates a Conserved Role for NANOG in Spermatogenesis. *PLoS ONE* **5**, e10987.

Kuijk, E. W., van Tol, L. T. A., van de Velde, H., Wubbolts, R., Welling, M., Geijsen, N. and Roelen, B. A. J. (2012). The roles of FGF and MAP kinase signaling in the segregation of the epiblast and hypoblast cell lineages in bovine and human embryos. *Development* **139**, 871–882.

Kurilo, L. F. (1981). Oogenesis in antenatal development in man. *Human Genetics* **57**, 86–92.

Kurosawa, H. (2007). Methods for inducing embryoid body formation: in vitro differentiation system of embryonic stem cells. *Journal of Bioscience and Bioengineering* **103**, 389–398.

Lavi, O. (2015). Redundancy: A Critical Obstacle to Improving Cancer Therapy. *Cancer Research* **75**, 808–812.

Lavial, F., Acloque, H., Bertocchini, F., MacLeod, D. J., Boast, S., Bachelard, E., Montillet, G., Thenot, S., Sang, H. M., Stern, C. D., et al. (2007). The Oct4 homologue PouV and Nanog regulate pluripotency in chicken embryonic stem cells. *Development* **134**, 3549–3563.

Lawson, K. A., Meneses, J. J. and Pedersen, R. A. (1991). Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891–911.

le Bin, G. C., Muñoz-Descalzo, S., Kurowski, A., Leitch, H., Lou, X., Mansfield, W., Etienne-Dumeau, C., Grabole, N., Mulas, C., Niwa, H., et al. (2014). Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst. *Development* **141**, 1001–1010.

Lea, R. A., McCarthy, A., Boeing, S., Fallesen, T., Elder, K., Snell, P., Christie, L., Adkins, S., Shaikly, V., Taranissi, M., et al. (2021). KLF17 promotes human naïve pluripotency but is not required for its establishment. *Development* **148**, dev199378.

Lee, L. H., Peerani, R., Ungrin, M., Joshi, C., Kumacheva, E. and Zandstra, PeterW. (2009). Micropatterning of human embryonic stem cells dissects the mesoderm and endoderm lineages. *Stem Cell Research* **2**, 155–162.

Lee, J.-H., Laronde, S., Collins, T. J., Shapovalova, Z., Tanasijevic, B., McNicol, J. D., Fiebig-Comyn, A., Benoit, Y. D., Lee, J. B., Mitchell, R. R., et al. (2017). Lineage-Specific Differentiation Is Influenced by State of Human Pluripotency. *Cell Reports* **19**, 20–35.

Leeton, J., Trounson, A., Jessup, D. and Wood, C. (1982). The technique for human embryo transfer. *Fertility and Sterility* **38**, 156–161.

Lei, Y. and Schaffer, D. (2013). A fully defined and scalable 3D culture system for human pluripotent stem cell expansion and differentiation. *Proceedings of the National Academy of Sciences* **110**, E5039-5048.

Leitch, H. G., Tang, W. W. C. and Surani, M. A. (2013). Primordial Germ-Cell Development and Epigenetic Reprogramming in Mammals. *Current Topics in Developmental Biology* **104**, 149–187.

Lemay, D. G., Lynn, D. J., Martin, W. F., Neville, M. C., Casey, T. M., Rincon, G., Kriventseva, E. v, Barris, W. C., Hinrichs, A. S., Molenaar, A. J., et al. (2009). The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biology* **10**, R43.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.

Li, J., Wang, G., Wang, C., Zhao, Y., Zhang, H., Tan, Z., Song, Z., Ding, M. and Deng, H. (2007). MEK/ERK signaling contributes to the maintenance of human embryonic stem cell self-renewal. *Differentiation* **75**, 299–307.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Li, L., Dong, J., Yan, L., Yong, J., Liu, X., Hu, Y., Fan, X., Wu, X., Guo, H., Wang, X., et al. (2017). Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858–873.

Lie, K.-H., Tuch, B. E. and Sidhu, K. S. (2012). Suppression of NANOG Induces Efficient Differentiation of Human Embryonic Stem Cells to Pancreatic Endoderm. *Pancreas* **41**, 54–64.

Lin, T. and Lin, Y. (2017). p53 switches off pluripotency on differentiation. *Stem Cell Research & Therapy* **8**, 44.

Linneberg-Agerholm, M., Wong, Y. F., Romero Herrera, J. A., Monteiro, R. S., Anderson, K. G. v. and Brickman, J. M. (2019). Naïve human pluripotent stem cells respond to Wnt, Nodal and LIF signalling to produce expandable naïve extra-embryonic endoderm. *Development* **146**, dev180620.

Liu, X., Nefzger, C. M., Rossello, F. J., Chen, J., Knaupp, A. S., Firas, J., Ford, E., Pflueger, J., Paynter, J. M., Chy, H. S., et al. (2017). Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming. *Nature Methods* **14**, 1055–1062.

Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., et al. (2019). An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nature Communications* **10**, 364.

Liu, G., David, B. T., Trawczynski, M. and Fessler, R. G. (2020). Advances in Pluripotent Stem Cells: History, Mechanisms, Technologies, and Applications. *Stem Cell Reviews and Reports* **16**, 3–32.

Liu, X., Tan, J. P., Schröder, J., Aberkane, A., Ouyang, J. F., Mohenska, M., Lim, S. M., Sun, Y. B. Y., Chen, J., Sun, G., et al. (2021). Modelling human blastocysts by reprogramming fibroblasts into iBlastoids. *Nature* **591**, 627–632.

Lobbestael, E., Reumers, V., Ibrahimi, A., Paesen, K., Thiry, I., Gijsbers, R., van den Haute, C., Debyser, Z., Baekelandt, V. and Taymans, J.-M. (2010). Immunohistochemical detection of transgene expression in the brain using small epitope tags. *BMC Biotechnology* **10**, 16.

Loehlin, D. W. and Carroll, S. B. (2016). Expression of tandem gene duplicates is often greater than twofold. *Proceedings of the National Academy of Sciences* **113**, 5988–5992.

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

Long, M., Betrán, E., Thornton, K. and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4**, 865–875.

Lopata, A. (1980). Successes and failures in human in vitro fertilization. *Nature* **288**, 642–643.

Lopata, A., McMaster, R., McBain, J. C. and Johnston, W. I. H. (1978). In-vitro fertilization of preovulatory human eggs. *Reproduction* **52**, 339–342.

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**, 550.

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I. and Young, R. A. (2013). Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* **153**, 320–334.

Lunyak, V. v. and Rosenfeld, M. G. (2005). No Rest for REST: REST/NRSF Regulation of Neurogenesis. *Cell* **121**, 499–501.

Madeira, F., Park, Y. mi, Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* **47**, W636–W641.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *J Genet* **92**, 155–161.

Mahmudi, O., Sennblad, B., Arvestad, L., Nowick, K. and Lagergren, J. (2015). Gene-pseudogene evolution: a probabilistic approach. *BMC Genomics* **16**, S12.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E. and Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science (1979)* **339**, 823–826.

Mamoor, S. (2020). Ribosomal Proteins and Ribosomal Pseudogenes Are Differentially Expressed by Medullary and Cortical Epithelial Cells of the Thymus. *OSF Preprints*.

Mandegar, M. A., Huebsch, N., Frolov, E. B., Shin, E., Truong, A., Olvera, M. P., Chan, A. H., Miyaoka, Y., Holmes, K., Spencer, C. I., et al. (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell* **18**, 541–553.

Manfrin, A., Tabata, Y., Paquet, E. R., Vuaridel, A. R., Rivest, F. R., Naef, F. and Lutolf, M. P. (2019). Engineered signaling centers for the spatially controlled patterning of human pluripotent stem cells. *Nature Methods* **16**, 640–648.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Francis Stewart, A., Smith, A., et al. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell* **149**, 590–604.

Marques-Bonet, T., Kidd, J. M., Ventura, M., Graves, T. A., Cheng, Z., Hillier, L. W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L. A., et al. (2009a). A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–881.

Marques-Bonet, T., Girirajan, S. and Eichler, E. E. (2009b). The origins and impact of primate segmental duplications. *Trends Genet* **25**, 443–454.

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**, 7634–7638.

Masaki, H., Kato-Itoh, M., Umino, A., Sato, H., Hamanaka, S., Kobayashi, T., Yamaguchi, T., Nishimura, K., Ohtaka, M., Nakanishi, M., et al. (2015). Interspecific in vitro assay for the chimera-forming ability of human pluripotent stem cells. *Development* **142**, 3222–3230.

Mascetti, V. L. and Pedersen, R. A. (2016a). Human-Mouse Chimerism Validates Human Stem Cell Pluripotency. *Cell Stem Cell* **18**, 67–72.

Mascetti, V. L. and Pedersen, R. A. (2016b). Contributions of Mammalian Chimeras to Pluripotent Stem Cell Research. *Cell Stem Cell* **19**, 163–175.

Masterson, J. (1994). Stomatal Size in Fossil Plants: Evidence for Polyploidy in Majority of Angiosperms. *Science (1979)* **264**, 421–424.

Mattenberger, F., Sabater-Muñoz, B., Toft, C. and Fares, M. A. (2017). The Phenotypic Plasticity of Duplicated Genes in *Saccharomyces cerevisiae* and the Origin of Adaptations. *G3 Genes|Genomes|Genetics* **7**, 63–75.

Maucuer, A., le Caer, J.-P., Manceau, V. and Sobel, A. (2000). Specific Ser-Pro phosphorylation by the RNA-recognition motif containing kinase KIS. *European Journal of Biochemistry* **267**, 4456–4464.

Meistermann, D., Bruneau, A., Loubersac, S., Reignier, A., Firmin, J., François-Campion, V., Kilens, S., Lelièvre, Y., Lammers, J., Feyeux, M., et al. (2021). Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. *Cell Stem Cell* **28**, 1625–1640.

Mendjan, S., Mascetti, V. L., Ortmann, D., Ortiz, M., Karjosukarso, D. W., Ng, Y., Moreau, T. and Pedersen, R. A. (2014). NANOG and CDX2 Pattern Distinct Subtypes of Human Mesoderm during Exit from Pluripotency. *Cell Stem Cell* **15**, 310–325.

Messmer, T., von Meyenn, F., Savino, A., Santos, F., Mohammed, H., Lun, A. T. L., Marioni, J. C. and Reik, W. (2019). Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Reports* **26**, 815–824.

Minn, K. T., Fu, Y. C., He, S., Dietmann, S., George, S. C., Anastasio, M. A., Morris, S. A. and Solnica-Krezel, L. (2020). High-resolution transcriptional and morphogenetic profiling of cells from micropatterned human ESC gastruloid cultures. *Elife* **9**, e59445.

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

Miura, H., Gurumurthy, C. B., Sato, T., Sato, M. and Ohtsuka, M. (2015). CRISPR/Cas9-based generation of knockdown mice by intronic insertion of artificial microRNA using longer single-stranded DNA. *Scientific Reports* **5**, 12799.

Molè, M. A., Coorens, T. H. H., Shahbazi, M. N., Weberling, A., Weatherbee, B. A. T., Gantner, C. W., Sancho-Serra, C., Richardson, L., Drinkwater, A., Syed, N., et al. (2021). A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nature Communications* **12**, 3679.

Molyneaux, K. A., Stallock, J., Schaible, K. and Wylie, C. (2001). Time-Lapse Analysis of Living Mouse Germ Cell Migration. *Developmental Biology* **240**, 488–498.

Moretto-Zita, M., Jin, H., Shen, Z., Zhao, T., Briggs, S. P. and Xu, Y. (2010). Phosphorylation stabilizes Nanog by promoting its interaction with Pin1. *Proceedings of the National Academy of Sciences* **107**, 13312–13317.

Moris, N., Anlas, K., van den Brink, S. C., Alemany, A., Schröder, J., Ghimire, S., Balayo, T., van Oudenaarden, A. and Martinez Arias, A. (2020). An in vitro model of early anteroposterior organization during human development. *Nature* **582**, 410–415.

Morris, S. A., Teo, R. T. Y., Li, H., Robson, P., Glover, D. M. and Zernicka-Goetz, M. (2010). Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proceedings of the National Academy of Sciences* **107**, 6364–6369.

Mostert, M., Rosenberg, C., Stoop, H., Schuyer, M., Timmer, A., Oosterhuis, W. and Looijenga, L. (2000). Comparative Genomic and In Situ Hybridization of Germ Cell Tumors of the Infantile Testis. *Laboratory Investigation* **80**, 1055–1064.

Mulas, C., Kalkan, T. and Smith, A. (2017). NODAL Secures Pluripotency upon Embryonic Stem Cell Progression from the Ground State. *Stem Cell Reports* **9**, 77–91.

Mullin, N. P., Yates, A., Rowe, A. J., Nijmeijer, B., Colby, D., Barlow, P. N., Walkinshaw, M. D. and Chambers, I. (2008). The pluripotency rheostat Nanog functions as a dimer. *Biochemical Journal* **411**, 227–231.

Mullin, N. P., Gagliardi, A., Khoa, L. T. P., Colby, D., Hall-Ponsele, E., Rowe, A. J. and Chambers, I. (2017). Distinct Contributions of Tryptophan Residues within the Dimerization Domain to Nanog Function. *Journal of Molecular Biology* **429**, 1544–1553.

Mullin, N. P., Varghese, J., Colby, D., Richardson, J. M., Findlay, G. M. and Chambers, I. (2020). Phosphorylation of NANOG by casein kinase I regulates embryonic stem cell self-renewal. *FEBS Letters* 1873-3468.13969.

Murty, V. V. V. S., Dmitrovsky, E., Bosl, G. J. and Chaganti, R. S. K. (1990). Nonrandom chromosome abnormalities in testicular and ovarian germ cell tumor cell lines. *Cancer Genetics and Cytogenetics* **50**, 67–73.

Nagy, A., Gócza, E., Diaz, E. M., Prideaux, V. R., Iványi, E., Markkula, M. and Rossant, J. (1990). Embryonic stem cells alone are able to support fetal development in the mouse. *Development* **110**, 815–821.

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**, 8424–8428.

Nakamura, T., Okamoto, I., Sasaki, K., Yabuta, Y., Iwatani, C., Tsuchiya, H., Seita, Y., Nakamura, S., Yamamoto, T. and Saitou, M. (2016). A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature* **537**, 57–62.

Nakamura, T., Fujiwara, K., Saitou, M. and Tsukiyama, T. (2021). Non-human primates as a model for human development. *Stem Cell Reports* **16**, 1093–1103.

Nakashima, Y. and Omasa, T. (2016). What Kind of Signaling Maintains Pluripotency and Viability in Human-Induced Pluripotent Stem Cells Cultured on Laminin-511 with Serum-Free Medium? *BioResearch Open Access* **5**, 84–93.

Navarro, P., Festuccia, N., Colby, D., Gagliardi, A., Mullin, N. P., Zhang, W., Karwacki-Neisius, V., Osorno, R., Kelly, D., Robertson, M., et al. (2012). OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. *The EMBO Journal* **31**, 4547–4562.

Niakan, K. K. and Eggan, K. (2013). Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental Biology* **375**, 54–64.

Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. and Pera, R. A. R. (2012). Human pre-implantation embryo development. *Development* **139**, 829–841.

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

Niimura, Y. and Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences* **100**, 12235–12240.

Niimura, Y. and Nei, M. (2005). Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* **346**, 23–28.

Niimura, Y. and Nei, M. (2007). Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution. *PLoS ONE* **2**, e708.

Niwa, H., Yamamura, K. and Miyazaki, J. (1991). Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene* **108**, 193–199.

Norwitz, E. R., Schust, D. J. and Fisher, S. J. (2001). Implantation and the Survival of Early Pregnancy. *New England Journal of Medicine* **345**, 1400–1408.

Nowak, D. E., Tian, B. and Brasier, A. R. (2005). Two-step cross-linking method for identification of NF-κB gene network by chromatin immunoprecipitation. *Biotechniques* **39**, 715–725.

Nowick, K., Hamilton, A. T., Zhang, H. and Stubbs, L. (2010). Rapid Sequence and Expression Divergence Suggest Selection for Novel Function in Primate-Specific KRAB-ZNF Genes. *Molecular Biology and Evolution* **27**, 2606–2617.

Oh, J.-H., Do, H.-J., Yang, H.-M., Moon, S.-Y., Cha, K.-Y., Chung, H.-M. and Kim, J.-H. (2005). Identification of a putative transactivation domain in human Nanog. *Exp Mol Med* **37**, 250–254.

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ohta, T. (2000). Evolution of gene families. *Gene* **259**, 45–52.

Okamoto, I., Patrat, C., Thépot, D., Peynot, N., Fauque, P., Daniel, N., Diabangouaya, P., Wolf, J.-P., Renard, J.-P., Duranthon, V., et al. (2011). Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* **472**, 370–374.

Oosterhuis, J. W. and Looijenga, L. H. J. (2005). Testicular germ-cell tumours in a broader perspective. *Nature Reviews Cancer* **5**, 210–222.

Ophir, R. and Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**, 191–202.

O'Rahilly, R. and Müller, F. (2010). Developmental Stages in Human Embryos: Revised and New Measurements. *Cells Tissues Organs* **192**, 73–84.

Osafune, K., Caron, L., Borowiak, M., Martinez, R. J., Fitz-Gerald, C. S., Sato, Y., Cowan, C. A., Chien, K. R. and Melton, D. A. (2008). Marked differences in differentiation propensity among human embryonic stem cell lines. *Nature Biotechnology* **26**, 313–315.

Paik, I., Scurr, D. J., Morris, B., Hall, G., Denning, C., Alexander, M. R., Shakesheff, K. M. and Dixon, J. E. (2012). Rapid micropatterning of cell lines and human pluripotent stem cells on elastomeric membranes. *Biotechnology and Bioengineering* **109**, 2630–2641.

Pain, D., Chirn, G.-W., Strassel, C. and Kemp, D. M. (2005). Multiple Retropseudogenes from Pluripotent Cell-specific Gene Expression Indicates a Potential Signature for Novel Gene Identification. *Journal of Biological Chemistry* **280**, 6265–6268.

Palla, A. R., Piazzolla, D., Abad, M., Li, H., Dominguez, O., Schonthaler, H. B., Wagner, E. F. and Serrano, M. (2014). Reprogramming activity of NANOGP8, a NANOG family member widely expressed in cancer. *Oncogene* **33**, 2513–2519.

Palmieri, S. L., Peter, W., Hess, H. and Schöler, H. R. (1994). Oct-4 Transcription Factor Is Differentially Expressed in the Mouse Embryo during Establishment of the First Two Extraembryonic Cell Lineages Involved in Implantation. *Developmental Biology* **166**, 259–267.

Papaioannou, I., Disterer, P. and Owen, J. S. (2009). Use of internally nuclease-protected single-strand DNA oligonucleotides and silencing of the mismatch repair protein, MSH2, enhances the replication of corrected cells following gene editing. *The Journal of Gene Medicine* **11**, 267–274.

Parsons, J. D. (1995). Miropeats: graphical DNA sequence comparisons. *Bioinformatics* **11**, 615–619.

Pastor, W. A., Chen, D., Liu, W., Kim, R., Sahakyan, A., Lukianchikov, A., Plath, K., Jacobsen, S. E. and Clark, A. T. (2016). Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. *Cell Stem Cell* **18**, 323–329.

Pastor-Satorras, R., Smith, E. and Solé, R. v. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology* **222**, 199–210.

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

Pickett, F. B. and Meeks-Wagner, D. R. (1995). Seeing double: appreciating genetic redundancy. *The Plant Cell* **7**, 1347–1356.

Piper, D. E., Batchelor, A. H., Chang, C.-P., Cleary, M. L. and Wolberger, C. (1999). Structure of a HoxB1–Pbx1 Heterodimer Bound to DNA. *Cell* **96**, 587–597.

Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. and Hadjantonakis, A.-K. (2008). Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* **135**, 3081–3091.

Politzer, G. (1930). Über einen menschlichen Embryo mit sieben Urwirbelpaaren. *Zeitschrift für Anatomie und Entwicklungsgeschichte* **93.3**, 386–428.

Politzer, G. (1933). Die keimbahn des menschen. *Zeitschrift für Anatomie und Entwicklungsgeschichte* **100**, 331–361.

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T. W., Jaenisch, R. and Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724-735.e5.

Poursani, E. M., Mohammad Soltani, B. and Mowla, S. J. (2016). Differential Expression of OCT4 Pseudogenes in Pluripotent and Tumor Cell Lines. *Cell J* **18**, 28–36.

Qian, X., Kim, J. K., Tong, W., Villa-Diaz, L. G. and Krebsbach, P. H. (2016). DPPA5 Supports Pluripotency and Reprogramming by Regulating NANOG Turnover. *STEM CELLS* **34**, 588–600.

Qin, H., Hejna, M., Liu, Y., Percharde, M., Wossidlo, M., Blouin, L., Durruthy-Durruthy, J., Wong, P., Qi, Z., Yu, J., et al. (2016). YAP Induces Human Naive Pluripotency. *Cell Reports* **14**, 2301–2312.

Quadros, R. M., Miura, H., Harms, D. W., Akatsuka, H., Sato, T., Aida, T., Redder, R., Richardson, G. P., Inagaki, Y., Sakai, D., et al. (2017). Easi-CRISPR: a robust method for one-step generation of mice carrying conditional and insertion alleles using long ssDNA donors and CRISPR ribonucleoproteins. *Genome Biology* **18**, 92.

Ramakrishna, S., Suresh, B., Lim, K.-H., Cha, B.-H., Lee, S.-H., Kim, K.-S. and Baek, K.-H. (2011). PEST Motif Sequence Regulating Human NANOG for Proteasomal Degradation. *Stem Cells and Development* **20**, 1511–1519.

Reams, A. B., Kofoid, E., Kugelberg, E. and Roth, J. R. (2012). Multiple Pathways of Duplication Formation with and Without Recombination (RecA) in *Salmonella enterica*. *Genetics* **192**, 397–415.

Renaud, J.-B., Boix, C., Charpentier, M., De Cian, A., Cochennec, J., Duvernois-Berthet, E., Perrouault, L., Tesson, L., Edouard, J., Thinard, R., et al. (2016). Improved Genome Editing Efficiency and Flexibility Using Modified Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases. *Cell Reports* **14**, 2263–2272.

Riesenberg, S., Chintalapati, M., Macak, D., Kanis, P., Maricic, T. and Pääbo, S. (2019). Simultaneous precise editing of multiple genes in human cells. *Nucleic Acids Research* **47**, e116–e116.

Rivlin, N., Brosh, R., Oren, M. and Rotter, V. (2011). Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes & Cancer* **2**, 466–474.

Robertson, M., Stenhouse, F., Colby, D., Marland, J. R. K., Nichols, J., Tweedie, S. and Chambers, I. (2006). Nanog retrotransposed genes with functionally conserved open reading frames. *Mammalian Genome* **17**, 732–743.

Rogers, R. L., Shao, L. and Thornton, K. R. (2017). Tandem duplications lead to novel expression patterns through exon shuffling in Drosophila yakuba. *PLOS Genetics* **13**, e1006795.

Roode, M., Blair, K., Snell, P., Elder, K., Marchant, S., Smith, A. and Nichols, J. (2012). Human hypoblast formation is not dependent on FGF signalling. *Developmental Biology* **361**, 358–363.

Rosner, M. H., Vigano, M. A., Ozato, K., Timmons, P. M., Poirie, F., Rigby, P. W. J. and Staudt, L. M. (1990). A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. *Nature* **345**, 686–692.

Rossant, J. and Tam, P. P. L. (2017). New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell Stem Cell* **20**, 18–28.

Rossant, J., Chazaud, C. and Yamanaka, Y. (2003). Lineage allocation and asymmetries in the early mouse embryo. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **358**, 1341–1349.

Rostovskaya, M., Stirparo, G. G. and Smith, A. (2019). Capacitation of human naïve pluripotent stem cells for multi-lineage differentiation. *Development* **146**, dev172916.

Sahakyan, A., Kim, R., Chronis, C., Sabri, S., Bonora, G., Theunissen, T. W., Kuoy, E., Langerman, J., Clark, A. T., Jaenisch, R., et al. (2017). Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. *Cell Stem Cell* **20**, 87–101.

Saito, H., Ishida, G. M., Kaneko, T., Kawachiya, S., Ohta, N., Takahashi, T., Saito, T. and Hiroi, M. (2000). Application of Vitrification to Human Embryo Freezing. *Gynecologic and Obstetric Investigation* **49**, 145–149.

Saitou, M., Barton, S. C. and Surani, M. A. (2002). A molecular programme for the specification of germ cell fate in mice. *Nature* **418**, 293–300.

Sant, S., Hancock, M. J., Donnelly, J. P., Iyer, D. and Khademhosseini, A. (2010). Biomimetic gradient hydrogels for tissue engineering. *The Canadian Journal of Chemical Engineering* **88**, 899–911.

Sasidharan, R. and Gerstein, M. (2008). Protein fossils live on as RNA. *Nature* **453**, 729–731.

Sathananthan, A. H. (1984). Ultrastructural morphology of fertilization and early cleavage in the human. In *In Vitro Fertilization and Embryo Transfer* (ed. Trounson, A.) and Wood, C.), pp. 131–171. Edinburgh: Churchill Livingstone.

Sathananthan, A. H. (1990). Abnormal nuclear configurations encountered in human IVF: possile genetic implications. *Assisted Reprod.Technol. /Androl.* **1**, 115–133.

Sathananthan, A. H. (1993). Ultrastructure of fertilization and embryo development. In *Handbook of in vitro fertilization* (ed. Trounson, A.) and Gardner, D. K.), pp. 237–262. Boca Raton, Florida: CRC Press.

Sathananthan, A. H. (1994). Functional competence of abnormal spermatozoa. In *Baillieres Clinical Obstetrics and Gynaecology Micromanipulation Techniques* (ed. Fishel, S.), pp. 141–156. London: Bailliere Tindall.

Sathananthan, A. H. (1997). Mitosis in the human embryo: the vital role of the sperm centrosome (centriole). *Histol. Histopathol.* **12**, 827–856.

Sathananthan, A. H. (1998). Paternal centrosomal dynamics in early human development and infertility. *Journal of Assisted Reproduction and Genetics* **15**, 129–139.

Sathananthan, A. H. and Trounson, A. O. (1985). The human pronuclear ovum: Fine strcture of monospermic and polyspermic fertilization in vitro. *Gamete Research* **12**, 385–398.

Sathananthan, A. H. and Trounson, A. O. (1989). Effects of culutre and cryopreservation on human oocyte and emrbyo ultrastructure and function. In *Ultrastructure of human gametogenesis and early embryogenesis* (ed. In Van Blerkom, J.) and Motta, P. M.), pp. 181–200. Boston: Kluwer Academic.

Sathananthan, A., Wood, C. and Leeton, J. F. (1982). Ultrastructural evaluation of 8–16 cell human embryos cultured in vitro. *Micron (1969)* **13**, 193–203.

Sathananthan, A. H., Trounson, A. and Wood, C. (1986). *Atlas of fine structure of human sperm penetration, eggs, and embryos cultured in vitro*. Philadelphia: Praeger Scientific.

Sathananthan, A., Bongso, A., Ng, S.-C., Ho, J., Mok, H. and Ratnam, S. (1990). Ultrastructure of preimplantation human embryos co-cultured with human ampullary cells. *Human Reproduction* **5**, 309–318.

Sathananthan, A. H., Ratnam, S. S. and Trounson, A. (1999). Early human developement. VIDEO. *Human Reprod. Update* **5**, 89.

Sathananthan. A, Ng, S. C. and Bongso, A. (1993). *Visual atlas of early human development for assisted reproductive technology*. Singapore: Serono.

Scerbo, P., Markov, G. v., Vivien, C., Kodjabachian, L., Demeneix, B., Coen, L. and Girardot, F. (2014). On the Origin and Evolutionary History of NANOG. *PLoS ONE* **9**, e85104.

Schneider, D. T., Schuster, A. E., Fritsch, M. K., Calaminus, G., Göbel, U., Harms, D., Lauer, S., Olson, T. and Perlman, E. J. (2002). Genetic analysis of mediastinal nonseminomatous germ cell tumors in children and adolescents. *Genes, Chromosomes and Cancer* **34**, 115–125.

Schoenfelder, K. P. and Fox, D. T. (2015). The expanding implications of polyploidy. *Journal of Cell Biology* **209**, 485–491.

Schoenherr, C. J. and Anderson, D. J. (1995). The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**, 1360–1363.

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* **24**, 347–350.

Seki, Y., Yamaji, M., Yabuta, Y., Sano, M., Shigeta, M., Matsui, Y., Saga, Y., Tachibana, M., Shinkai, Y. and Saitou, M. (2007). Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. *Development* **134**, 2627–2638.

Shahbazi, M. N., Jedrusik, A., Vuoristo, S., Recher, G., Hupalowska, A., Bolton, V., Fogarty, N. M. E., Campbell, A., Devito, L. G., Ilic, D., et al. (2016). Self-organization of the human embryo in the absence of maternal tissues. *Nature Cell Biology* **18**, 700–708.

Shakiba, N., White, C. A., Lipsitz, Y. Y., Yachie-Kinoshita, A., Tonge, P. D., Hussein, S. M. I., Puri, M. C., Elbaz, J., Morrissey-Scoot, J., Li, M., et al. (2015). CD24 tracks divergent pluripotent states in mouse and human cells. *Nature Communications* **6**, 7329.

Shay, J. W. and Werbin, H. (1992). New evidence for the insertion of mitochondrial DNA into the human genome: significance for cancer and aging. *Mutation Research/DNAging* **275**, 227–235.

Shepherd, G. M. (2004). The Human Sense of Smell: Are We Better Than We Think? *PLoS Biology* **2**, e146.

Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeliger, S. R., Fried, Y. C., Ainbinder, E., Friedman, N. and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119.

Silva, J., Nichols, J., Theunissen, T. W., Guo, G., van Oosten, A. L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I. and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. *Cell* **138**, 722–37.

Simmet, K., Zakhartchenko, V., Philippou-Massier, J., Blum, H., Klymiuk, N. and Wolf, E. (2018). OCT4/POU5F1 is required for NANOG expression in bovine blastocysts. *Proceedings of the National Academy of Sciences* **115**, 2770–2775.

Skarnes, W. C., Pellegrino, E. and McDonough, J. A. (2019). Improving homology-directed repair efficiency in human stem cells. *Methods* **164–165**, 18–28.

Smith, J. R., Vallier, L., Lupo, G., Alexander, M., Harris, W. A. and Pedersen, R. A. (2008). Inhibition of Activin/Nodal signaling promotes specification of human embryonic stem cells into neuroectoderm. *Developmental Biology* **313**, 107–117.

Sozen, B., Amadei, G., Cox, A., Wang, R., Na, E., Czukiewska, S., Chappell, L., Voet, T., Michel, G., Jing, N., et al. (2018). Self-assembly of embryonic and two extra-embryonic stem cell types into gastrulating embryo-like structures. *Nature Cell Biology* **20**, 979–989.

Sozen, B., Jorgensen, V., Weatherbee, B. A. T., Chen, S., Zhu, M. and Zernicka-Goetz, M. (2021). Reconstructing aspects of human embryogenesis with pluripotent stem cells. *Nature Communications* **12**, 5550.

Speed, R. M. (1982). Meiosis in the foetal mouse ovary. *Chromosoma* **85**, 427–437.

Štefková, K., Procházková, J. and Pacherník, J. (2015). Alkaline Phosphatase in Stem Cells. *Stem Cells International* **2015**, 1–11.

Steptoe, P. C., Edwards, R. G. and Purdy, J. M. (1971). Human Blastocysts grown in Culture. *Nature* **229**, 132–133.

Steptoe, P. C., Edwards, R. G. and Purdy, J. M. (1980). Clinical aspects of pregnancies established with cleaving embryos grown in vitro*. *BJOG: An International Journal of Obstetrics and Gynaecology* **87**, 757–768.

Stevens, A., Smith, H., Garner, T., Minogue, B., Sneddon, S., Shaw, L., Keramari, M., Oldershaw, R., Bates, N., Brison, D. R., et al. (2019). Interactome comparison of human embryonic stem cell lines with the inner cell mass and trophectoderm. *bioRxiv* 411439.

Stewart, C. L. (1993). [50] Production of chimeras between embryonic stem cells and embryos.pp. 823–856.

Stirparo, G. G., Boroviak, T., Guo, G., Nichols, J., Smith, A. and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human preimplantation epiblast. *Development* **145**, dev158501.

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**, 2093–2102.

Stuart, H. T. (2019). Studying the principles of cell identity transitions using naïve pluripotency induction as a model (Doctoral thesis).

Su, Z., Zhang, Y., Liao, B., Zhong, X., Chen, X., Wang, H., Guo, Y., Shan, Y., Wang, L. and Pan, G. (2018). Antagonism between the transcription factors NANOG and OTX2 specifies rostral or caudal cell fate during neural patterning transition. *Journal of Biological Chemistry* **293**, 4445–4455.

Swijnenburg, R.-J., Tanaka, M., Vogel, H., Baker, J., Kofidis, T., Gunawan, F., Lebl, D. R., Caffarelli, A. D., de Bruin, J. L., Fedoseyeva, E. v., et al. (2005). Embryonic Stem Cell Immunogenicity Increases Upon Differentiation After Transplantation Into Ischemic Myocardium. *Circulation* **112**, 166–172.

Swijnenburg, R.-J., Schrepfer, S., Cao, F., Pearl, J. I., Xie, X., Connolly, A. J., Robbins, R. C. and Wu, J. C. (2008). In Vivo Imaging of Embryonic Stem Cells Reveals Patterns of Survival and Immune Rejection Following Transplantation. *Stem Cells and Development* **17**, 1023–1029.

Tachibana, M., Clepper, L. L., Sparman, M. L., Sritanaudomchai, H., Ramsey, C. M. and Mitalipov, S. M. (2009). NANOG regulates pluripotency in the early primate embryo. *Fertility and Sterility* **92**, S226–S227.

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

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**, 861–872.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell* **158**, 1254–1269.

Tam, P. P. and Snow, M. H. (1981). Proliferation and migration of primordial germ cells during compensatory growth in mouse embryos. *J Embryol Exp Morphol* **64**, 133–147.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution* **24**, 1596–1599.

Tapia, N., Reinhardt, P., Duemmler, A., Wu, G., Araúzo-Bravo, M. J., Esch, D., Greber, B., Cojocaru, V., Rascon, C. A., Tazaki, A., et al. (2012). Reprogramming to pluripotency is an ancient trait of vertebrate Oct4 and Pou2 proteins. *Nature Communications* **3**, 1279.

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**, 196–199.

Testart, J., LASSALLE, B., BELAISCH-ALLART, J., FORMAN, R., HAZOUT, A., FRIES, N. and FRYDMAN, R. (1988). Human Embryo Freezing. *Ann N Y Acad Sci* **541**, 532–540.

Tewary, M., Ostblom, J., Prochazka, L., Zulueta-Coarasa, T., Shakiba, N., Fernandez-Gonzalez, R. and Zandstra, P. W. (2017). A stepwise model of Reaction-Diffusion and Positional-Information governs self-organized human peri-gastrulation-like patterning. *Development* **144**, 4298–4312.

Tewary, M., Dziedzicka, D., Ostblom, J., Prochazka, L., Shakiba, N., Heydari, T., Aguilar-Hidalgo, D., Woodford, C., Piccinini, E., Becerra-Alonso, D., et al. (2019). High-throughput micropatterning platform reveals Nodal-dependent bisection of peri-gastrulation–associated versus preneurulation-associated fate patterning. *PLOS Biology* **17**, e3000081.

Thakore-Shah, K., Koleilat, T., Jan, M., John, A. and Pyle, A. D. (2015). REST/NRSF Knockdown Alters Survival, Lineage Differentiation and Signaling in Human Embryonic Stem Cells. *PLOS ONE* **10**, e0145280.

Theunissen, T. W., van Oosten, A. L., Castelo-Branco, G., Hall, J., Smith, A. and Silva, J. C. R. (2011a). Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. *Current Biology* **21**, 65–71.

Theunissen, T. W., Costa, Y., Radzisheuskaya, A., van Oosten, A. L., Lavial, F., Pain, B., Castro, L. F. C. and Silva, J. C. R. (2011b). Reprogramming capacity of Nanog is functionally conserved in vertebrates and resides in a unique homeodomain. *Development* **138**, 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., et al. (2014). Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487.

Theunissen, T. W., Friedli, M., He, Y., Planet, E., O'Neil, R. C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., et al. (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**, 502–515.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Thomson, J. A. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science (1979)* **282**, 1145–1147.

Thomson, J. A., Kalishman, J., Golos, T. G., Durning, M., Harris, C. P., Becker, R. A. and Hearn, J. P. (1995). Isolation of a primate embryonic stem cell line. *Proceedings of the National Academy of Sciences* **92**, 7844–7848.

Thomson, J. A., Kalishman, J., Golos, T. G., Durning, M., Harris, C. P. and Hearn, J. P. (1996). Pluripotent Cell Lines Derived from Common Marmoset (Callithrix jacchus) Blastocysts. *Biology of Reproduction* **55**, 254–259.

Tonner, P., Srinivasasainagendra, V., Zhang, S. and Zhi, D. (2012). Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* **13**, 412.

Torres, J. and Watt, F. M. (2008). Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFκB and cooperating with Stat3. *Nature Cell Biology* **10**, 194–201.

Tosolini, M. and Jouneau, A. (2015). Acquiring Ground State Pluripotency: Switching Mouse Embryonic Stem Cells from Serum/LIF Medium to 2i/LIF Medium.pp. 41–48.

Tsakiridis, A., Huang, Y., Blin, G., Skylaki, S., Wymeersch, F., Osorno, R., Economou, C., Karagianni, E., Zhao, S., Lowell, S., et al. (2014). Distinct Wnt-driven primitive streak-like populations reflect *in vivo* lineage precursors. *Development* **141**, 1209–1221.

Tyser, R. C. v., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A. and Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285–289.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. and Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**, e115–e115.

Valamehr, B., Robinson, M., Abujarour, R., Rezner, B., Vranceanu, F., Le, T., Medcalf, A., Lee, T. T., Fitch, M., Robbins, D., et al. (2014). Platform for Induction and Maintenance of Transgene-free hiPSCs Resembling Ground State Pluripotent Stem Cells. *Stem Cell Reports* **2**, 366–381.

Vallier, L., Alexander, M. and Pedersen, R. A. (2005). Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *Journal of Cell Science* **118**, 4495–4509.

Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L. E., Trotter, M. W. B., Cho, C. H.-H., Martinez, A., Rugg-Gunn, P., et al. (2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development* **136**, 1339–1349.

Vallot, C., Patrat, C., Collier, A. J., Huret, C., Casanova, M., Liyakat Ali, T. M., Tosolini, M., Frydman, N., Heard, E., Rugg-Gunn, P. J., et al. (2017). XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. *Cell Stem Cell* **20**, 102–111.

van Royen, E., Mangelschots, K., de Neubourg, D., Valkenburg, M., van de Meerssche, M., Ryckaert, G., Eestermans, W. and Gerris, J. (1999). Characterization of a top quality embryo, a step towards single-embryo transfer. *Human Reproduction* **14**, 2345–2349.

Vázquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003). Modeling of Protein Interaction Networks. *Complexus* **1**, 38–44.

Verma, R., Liu, J., Holland, M. K., Temple-Smith, P., Williamson, M. and Verma, P. J. (2013). Nanog Is an Essential Factor for Induction of Pluripotency in Somatic Cells from Endangered Felids. *BioResearch Open Access* **2**, 72–76.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., et al. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566.

Vinckenbosch, N., Dupanloup, I. and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences* **103**, 3220–3225.

Wang, Z. (2009). Epitope Tagging of Endogenous Proteins for Genome-Wide Chromatin Immunoprecipitation Analysis.pp. 87–98.

Wang, S.-H., Tsai, M.-S., Chiang, M.-F. and Li, H. (2003). A novel NK-type homeobox gene, ENK (early embryo specific NK), preferentially expressed in embryonic stem cells. *Gene Expression Patterns* **3**, 99–103.

Wang, G., Zhang, H., Zhao, Y., Li, J., Cai, J., Wang, P., Meng, S., Feng, J., Miao, C., Ding, M., et al. (2005). Noggin and bFGF cooperate to maintain the pluripotency of human embryonic stem cells in the absence of feeder layers. *Biochemical and Biophysical Research Communications* **330**, 934–942.

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

Wang, J., Levasseur, D. N. and Orkin, S. H. (2008a). Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences* **105**, 6326–6331.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A. and Liu, P. (2008b). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences* **105**, 9290–9295.

Wang, X., Jin, J., Wan, F., Zhao, L., Chu, H., Chen, C., Liao, G., Liu, J., Yu, Y., Teng, H., et al. (2019). AMPK Promotes SPOP-Mediated NANOG Degradation to Regulate Prostate Cancer Cell Stemness. *Developmental Cell* **48**, 345-360.e7.

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., et al. (2014). Derivation of naive human embryonic stem cells. *Proceedings of the National Academy of Sciences* **111**, 4484–4489.

Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. and Brivanlou, A. H. (2014). A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nature Methods* **11**, 847–854.

Watanabe, K., Ueno, M., Kamiya, D., Nishiyama, A., Matsumura, M., Wataya, T., Takahashi, J. B., Nishikawa, S., Nishikawa, S., Muguruma, K., et al. (2007). A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nature Biotechnology* **25**, 681–686.

Weiler, S., Tsao, D., Gruschus, J., Yu, L., Wang, L. H., Nirenberg, M. and Ferretti, J. (1996). Role of amino acid residues in the recognition helix of the NK-2 homeodomain on structure, function and thermal stability. *Biophys. J.* **70**, A342.

Weiler, S., Gruschus, J. M., Tsao, D. H. H., Yu, L., Wang, L.-H., Nirenberg, M. and Ferretti, J. A. (1998). Site-directed Mutations in the vnd/NK-2 Homeodomain. *Journal of Biological Chemistry* **273**, 10994–11000.

Weinberger, L., Ayyash, M., Novershtern, N. and Hanna, J. H. (2016). Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nature Reviews Molecular Cell Biology* **17**, 155–169.

Whittington, C. M., Papenfuss, A. T., Bansal, P., Torres, A. M., Wong, E. S. W., Deakin, J. E., Graves, T., Alsop, A., Schatzkamer, K., Kremitzki, C., et al. (2008). Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Research* **18**, 986–994.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. and Young, R. A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307–319.

Witschi, E. (1948). Migration of the germ cells of human embryos from the yolk sac to the primitive gonadal folds. *Carnegie Institute Contributions to Embryology* **32**, 67–80.

Woischnik, M. and Moraes, C. T. (2002). Pattern of Organization of Human Mitochondrial Pseudogenes in the Nuclear Genome. *Genome Research* **12**, 885–893.

Woltjen, K., Michael, I. P., Mohseni, P., Desai, R., Mileikovsky, M., Hämäläinen, R., Cowling, R., Wang, W., Liu, P., Gertsenstein, M., et al. (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458**, 766–770.

Wong, C. C., Loewke, K. E., Bossert, N. L., Behr, B., de Jonge, C. J., Baer, T. M. and Pera, R. A. R. (2010). Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nature Biotechnology* **28**, 1115–1121.

Wray, J., Kalkan, T. and Smith, A. G. (2010). The ground state of pluripotency. *Biochemical Society Transactions* **38**, 1027–1032.

Xiang, L., Yin, Y., Zheng, Y., Ma, Y., Li, Y., Zhao, Z., Guo, J., Ai, Z., Niu, Y., Duan, K., et al. (2020). A developmental landscape of 3D-cultured human pre-gastrulation embryos. *Nature* **577**, 537–542.

Xie, X., Piao, L., Cavey, G. S., Old, M., Teknos, T. N., Mapp, A. K. and Pan, Q. (2014). Phosphorylation of Nanog is essential to regulate Bmi1 and promote tumorigenesis. *Oncogene* **33**, 2040–2052.

Xu, R.-H., Chen, X., Li, D. S., Li, R., Addicks, G. C., Glennon, C., Zwaka, T. P. and Thomson, J. A. (2002). BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nature Biotechnology* **20**, 1261–1264.

Xu, R.-H., Peck, R. M., Li, D. S., Feng, X., Ludwig, T. and Thomson, J. A. (2005). Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nature Methods* **2**, 185–190.

Xu, R.-H., Sampsell-Barron, T. L., Gu, F., Root, S., Peck, R. M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., Stewart, R., et al. (2008). NANOG Is a Direct Target of TGFβ/Activin-Mediated SMAD Signaling in Human ESCs. *Cell Stem Cell* **3**, 196–206.

Xu, Z., Robitaille, A. M., Berndt, J. D., Davidson, K. C., Fischer, K. A., Mathieu, J., Potter, J. C., Ruohola-Baker, H. and Moon, R. T. (2016). Wnt/β-catenin signaling promotes self-renewal and inhibits the primed state transition in naïve human embryonic stem cells. *Proceedings of the National Academy of Sciences* **113**, E6382–E6390.

Yamaguchi, S., Kimura, H., Tada, M., Nakatsuji, N. and Tada, T. (2005). Nanog expression in mouse germ cell development. *Gene Expression Patterns* **5**, 639–646.

Yamaguchi, S., Kurimoto, K., Yabuta, Y., Sasaki, H., Nakatsuji, N., Saitou, M. and Tada, T. (2009). Conditional knockdown of Nanog induces apoptotic cell death in mouse migrating primordial germ cells. *Development* **136**, 4011–4020.

Yamane, M., Ohtsuka, S., Matsuura, K., Nakamura, A. and Niwa, H. (2018). Overlapping function of klf family targets multiple transcription factors to maintain naïve pluripotency of ES cells. *Development* **145**, dev162404.

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

Yanagida, A., Spindlow, D., Nichols, J., Dattani, A., Smith, A. and Guo, G. (2021). Naive stem cell blastocyst model captures human embryo lineage segregation. *Cell Stem Cell* **28**, 1016-1022.e4.

Yang, Z. and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution* **17**, 32–43.

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

Yoshimi, K., Kunihiro, Y., Kaneko, T., Nagahora, H., Voigt, B. and Mashimo, T. (2016). ssODN-mediated knock-in with CRISPR-Cas for large genomic regions in zygotes. *Nature Communications* **7**, 10431.

Young, J. M. (2002). Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Human Molecular Genetics* **11**, 535–546.

Yu, L., Wei, Y., Duan, J., Schmitz, D. A., Sakurai, M., Wang, L., Wang, K., Zhao, S., Hon, G. C. and Wu, J. (2021). Blastocyst-like structures generated from human pluripotent stem cells. *Nature* **591**, 620–626.

Yuan, J. D., Shi, J. X., Meng, G. X., An, L. G. and Hu, G. X. (1999). Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. *Cell Research* **9**, 281–290.

Zaehres, H., Lensch, M. W., Daheron, L., Stewart, S. A., Itskovitz-Eldor, J. and Daley, G. Q. (2005). High-efficiency RNA interference in human embryonic stem cells. *Stem Cells* **23**, 299–305.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771.

Zhang, Z. (2002). Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome. *Genome Research* **12**, 1466–1482.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292–298.

Zhang, X. and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience* **5**, 124–133.

Zhang, J., Wang, X., Li, M., Han, J., Chen, B., Wang, B. and Dai, J. (2006). NANOGP8 is a retrogene expressed in cancers. *FEBS Journal* **273**, 1723–1730.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137.

Zhao, C., Reyes, A. P., Schell, J. P., Weltner, J., Ortega, N., Zheng, Y., Björklund, Å. K., Rossant, J., Fu, J., Petropoulos, S., et al. (2021). Reprogrammed iBlastoids contain amnion-like cells but not trophectoderm. *bioRxiv* 2021.05.07.442980.

Zhong, Y. and Holland, P. W. H. (2011a). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development* **13**, 567–568.

Zhong, Y. and Holland, P. W. (2011b). The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. *BMC Evolutionary Biology* **11**, 169.

Zozulya, S., Echeverri, F. and Nguyen, T. (2001). The human olfactory receptor repertoire. *Genome Biology* **2**, RESEARCH0018.