Chapter 2

Materials and Methods

2.1 Cell culture conditions

Cell cultures were done in collaboration with Dr. Jason Tsang. The G4 (C57BL/6Ncr x 129S6/SvEvTac) mouse hybrid (George et al., 2007) embryonic stem cells were obtained from Mount Sinai Hospital and were maintained on STO feeders in serum-containing media at 5% CO2 and 37°C. They were sub-cloned, and a line with normal karyotype was selected based on spectral karyotyping analysis performed at the Molecular Cytogenetics core facility at the Sanger Institute for further analysis. The cells were split onto gelatinized plates (10cm, Corning) and expanded in serum-containing media or chemically defined media (standard 2i or alternative 2i) for at least three passages.

The three media are as follows:

1) Serum-containing media: Knockout DMEM (Gibco), 1X penicillin-streptomycin-glutamine (Gibco), 1X non-essential amino acids (Gibco), 100

U/ml recombinant human leukaemia inhibitory factor (Millipore), 15% foetal bovine serum (HyClone), 0.1 mM β-mercaptoethanol (Sigma).

2) Standard 2i media: N2B27 basal media (NDiff 227, StemCells), 100 U/ml recombinant human leukaemia inhibitory factor (Millipore), 1 μM PD0325901 (Stemgent), 3 μM CHIR99021 (Stemgent).

3) Alternative 2i media: N2B27 basal media (NDiff 227, StemCells), 100 U/ml recombinant human leukaemia inhibitory factor (Millipore), 1 μM CGP77675 (Sigma), 3 μM CHIR99021 (Stemgent).

Dr. Alex Tuck performed NPC differentiation time course using protocol published by Bibel et al., 2007 and harvested cells at day 6 and day 8. He prepared libraries for single cell mRNA sequencing using the protocol described in the section 2.2 and these samples were sequenced 150bp paired end on Illumina HiSeq2000. Mapping and downstream analysis was preformed as described in the section 2.5.

2.2 Single cell mRNA-seq using SmartSeq and Fluidigm C1

2.2.1 Single cell suspension preparation

Cells were harvested by trypsinisation (0.05% trypsin/EDTA, Gibco) for 10 minutes, when they reach 70-80% confluence for single cell capture. Subsequently they were inspected under the microscope to assure the cells are a single cell suspension, counted and diluted to 1.3×10^{-6} cells per millilitre.

2.2.2 cDNA synthesis and amplification

For each culture condition, 4000 cells were loaded on to a 10-17 micron Fluidigm C1 Single-Cell Auto Prep IFC, and cell capture was performed according to the manufacturer's instructions. The capture efficiency was determined using a microscope to exclude samples from the analysis with no or more than one cell captured or samples where in addition to cell there was cellular debris visible. Upon capture, reverse transcription and cDNA preamplification were performed in the 10-17 microns Fluidigm C1 Single-Cell Auto Prep IFC using the SMARTer PCR cDNA Synthesis kit (Clontech) and the Advantage 2 PCR kit (Ramskold et al., 2012).

Within the C1 cells are first lysed to release RNA using Triton-X 100 in the lysis buffer. Subsequently reverse transcription mix is added to perform reverse transcription. Importantly template-switching mechanism is used to avoid additional steps of adapter ligation and second strand synthesis.

The yield of the cDNA from a single cell is low, so it needs to be amplified before library preparation can be performed. During reverse transcription, adaptors are incorporated within the primers to allow amplification of full-length transcript by PCR. Reverse transcription is primed using a poly-T oligonucleotide, which allows selection of polyadenylated RNA species i.e. mRNA and some lncRNAs; this avoids sequencing abundant rRNAs. Full-length amplified cDNA was harvested, assessed and quantified using High Sensitivity DNA Kit (Agilent) and stored at -20°C.

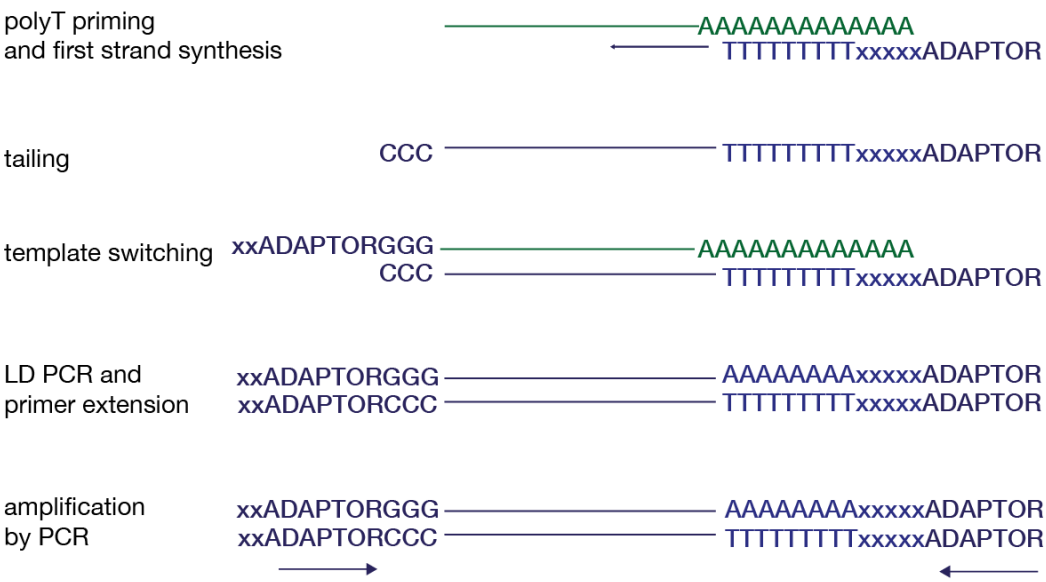

Figure 2.1 Schematic of cDNA synthesis and amplification
Polyadenylated RNAs are selected by reverse transcription with polyT primer; second strand is synthesized with template switching reaction. cDNA is amplified by PCR.

2.2.3 Illumina library preparation using Nextera XT

cDNA was diluted to a range of 0.1-0.3 ng/μl and Nextera libraries were prepared using the Nextera XT DNA Sample Preparation Kit and the Nextera Index Kit (Illumina) following the instructions in the Fluidigm manual "Using the C1™ Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing". Libraries from one chip were pooled, and paired-end 100bp sequencing was performed on 4 lanes of an Illumina HiSeq2000.

2.3 mRNA sequencing of bulk controls

Bulk mRNA sequencing libraries were prepared and sequenced using the Wellcome Trust Sanger Institute sample preparation pipeline with the TruSeq RNA Sample Preparation v2 kit (Illumina). RNA was extracted from 1-2 million cells using the Qiagen RNA Purification Kit on a QiaCube robot. The quality of the RNA sample was checked using gel electrophoresis. For library

preparation, poly-A RNA was purified from total RNA using oligo-dT magnetic pull-down. Subsequently, mRNA was fragmented using metal-ion catalysed hydrolysis. The cDNA was synthesized using random hexamer priming, and end repair was performed to obtain blunt ends. A-tailing was done to enable subsequent ligation of Illumina paired-end sequencing adapters, and samples were multiplexed at this stage. The resulting library was amplified using 10 cycles of PCR, substituting the Kapa Hifi polymerase for the polymerase in the Illumina TruSeq kit. Samples were diluted to 4nM, and 100bp paired end sequencing carried out on an Illumina HiSeq2000. The Sanger sequencing facility performed Sequencing Quality Control.

2.4 Candidate gene expression downregulation using CRISPR repressor

2.4.1 CRISPRi plasmids and cloning

Expression of candidate pluripotency regulators was downregulated with CRISPRi technology. I obtained three plasmids necessary for genome integration and expression of dCas9-KRAB and gRNA from Dr. Xuefei Gao. Two plasmids were used, one bearing gRNA linked to mCherry (Figure 2.2) and the second one dCas9-KRAB linked to BFP (Figure 2.3). Both expression cassettes are within LTR sites that are integrated into the genome using the hyperactive piggyBac transposase (Yusa et al., 2011) expressed from the third plasmid (Gao et al., 2014) (Figure 2.4).

Oligonucleotides targeting sites at promoters of candidate genes were ordered from Sigma-Aldrich (Table 2.1). I diluted the oligos to 1 mM in water and mixed them 1:1. I took 10 μl of oligo mix and heated it up to 98°C in the

thermo-cycler and then lowered the temperature by 1°C every minute until it reached 20°C to anneal the oligos and create sticky ends for the ligation to the backbone. pPB-gRNA-BsaI backbone was designed in a way that there are two BsaI cutting sites in the position where annealed oligos need to be ligated.

I performed restriction digestion of the plasmid using BsaI enzyme from New England Biolabs for 2h at 37°C. In 50 μl reaction I digested 2 μg of plasmid using 20U of the enzyme in 1x CutSmart buffer. Subsequently I ran a 2% agarose gel, cut the band corresponding to the double cut plasmid and purified the DNA using Qiagen Gel Extraction kit. Ligation was performed for each insert in the same way. 0.05 μg of plasmid was mixed with 5 μl of 5 mM annealing product, 1U of T4 DNA ligase from Thermo Fisher in 20 μl reaction containing 1x ligation buffer. Ligation was done for 1h at room temperature. 1 μl of ligation reaction was used for heat shock transformation of 25 μl of DH5α cells. Cells were plated on ampicillin for selection of successfully transformed cells and subsequently colonies were picked and grown in LB media and then I purified plasmids using MiniPrep kits from Qiagen. To check if ligation was successful I performed test digestions with BglII and XhoI (if successful 0.5k, 1.7kb and 3.9kb fragments were observed, if not: 0.9kb, 1.7kb and 3.9kb fragments) and subsequently sent plasmids for Sanger sequencing.

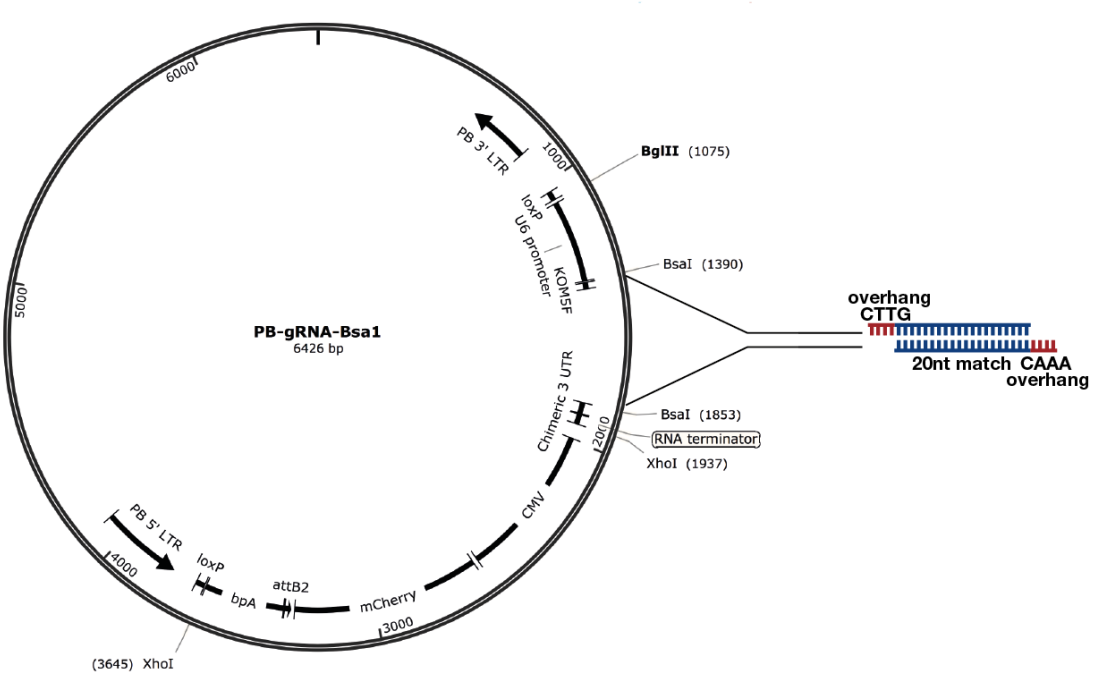

Figure 2.2 Schematic of gRNA plasmid

Column	Name	PRIMER 1	PRIMER 2
Ptma	Ptma-1	cttggcgccgcgtgagtcccccac	aaacgtgggggactcacgcggcgc
	Ptma-2	cttgcaatagcgccgggactaggg	aaacccctagtcccggcgctattg
	Ptma-3	cttgctgcgctcagccaatagcgc	aaacgcgctattggctgagcgcag
	Ptma-4	cttgttcggaatcgagccaatgag	aaacctcattggctcgattccgaa
	Ptma-5	cttggcgcagcgcgcgccaagccg	aaaccggcttggcgcgcgctgcgc
Set	Set-1	cttgtgctgattggagggagggcg	aaaccgccctccctccaatcagca
	Set-2	cttgtcaaagaagtttctgctgat	aaacatcagcagaaacttctttga
	Set-3	cttggccgccccctttctccatcgc	aaacgcgatggagaaggggggcggc
	Set-4	cttgcccggcgcgcctgcgctctg	aaaccagagcgcaggcgcgccggg
	Set-5	cttggccgggggcgggacttgcgc	aaacgcgcaagtcccgccccggc
	Set-6	cttgacggcgcgagcctctccggc	aaacgccggagaggctcgcgccgt
	Set-7	cttgggggagcaccgcgcgggggc	aaacgcccccgcgcggtgctcccc
Zfp710	Zfp710-1	cttgggagagcagggaagtgtggg	aaaccccacacttccctgctctcc
	Zfp710-2	cttggatgagaaggggtggagcca	aaactggctccacccccttctcatc
	Zfp710-3	cttgtgtgggaggaattgatgaga	aaactctcatcaattcctcccaca
	Zfp710-4	cttgccagggagagcagggaagtg	aaaccacttccctgctctccctgg
	Zfp710-5	cttgcctctgcgagcaggcttagg	aaaccctaagcctgctcgcagagg
	Zfp710-6	cttggaaaacaaaagagagataaa	aaactttatctctcttttgttttc
	Zfp710-7	cttgaagaagaaaaatcctctctg	aaaccagagaggatttttcttctt
	Zfp710-8	cttgtccaggcttgcaattcgagt	aaacactcgaattgcaagcctgga
Zfp640	Zfp640-1	cttgcaagatcactgtggctgtgc	aaacgcacagccacagtgatcttg
	Zfp640-2	cttggacaaagaggcgggatcttc	aaacgaagatcccgcctctttgtc
	Zfp640-3	cttgggaagcaaactttaacatta	aaactaatgttaaagtttgcttcc
	Zfp640-4	cttgactggccaatcaagttcgcc	aaacggcgaacttgattggccagt
Kat6b	Kat6b-1	cttggggctctgtgcgctgcagcc	aaacggctgcagcgcacagagccc
	Kat6b-2	cttgcctcccctgagggcggtgag	aaacctcaccgccctcaggggagg
	Kat6b-3	cttgcgggtgacggacagacccgt	aaacacgggtctgtccgtcacccg
	Kat6b-4	cttgggcatccccgccctcccctg	aaaccaggggagggcggggatgcc
Etv5	Etv5-1	cttgccggaggccggcgcgcagag	aaacctctgcgcgccggcctccgg
	Etv5-2	cttggacgtgtgtgctctgggctg	aaaccagcccagagcacacacgtc
	Etv5-3	cttgcggggatggccgccgaccaa	aaacttggtcggcggccatccccg
	Etv5-4	cttgcaagaggtgatgggcagccg	aaaccggctgcccatcacctcttg
	Etv5-5	cttgaaggtggctacacaggcaag	aaaccttgcctgtgtagccaccttt
	Etv5-6	cttgtttttcagtgcaagtaaggg	aaaccccttacttgcactgaaaaa
	Etv5-7	cttgggcttttgtggtagacaggc	aaacgcctgtctaccacaaaagcc
	Etv5-8	cttgttggttggttttggcttttg	aaaccaaaagccaaaaccaaccaa
Dpy30	Dpy30-1	cttggtctgctgcccgcgggggtg	aaaccacccccgcgggcagcagac
	Dpy30-2	cttgcgacgaggacggccagtcgg	aaacccgactggccgtcctcgtcg
	Dpy30-3	cttgccgagcctcgcgatgcgacg	aaaccgtcgcatcgcgaggctcgg
	Dpy30-4	cttgtcctcccaccgctacatcct	aaacaggatgtagcggtgggagga
	Dpy30-5	cttgatttgcctcaagtctgtaaa	aaactttacagacttgaggcaaat
	Dpy30-6	cttgatacatacttcttgaacaat	aaacattgttcaagaagtatgtat

Table 2.1 Sequences of oligonucleotides used to construct insert gRNA plasmid

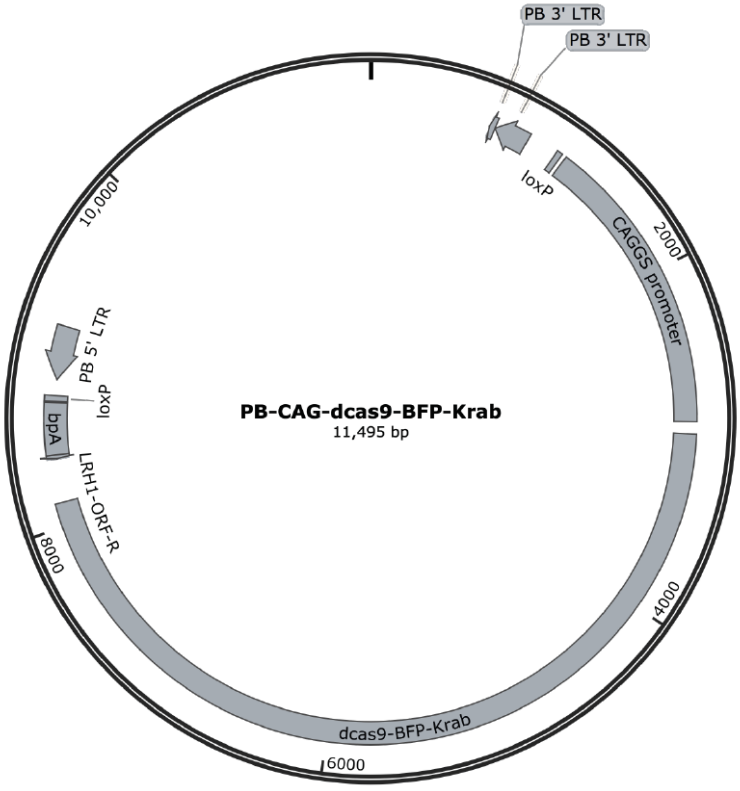

Figure 2.3 Schematic of dCas9-Krab plasmid

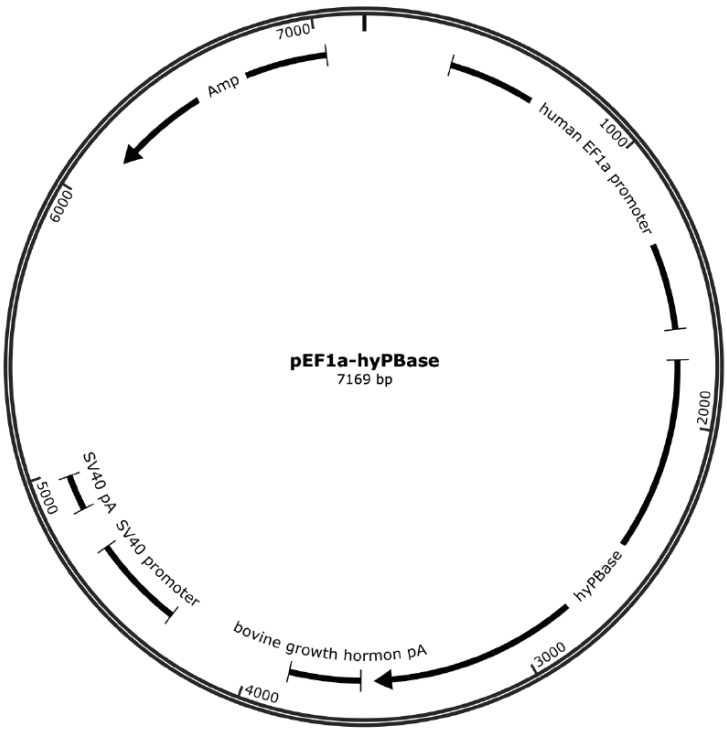

Figure 2.4 Schematic of piggyBac plasmid

2.4.2 Downregulation of target gene expression and cell sorting

GFP-OCT4 reporter strain ES cells (Silva et al., 2008) were grown by Dr. Xuefei Gao in 6-well plates and transfected with plasmids (1) 5 μg of dCas9-KRAB-BFP plasmid (2) 1 μg of hyPBase plasmid and (3) 1 μg cocktail of gRNA plasmids (Gao et al., 2014) targeting the gene of interest in a 1:1 ratio using Lipofectamine2000, Life Technologies. Subsequently, cells were cultured in knockout DMEM (Gibco) medium containing 15% serum and 100 U/ml LIF for 4 days. After 4 days cells were harvested from the culture dish using trypsin (0.05% trypsin/EDTA, Gibco) and the Cytometry Core Facility at Sanger Institute sorted BFP and mCherry double positive cells.

2.4.3 Library preparation

RNA was extracted using Qiagen RNeasy Mini kit from 10,000 mCherry and BFP positive cells that were sorted for each sample in triplicates. Modified SmartSeq2 protocol was used for reverse transcription and amplification of cDNA (Picelli et al., 2014), because the amount of RNA from 10,000 cells is not sufficient for conventional bulk library preparation protocols, which involve polyA species enrichment where a substantial amount of material is lost. Sequencing libraries were prepared using Nextera XT kit according to manufacturers guidelines, barcoded with Nextera XT Dual Index kit and sequenced on an Illumina HiSeq2500 in rapid mode.

2.5 Data analysis

2.5.1 Sequencing reads alignment

For each cell, 100bp paired-end reads were aligned to the *Mus musculus* genome (GRCm38) using GSNAP (version gmap-2014-05-15_v2) with default options (Wu and Nacu, 2010). To detect splice junctions in reads, I used a set of known splice sites from the GTF file for GRCm38 provided by Ensembl (release 73). Only reads uniquely mapped to the genome were counted for each gene using htseq-count and the same GTF file (Anders et al., 2014).

Dr. Jong Kyoung Kim additionally applied location and scale adjustments to the normalized read counts to remove technical variation among multiple batches as described below.

2.5.2 Normalisation and batch correction

To remove technical variation across multiple batches, Dr. Jong Kyoung Kim applied location and scale adjustments to the normalized read counts by using the ComBat function of the sva package of R with default options (Johnson et al., 2007). He first $\log_{10}$-transformed the normalized read counts (after removing lowly expressed genes whose mean normalised read counts are less than 10) and after adding a pseudo count of 1. Secondly, he adjusted for batch effects using ComBat with the known batch covariate and sample conditions. Finally, he re-transformed the batch-adjusted expression values x back to the original scale (10^{x-1}). If the re-transformed values were less than 0 or the original read counts are 0, we set the re-transformed values to 0.

2.5.3 Quality control of cells

To exclude poor quality libraries from downstream analysis, first I removed cells that correspond to empty capture sites, capture sites with multiple cells, or capture sites containing cell debris on the C1 chip by visually inspecting them under microscope. Second, it has been known that some cells suffer from cell rupture during the process of microfluidic cell capture (Islam et al., 2014). To identify these abnormal cells, I calculated two quantities for each cell: the number of reads mapped to exons, and the proportion of reads (of all reads mapped to exons) mapped to 37 genes on the mitochondrial chromosome. I identified two populations of cells in terms of the above two quantities and most of the cells corresponding to empty cells or cell debris are in one of the two populations. Biologically when cell is ruptured cytoplasm leaks out and there is a relative increase in abundance of transcripts that are enclosed within the mitochondria. Based on this, I set the following criteria to remove abnormal cells:

1) Cells that have fewer than 500,000 reads mapped to exons

2) Cells that have greater than 10% reads mapped to mitochondrial genes

Finally, I compared the normalised read counts of genes between cells in the same condition, and found that in one cell (cell "85" in the first replicate of serum) there was a problem in library preparation and many genes were abnormally amplified (Figure 2.5). I removed the cell from further analysis. In summary, I have the following number of cells for the analysis: 81, 90 and 79 for serum replicates; 82, 59, 72 and 82 for 2i replicates, 93 and 66 for a2i replicates, where the total number of cells across conditions is 704.

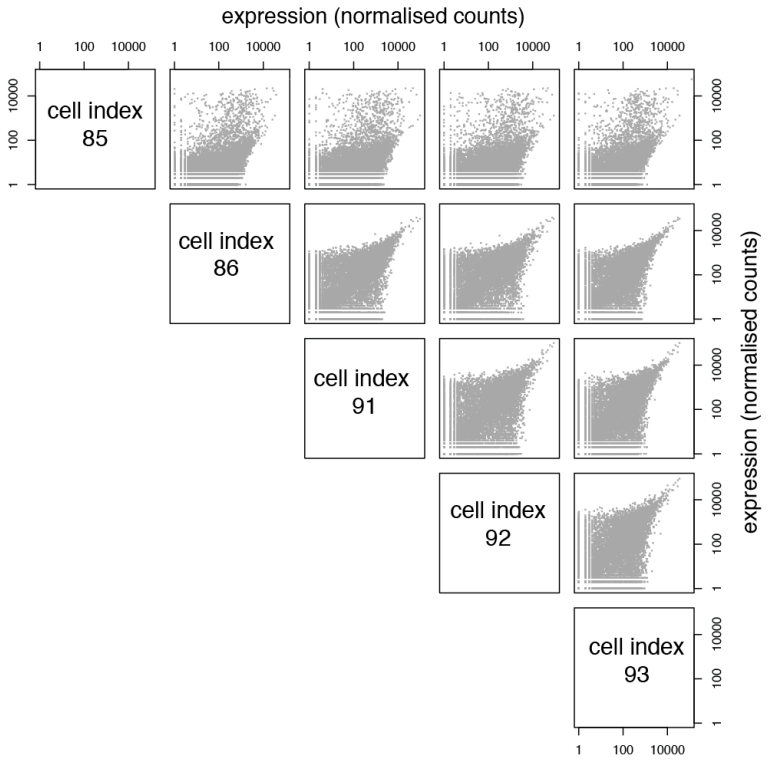

Figure 2.5 Correlation of gene expression levels in single cells.
Expression levels correlate with each other as shown representatively for cells index 86, 91, 92 and 93 (serum 1). Cell 85, as it is substantially different than any other cell, suggesting failure of the experimental protocol.

2.5.4 Calculating DM as a measure of noise

To account for the confounding effects of gene length and mean expression level on the CV, Dr. Jong Kyoung Kim computed the DM values for each gene using rolling medians of the squared CV. First, he computed gene lengths by taking the union of all exons within a gene based on the Ensembl annotation. He excluded all exons annotated as "retained_intron". He also removed lowly expressed genes whose mean normalised read counts (reads per million) are less than 10, since we cannot distinguish biological noise from technical noise for these genes. Second, he computed rolling medians from the scatter plot between the mean normalised read counts and the squared CV values, where

the x- and y-axis are log10 transformed. Third, we calculated the mean-corrected residual of the squared CV of gene i to its corresponding rolling median $f(i)$ such that

$$r(i) = \log_{10} CV(i)^2 - f(i).$$

Finally, to correct for the effect of gene length on the mean corrected residual, he calculated the difference between the mean corrected residual of the squared CV of gene i and its expected residual by using the following formula

$$DM(i) = r(i) - g(i),$$

where $g(i)$ is the rolling median of gene i from the scatter plot between $r(i)$ and $\log_{10}$ transformed gene lengths. To compute the rolling medians, he used the rollapply function of the zoo package of R (Zeileis and Grothendieck, 2005) and the following parameters: the number of genes in the window is 50 and the number of overlapping genes between adjacent windows is 25. This relative noise measure, which is referred to as DM, does not depend on either gene expression levels or gene lengths (Spearman's ϱ=0.0200 for gene expression levels and ϱ=0.0206 for gene length in the serum condition) (Kolodziejczyk et al., 2015b).

2.5.5 Testing the absolute level of cell-to-cell variation of a functional category within a culture condition

To test whether genes belonging to a defined functional category have a high or low level of expression heterogeneity within a culture condition, Dr. Jong Kyoung Kim performed gene set enrichment analysis using the Piano package of Bioconductor (Varemo et al., 2013). He used the DM values for gene-level statistics and calculated the mean DM values as a gene-set statistic

for each GO term. The associations between Ensembl gene IDs and GO terms were obtained from the biomaRt package of Bioconductor (Kasprzyk, 2011). Since gene set enrichment analysis tends to bias towards large or small categories in terms of their number of genes, he considered only gene sets with between 3 and 2,000 genes. The P-value for each GO term was then computed by randomly taking a set of genes of the same size as in the GO term, and by repeating this 10,000 times.

2.5.6 Testing the relative difference in expression heterogeneity of a functional category across culture conditions

To explore further the difference of the three culture conditions in terms of gene expression noise, Dr. Jong Kyoung Kim compared two sets of DM values for each GO term between two culture conditions using the two-sided paired t-test. He only considered GO terms with at least 2 genes having DM values. The associations between GO terms and their offspring terms were obtained from the GO.db annotation package of Bioconductor

(http://www.bioconductor.org/packages/release/data/annotation/html/GO.db.html).

2.5.7 Differential expression analysis

I identified differentially expressed genes from bulk data and single cell data using the DESeq package (Anders and Huber, 2010). I considered genes that differed in expression by two-fold and with a multiple testing adjusted *p*-value was < 0.05 to be differentially expressed. For single cell differential expression analysis I used each as a replicate of the condition it came from and I removed genes that had mean expression below 50 counts.

2.6 Doubling time estimation of mouse embryonic stem cells in different conditions

Fifty thousands G4 mouse ES cells were plated by Dr. Jason Tsang in single wells on gelatinized 6-well plates, and maintained in the three culture conditions of interest (total 12 wells for each culture condition): serum-containing media, standard 2i media and alternative 2i media. Three wells were harvested and quantified on a haemocytometer every 24 hours for 4 days to estimate the doubling time of mouse ES cells in each condition.

2.7 Datasets

	Generated by:	Data accession numbers
Single cell mRNA seq data of mESC cultured in three conditions (2i, a2i, serum)	Kolodziejczyk et al., 2015, Cell Stem Cell	Array Express E-MTAB-2600
Single cell mRNA seq data of early mouse embryo development	Deng et al., 2014, Science	Gene Expression Omnibus GSE45719
2C-like cell gene expression profiles (microarray data)	Macfarlan et al., 2012 Nature	DE count tables from Supplementary Table 4
NPC differentiation time course	Dr. Alex Tuck (unpublished)	unpublished