Supplementary Information

Supplementary Methods

Cell culture and Secondary Reprogramming

ESCs and iPSCs were cultured in 5% CO_2 at 37°C on irradiated MEFs in DMEM containing 15% FCS, leukemia-inhibiting factor, penicillin/streptomycin, l-glutamine, nonessential amino acids, sodium pyruvate, and 2-mercaptoethanol. 1B 1° iPS cells were aggregated with tetraploid host CD-1 embryos as described[43] (in compliance with Protocol # 009 at the Toronto Centre for Phenogenomics) and MEFs were established from E13.5 embryos [43]. High doxycycline cell samples (1500ng/mL dox) were collected at days 0, 2, 5, 8, 11, 16 and 18 (D2H, D5H, D8H, D11H, D16H, D18H). A subculture of the reprogramming cells was established from day 19 and cultured in the absence of dox, to develop a factor independent 2° iPS cell line by day 30 (2°iPSC). Low dox samples were maintained from day 8 to day 14 cells in 5ng/mL dox. At day 14 the culture diverged to two groups, with one group of the cells being cultured until day 21 in the absence of dox (D21Ø) and the other cultured in 5ng/mL of dox and collected at day 16 (D16L) and (D21L). ROSA26-rtTA-IRES-GFP mouse ES cells[12], and 1B 1° iPSCs[10] were collected as controls. All cell lines have been tested for mycoplasma and other pathogens.

Long RNA Sequencing and Alignment

Cells were scraped, harvested in ice cold PBS and stored in RNA-later (Ambion) at -80°C. Total RNA for transcriptome sequencing was prepared using Qiagen total RNA purification kit followed by two rounds of on column DNAseI treatment to remove contaminating DNA using the RNase-Free DNase set (Qiagen PN 79254) as per manufacturer's protocol. The total RNA was then analyzed using Agilent RNA 6000 Nano Kit (PN 5067-1511) on the Agilent Bioanalyzer 2100 (PN G2939AA) to quantify yield, qualify integrity and confirm removal of DNA contamination.

Following DNAseI treatment, 5µg total RNA from each sample was depleted of Ribosomal RNA using the Ribo-Zero™rRNA Removal Kit (Epicenter PN RZH110424) as per manufacturer's instructions. The rRNA depleted RNA was then run on an Agilent RNA 6000 Pico Kit (PN 5067-1513) on the Agilent Bioanalyzer 2100 to confirm rRNA depletion. Sequencing libraries where generated from the rRNA depleted RNA using the SOLiD™ Transcriptome Multiplexing Kit (PN 4427046) from Applied Biosystems following the manufacturer's publication. Final libraries were quantified and qualified using Agilent High Sensitivity DNA Kit (PN 5067-4626) on the Agilent Bioanalyzer 2100.

Sequencing libraries were subsequently pooled in equimolar ratios (four libraries per pool) and clonally amplified onto SOLiD Nanobeads. Clonal amplification was completed via emulsion PCR using the SOLiD EZ Bead System (PN 4448419, 4448418 and 4448420) coupled with SOLiD EZ Bead N200 amplification reagents (PN 4467267, 4457185, 4467281, 4467283, 4467282). Following emulsion PCR clonally amplified Nanobeads were enriched using the SOLiD EZ Bead Enricher Kits (4467276, 4444140, 4453073) before being deposited into SOLiD™ 6-Lane FlowChip (PN 4461826) using the SOLiD Flowchip Deposition Kit v2 (PN 4468081) as per the manufacturer's recommendations.

In total two flowchips were sequenced yielding a total of 8 lanes of data; with sequencing reads generated using the SOLiD 5500xl platform generating paired 75bp forward and 35bp reverse reads. To allow de-convolution of the pooled libraries a single 5bp index read was generated. A total of 1,204,676,394 fragments (2,409,352,788 reads) were generated post de-convolution, ranging from 35,714,748 to 147,282,580 fragments per library (**Supplementary Table 6**).

Sequence mapping was performed using Applied Biosystems LifeScope v2.5 whole transcriptome (paired-end) analysis pipeline against the NCBIM37 (mm9) genome and exon-junction libraries constructed from the Ensembl v64 gene model. Briefly, this pipeline first removes potential contaminant reads by aligning to a filter set containing rRNA, tRNA, adaptor sequences and retrotransposon sequences. Following filtering, LifeScope then aligns all reads to the genome and F3 reads to the junction library. F5 reads are additionally aligned at a higher sensitivity to exonic sequences within insert size distance from the paired (F3) read alignment. For RNAseq datasets from O'Malley *et al.* (2013)[8] (accession number E-MTAB-1654 in ArrayExpress) and Golipour *et al.* (2012)[7](GEO accession number GSE42100), Tophat (version 2.0.6) was used to map reads against the NCBIM37 (mm9) genome (Ensembl v67). Read alignments were merged and disambiguated, and a single BAM (Binary Alignment Mapped) file output per library or sample was used. BAM files were then additionally filtered to remove reads with a mapping quality (MAPQ) < 13, and all ribosomal and mitochondrial RNA reads. Alignments were assembled using Cufflinks (v2.1.1) using the –g parameter to construct a genome annotation file against the reference gene model (Ensembl v67) and to identify novel transcripts (refer to **Long RNA sequencing analysis pipeline** below for details).

Long RNA sequencing analysis pipeline:

- Run Lifescope

    - Align to filter set
    - Align to genome
    - Align to exon junction
    - Choose alignments
        - Refer to **Supplementary Table 6** for read counts.
- Remove reads < 13 MAPQ
- Remove chrMT reads
- Assemble with Cufflinks to create annotation file
    - Sequences assembled using -g parameter against Ensembl v67
- FPKM values are then calculated as detailed below.
- FPKM values for each gene are used for all subsequent analyses.

Read count and Differential gene expression

Raw read counts were obtained by mapping reads at the gene level using the Cufflinks assembled transcript annotation file with HTseq-count tool from the Python package HTSeq, at http://www-huber.embl.de/users/anders/HTSeq/doc/count.html, using Intersection-nonempty counting mode. EdgeR R-package (v3.0.8)[44] was then used to perform statistical analysis, samples were grouped as follows:

- Group 1: 2°MEF, group 2: D2H, group 3: D5H-D11H, group 4: D16H-D18H, D16L, group 5: D21L and D21ø, group 6: 2°iPSC, ESC, and 1°iPSC

A common biological coefficient of variation (BCV) and dispersion (variance) was estimated for each grouping scenario based on a negative binomial distribution model. The estimated dispersion values were combined to obtain a final BCV value. This value was then incorporated into the final EdgeR analysis for differential gene expression, and the exact test for negative binomial distribution was used for statistics, as described in EdgeR user guide.

Identification and characterization of novel transcripts

Transcripts that did not overlap with ensembl annotations were selected as candidate novel lncRNA genes. We only considered genes that fell under the following criteria: 1.) length of 200bp or more from cufflinks assembled transcriptome, 2.) intergenic transcripts, 3.) novel antisense transcripts, and 4.) novel transcripts that overlapped intergenic miRNAs. Novel transcript overlapping annotated genes or novel isoforms as defined by Cufflinks output were not considered. We used Coding Potential Calculator (CPC)[45] to calculate the likelihood of any of these transcripts to be part of a coding protein sequence (i.e. coding potential score). CPC accounts for quality and length of open reading frame, start and in-frame stop codons, and sequence homology with known protein coding genes. Transcripts with a negative coding potential score were considered non-coding (**Extended Data Fig. 9d**). To identify multi-exonic novel transcripts, we relied on H3K4me3/H3K36me3 chromatin domains derived from our ChIPseq dataset, as previously described[41], to determine whether single transcripts with the same orientation as identified by Cufflinks were exons of a larger novel transcript. If these single transcripts were within an H3K4me3/H3K36me3 chromatin domain and showed a similar expression pattern, they were considered putative exons of one novel transcript.

Transcriptional start site (TSS)/gene promoter identification and FPKM calculation

To properly identify gene promoters and gene size for Fragments per Kilobase per million reads (FPKM) calculations, we first examined all possible annotated gene isoforms identified by Cufflinks and novel transcripts that passed the criteria described above, and restricted our analysis as follows:

1.) We divided every exon identified by Cufflinks assembled transcriptome into bins based on exon boundaries derived from all isoforms of a gene, identical to the method employed in DEXseq R package for differential exon usage[46]. We then mapped reads to these features using HTseq-count tool in Intersection-nonempty counting mode. EdgeR R-package (v3.0.8)[44] was then used to normalize the data and calculate counts per million reads (CPM) values.

2.) For annotated isoforms, we only considered isoforms that showed at least 10 reads or a value of 0.2 CPM in the first exon bin in at least two samples, except when reads were only detected in 2°MEFs. Isoforms that failed to show expression in their corresponding first exon bin were filtered out.

3.) If multiple isoforms were detected based on the first exon bin strategy above, we examined the number of reads in the subsequent exon bin, sequentially scanning each exon bin of a gene for read count. Any exon bin that failed to have at least 10 reads or a value of 0.2 CPM was excluded along with its corresponding isoform. We followed this strategy until we identified the most abundantly expressed isoform per gene. We were able to robustly detect the proper isoform, with the exception of a few cases described below. The most abundant isoform size and TSS were used for FPKM calculations (FPKM =1000 * CPM/(size of gene)) and subsequent analysis.

4.) The low number of genes, where this strategy failed, were genes with very low level expression. For such genes, we used Ensembl NCBI37/mm9 annotated mouse reference for TSS coordinates and calculated the gene size based on the sum of exon sizes where an expressed exon bin could be detected, and incorporated this into our FPKM calculations. In cases where a gene was not expressed in any of the samples, we used Ensembl NCBI37/mm9 annotated mouse reference for gene and TSS coordinates.

5.) For novel transcripts, we used annotations and gene sizes defined by Cufflinks *de novo* assembly, except in cases where we defined a novel transcript as multi-exonic. In these cases, the TSS of the first transcript (i.e. first exon) within a novel multi-exonic gene was considered to be the start site, and the sum of transcript sizes as the final gene size.

Identification of stage-specific genes

To identify stage-specific genes, we used Shannon entropy modeling to compute a stage-specificity index for each gene, as previously described [18,47]. Briefly, for each gene, a relative expression (Ri) value was calculated per sample and per grouping as described above and in **Extended Data Fig. 3a**, where Ri = (Expression per sample or average expression per group)/sum of FPKM values in all the samples or all the groups. The entropy index score (Hi) across all samples or groups was calculated as follows:

- $Hi = -1 * \Sigma (Ri * \log_2(Ri))$

An entropy score close to 0 indicates high stage specificity, whereas, a score closer to $\log_2$ of the total number of samples (13) or groups (6) indicates ubiquitous expression. As a threshold for selecting candidate genes for stage-specificity, we used a 10th percentile cutoff (indicated by dashed lines in **Extended Data Figs 4a** and **9b**) of the entropy scores distribution curve. Genes below this threshold were considered stage specific.

Analysis of repeat elements

We downloaded the RepeatMasker annotation file from the UCSC genome browser. We excluded repeats that overlapped Ensembl NCBI37/mm9 annotated mouse reference genes and considered repeats that are 200bp or greater in length. Reads were mapped to these features using HTseq-count tool in Intersection-nonempty counting mode. EdgeR R-package (v3.0.8)[44] was then used to normalize the data and calculate CPM values. CPM values were then divided by the length of the repeats and multiplied by 1000 to obtain FPKM values. Repeats with values > 0.5 FPKM were considered expressed.

Calculation of transgene versus endogenous gene expression for Yamanaka factors

To obtain endogenous expression of the reprogramming factors, we followed two strategies:

1.) We mapped reads to the 5'UTR and 3'UTR for Sox2, Klf4 and c-Myc using these coordinates:
    a. Sox2: 5'UTR-chr3:34548929-34549232, 3'UTR-chr3:34550301-34551414.
    b. Klf4: 5'UTR - chr4:55544734-55545078, 3'UTR - chr4:55540033-55540942.
    c. Myc: 5'UTR - chr15:61817049-61819045, 3'UTR - chr15:61821469-61821815
2.) For Pou5f1 (also known as Oct4), we identified a C/T single nucleotide polymorphism (SNP) at chr17:35643135 that differentiates between endogenous (C/G base pair) and exogenous (T/A base pair) expression. By mapping reads to the different polymorphisms, we quantified the relative levels of exogenous versus endogenous expression.

Library normalized read counts (CPM values) obtained from endogenous locations were further scaled for comparison to total CPM values. Scaling factors for each reprogramming factor were calculated from 4 samples (2°MEF, ESC, 2° and 1° iPSCs) as follows:

- $$\text{Scaling factor} = \frac{\text{(Total expression of reprogramming factor in sample)}}{\text{(Endogenous expression of reprogramming factor in sample)}}$$

The scaling factor was averaged over the 4 samples and used to scale up endogenous CPM values for all samples. Exogenous reprogramming factor expression was determined as the difference between total and scaled endogenous expression.

Calculation of Intron Retention (IR)

Data were mapped to Ensembl assembly Mus_musculus.GRCm38.74. The same build was used to define gene structures.

The intron retention ratio was calculated for each intron as: depth of intron cover / (spliced reads + depth of intron cover).

Introns were listed as having significant IR if meeting the following conditions:

- All samples had >5 reads correctly spliced across the intron; one sample had at least 20 reads.
- One sample had reads covering >90% of non-excluded bases within the intron.
- One sample had reads supporting continuation from exon into intron at both ends with minimum of 5bp overhang.

- To ensure readings were above any background, intron read depth was at least 25% greater than any neighboring introns, or the neighboring intron itself had been determined to have significant IR.

The mean of introns with significant IR within that gene was calculated for each gene.

The following regions were excluded from intron retention analysis:

- Intronic regions that overlapped with: exons, lncRNA and all other non-intron annotated features.
- Regions of poor-mappability were excluded from statistics.
- Introns were excluded where a feature of opposite sense intersected.
- Introns were excluded if more than 30% of bases had been excluded or the length was less than 120bp.

Introns contain numerous repeat and low complexity regions to which software cannot uniquely map them. These regions of low mappability cause artificial "valleys" of expression where the number of mapped reads drops close to zero. These valleys occur frequently in introns and lead to an underestimation of IR. To compensate for this we created a mappability index to correct these artificial valleys of expression and normalize intronic expression in low mappability regions. This index was calculated based on the Mus_musculus.GRCm38.74 reference genome. A sliding window of 40bp and of step 10bp was used to tile the reference genome. The genomic sequence in each window was extracted, a one base random mis-read was substituted and prepared in the format of sequencing reads (fastq). These artificially generated reads were then mapped against the reference genome using the same parameters used for the input mRNA-Seq data. Regions with resultant coverage at, or poorer than 50% were considered to have poor-mappability.

Reads were prepared by trimming adapters with a custom paired-end aware colour-space trimmer. Reads were mapped in single-end mode with CUSHAW3 allowing multi-mapping against a combined genome and junction transcriptome; the junction transcriptome was built with USeq MakeTranscriptome, allowing mapping across canonical and non-canonical combinations of known splice sites within genes. For each read pair, a unique correctly paired read was selected by custom code on best match measured by: direction, distance separating reads, and mismatch count.

The depth of spliced read-pairs was counted at both ends of the intron and the maximum taken; reads with at least 5bp overhanging the splice site in both directions were considered.

The depth of intron cover was calculated from non-excluded bases within the intron. A trimmed mean of depth of cover was then calculated, including the center 20% of values. All counts were performed with Bedtools. Coverage was assessed per-molecule where read pairs had no more than 120bp separation.

Spearman correlation coefficients were calculated for each gene between IR values and their corresponding RNA-seq FPKM expression values across the 13 samples. For determining the effect of IR during the reprogramming stages described in **Extended Data Fig. 6e**, we performed Pearson Correlation analysis between IR values and their corresponding FPKM expression across a minimum of 4 samples. We also performed a similar correlation by randomizing gene expression values 10 times to IR values for each indicated reprogramming stage to obtain the random level of correlation between IR and gene expression. This was used to calculate statistical significance.

Microarray data processing

Affymetrix HT Mouse Genome 430A microarray data from Polo et al. (2012)[6] (GEO accession number GSE42379) and Illumina MouseWG-6 v2.0 expression beadchip array data from Kojima et al. (2013)[20] (GEO accession number GSE46227) were analyzed using R application and limma R package (v3.14.4). The probe intensity data were log transformed and quantile normalized and unannotated probes were removed.

miRNA sequencing and alignment

miRNA purification was performed according to the miRvana miRNA isolation kit (Ambion #1560) and quality validated using on a Bioanalyser before sequence library preparation. Small RNA libraries were prepared for SOLiD™ next generation sequencing, with libraries sequenced to a depth of 27,420,558-118,946,232 tags (average 55,816,766 tags; up to 35 nucleotides in length), yielding a total of 725,617,952 tags (**Supplementary Table 6**). These tags were then mapped to the mouse genome (NCBI37/mm9 assembly) and miRNA-mapped tags determined as those overlapping with known miRNA loci (miRbase v18). Thus, using the tools and parameters detailed below, we were able to map 347,190,702 tags across the thirteen libraries (47% of tags) (refer to **Small RNA sequencing analysis pipeline** below for details).

Small RNA sequencing analysis pipeline:

- Identify and remove the adaptor sequence (maximum 25% mismatch with adaptor sequence).

- Retain tags with at least 20nt length, and at least 18 mean quality across the tag.
- Map tags to the mouse genome (mm9, NCBI37) and rRNA sequences using Bowtie aligner:
    - Command:
      ```
      bowtie -f -C -Q Sample,CV,qual --integers-quals -l 20 --nomaqround --maxbts 800 -y --chunkmbs 2048 -M -a --best --strata --snpfrac 0.01 --col-cqual --col-keepends --sam --mapq 20 --offrate 2 --threads 12 --shmem ReferenceBowtieIndex.fa Sample.csfasta
      ```
    - Version: 0.12.8
    - Reference: Mouse Genome (mm9 assembly), 18S rRNA (gi|374088232), and 28S rRNA (gi|120444900)
        - Refer to **Supplementary Table 6** for tag counts.
- Count the number of tags that overlap annotated miRNA as defined in miRNase version 18 (Tag length set between 20 and 26nt). For example:
    - miRBase annotates mature miRNA mmu-miR-XYZ on chromosome 1, starting at position 1,347 on the sense strand. All tags comply with the criteria below are assumed to mmu-miR-XYZ miRNA tags:
        - 20-26nt long
        - map to the sense strand of chromosome 1
        - start position between 1,344 and 1,350 inclusive (1347 +/- 3)
- Scale the number of tags assigned to each miRNA to correct for different library sizes. For example:
    - Total tags mapped to miRNA loci in 1°iPSC library = 9,618,934 and total tags mapped to miRNA loci in 2°iPSC library = 9,107,222, then a miRNA expression value is calculated as follows:
        - If number of tags mapped to mmu-let-7a-5p miRNA in 1°iPSC library is 36,868 then:
            - Number of tags mapped to mmu-let-7a-5p miRNA in 1°iPSC library after correcting for library size is: (36,868 / 9,618,934) * 1,000,000 = 3,832.86
        - If number of tags mapped to mmu-let-7a-5p miRNA in 2°iPSC library is 47,890 then:
            - Number of tags mapped to mmu-let-7a-5p miRNA in 2°iPSC library after correcting for library size is: (47,890 / 9,107,222) * 1,000,000 = 5,258.46
- Re-scale the library size using TMM method to compensate for sequencing real-estate effect[48].
- Normalized counts for each miRNA are used for differential expression analysis.

Chromatin Immunoprecipitation sequencing (ChIP-Seq)

ChIP Library Generation. ChIP was carried out as described in [49]. 40-150 million cells were fixed with 1% formaldehyde for 10min at room temperature, scraped and stored as pellets (-80^{o}C). Samples were lysed at 20 million cells/mL in Farnham lysis buffer for 10 min followed by 10 million cells/mL in nuclear lysis buffer. The released chromatin was sheared to 100-500bp (250bp average) on ice using a SonicsVibraCell Sonicator equipped with a 3mm probe. For each sample, 50µL of solubilized chromatin was used as input DNA to normalize sequencing results and the remaining chromatin was immunoprecipitated with 10µg of H3K4me3 (ab8580)[50], 10µg H3K27me3 (Millipore 07-

449)[51] or 10μg H3K36me3 (ab9050)[51] antibodies, separately. Antibody-chromatin complexes were pulled down with 100μL magnetic Protein G Dynal beads (Invitrogen) and washed six times. The chromatin was then eluted, reverse cross-linked at 65°C overnight and subjected to RNaseA / proteinase K treatment. ChIP and input DNA was purified using a Qiagen Purification Column and quantified using a Quant-it dsDNA High Sensitivity Assay (Invitrogen).

High-Throughput Sequencing. Sequencing libraries were prepared according to Illumina ChIP-seq Library Preparation kit instructions. 50ng of immunoprecipitated or input DNA was end-repaired, followed by the 3′ addition of a single adenosine nucleotide and ligation to universal library adapters. Ligated material was separated on a 2.0% agarose gel, followed by the excision of a 250–350bp fragment and column purification using Qiagen gel purification kit. DNA libraries were prepared by PCR amplification (18 cycles). ChIP DNA libraries were sequenced using the Illumina HiSeq 2000 as per the manufacturer's instructions. Sequencing libraries was performed up to 2 x 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0.

Processing and alignment of ChIP-Seq data to identify H3K4me3, H3K27me3, and H3K36me3 enriched peaks. ChIP-Seq sequencing data was processed using the Illumina analysis pipeline and FastQ format reads were aligned to the NCBI37/mm9 mouse reference using the Bowtie alignment algorithm[52]. Bowtie version 2.1.0 was used with the preset sensitive parameter to align ChIP sequencing reads from this study (refer to **ChIP sequencing analysis pipeline** below for more details and **Supplementary Table 6** for read counts) and ChIPseq dataset from Polo et al. (2012)[6] (GEO accession number GSE42477).

Peak Calling Algorithm. The MACS version 2.0.10 (Model based analysis of ChIP-Seq)[53] peak finding algorithm was used to identify regions of ChIP-Seq enrichment over background. Default parameters were used for H3K4me3, and broad peak parameters were used for H3K27me3 and H3K36me3 data (refer to **ChIP sequencing analysis pipeline** below for details).

Peak Annotation and Processing.

Multicov command from Bedtools v2.17.0 was used to obtain raw read counts within each histone mark peak identified by MACS and input reads within these peaks. The number of reads per kilobase of peak per million reads (RPKM) was calculated for each peak and the corresponding input levels of that peak. The RPKM values for the histone mark peak were then subtracted by those of the input RPKM values to obtain a final and background adjusted RPKM value, as modified from[54]. Peak calls with background adjusted RPKM values less than or equal to 0 were excluded from further analysis. The background adjusted RPKM values were averaged across -2kb to +3kb of a gene TSS, as determined above, for downstream data analysis and visualization. Gene loci with an average background adjusted RPKM values less than 0.5 were considered negative for the presence of the histone mark. ngs.plot.r[55] software was used to generate read density heatmaps and profiles. Read densities and enrichment scores per locus, where defined, was normalized to the total number of million uniquely mapped reads producing values in units of reads per million mapped reads (RPM).

***Identification of differential histone mark changes associated with* Fig. 3c-g**. To determine a histone mark change during reprogramming, as shown in **Fig. 3c-g**, we first applied the following criteria for transcriptionally active and silent loci identification:

- Active locus:
    - H3K4me3^{+}/H3K27me3^{-}/H3K36me3$^{(+/-)}$, and gene expression values of log$_2$(FPKM) >= 0.7226907 for protein coding genes or log$_2$(FPKM) >= -1.515307 for lncRNAs, as determined in **Extended Data Figs 8a** and **9a**, respectively.
- Silent locus:
    - H3K4me3$^{(+/-)}$/H3K27me3^{+}/H3K36me3^{-}, and gene expression values of log$_2$(FPKM) < 0.7226907 for protein coding genes or log$_2$(FPKM) < -1.515307 for lncRNAs.
    - H3K4me3^{-}/H3K27me3^{-}/H3K36me3^{-} (i.e. no mark), and log$_2$(FPKM) < 0.7226907 for protein coding genes or log$_2$(FPKM) < -1.515307 for lncRNAs.

Only histone marks that follow the above-described criteria were considered for further analysis.

We next grouped samples as follows:

- Group 1: 2°MEF, group 2: D2H, group 3: D5H-D11H, group 4: D16H-D18H, D16L, group 5: D21L and D21ø, group 6: 2°iPSC, ESC, and 1°iPSC

We then only examined histone mark modifications where a change was observed from 2°MEF to a minimum of 2 samples from within group 3, 4 or 6. In cases where a gene switched transcriptional activity, for example, changing from active to silent or vice versa, our analysis only focused on genes showing stage-specific expression by RNA-seq.

ChIP sequencing analysis pipeline:

- Trim Sequence (filter out 3' adaptor, and remove last 2 bases and 3 extra bases if it matches with adaptor sequence).
- Mapping sequences to mouse genome (mm9/NCBI37) using Bowtie
    - Command: bowtie2 -p 8 --sensitive -x mm9/mm9 -1 sequence.reads_R1.fastq -2

sequence.reads _R2.fastq –S sample.sam

    - Refer to **Supplementary Table 6** for read counts.

- Peak calling algorithm MACS
  - Command for H3K4me3: macs2 callpeak -t chromatin.mark.file.bam -c input.sample.file.bam -f BAMPE -g mm -n [directory] --nomodel --shiftsize 73 -B
  - Command for H3K27me3 and H3K36me3: macs2 callpeak –t chromatin.mark.file.bam –c input.sample.file.bam --broad –f BAMPE -g mm -n [directory] –nomodel --shiftsize 73 –B
    - Refer to **Supplementary Table 6** for read counts.
- Normalize unique mapped read values to library size
- Annotate peaks to mouse genome (mm9/NCBI37)

DNA Methylation Analysis

MethylC-Seq Library Generation. Five micrograms of genomic DNA was mixed with unmethylated cl857 Sam7 Lambda DNA (Promega, Madison, WI, USA). The DNA was fragmented by sonication to 300–500bp with a Covaris S2 system (Covaris) followed by end repair with the End-It DNA End-Repair Kit (Epicenter). Paired-end universal library adaptors provided by Illumina (Illumina) were ligated to the sonicated DNA as per manufacturer's instructions for genomic DNA library construction. Ligated products were purified with AMPure XP beads (Beckman, Brea, CA). Adaptor-ligated DNA was bisulfite treated using the EpiTect Bisulfite Kit (QIAGEN) following the manufacturer's instructions and then PCR amplified using PfuTurboCx Hotstart DNA polymerase (Agilent, Santa Clara, CA) with the following PCR conditions (2min at 95°C, 4 cycles of 15s at 98°C, 30s at 60°C, 4min at 72°C then 10min at 72°C). The reaction products were purified using the MinElute gel purification kit (QIAGEN). The sodium bisulfite nonconversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the lambda genome.

High-Throughput Sequencing. MethylC-Seq DNA libraries were sequenced using the Illumina HiSeq 2000 as per manufacturer's instructions. Sequencing was performed up to 2x 101cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0.

Processing and alignment of MethylC-Seq data to identify methylated cytosines. MethylC-Seq sequencing data was processed using the Illumina analysis pipeline and FastQ format reads were aligned to the NCBI37/mm9 mouse reference using the Bismark/Bowtie alignment algorithm [52,56] Paired-read MethylC-Seq sequences produced by the Illumina pipeline in FastQ format were trimmed with trim threshold 1500, which removed the last 2 bases from sequences that were not trimmed, and removed 3 bases from sequences that were trimmed. The Bismark package version 0.7.7 was used as the aligner (refer to **Methylome sequencing analysis pipeline** below for more details).

Since up to six independent libraries from each biological replicate were sequenced, we first removed duplicate reads. Subsequently, the reads from all libraries of a particular sample were combined. Unique read alignments were then subjected to post-processing. The number of calls for each base at every reference sequence position and on each strand was calculated. All results of aligning a read to both the Watson and Crick converted genome sequences were combined. The CpG methylation levels were calculated using bisulfite conversion rates by (Number of not converted Cs/ read depth) for each position.

Identification of methylated cytosines. At each reference cytosine the binomial distribution was used to identify whether at least a subset of the genomes within the sample were methylated, using a 0.01 FDR corrected P-value. We identified methyl cytosines while keeping the number of false positives methylcytosine calls below 1% of the total number of methyl cytosines we identified. The probability p in the binomial distribution B (n,p.) was estimated from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome (referred to as the error rate: non-conversion plus sequencing error frequency). We interrogated the sequenced bases at each reference cytosine position one at a time, where read depth refers to the number of reads covering that position. For each position, the number of trials (n) in the binomial distribution was the read depth. For each possible value of n we calculated the number of cytosines sequenced (k) at which the probability of sequencing k cytosines out of n trials with an error rate of p was less than the value M, where M* (number of unmethylated cytosines) < 0.01* (number of methylated cytosines) and if the error rate of p was over 0.01, we assumed the cytosine was not methylated. In this way, we established the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated, so that out of all methyl cytosines identified no more than 1% would be due to the error rate.

Identification of differentially methylated regions (DMRs). DMRs were identified using sliding window approach of 300bp, sliding every 30bp. Windows showing differences above 45% between any sample and a minimum of 5 CpGs were considered differentially methylated. 131540 differentially methylated windows were identified. Differentially methylated windows were merged to obtain an average methylation level or differential methylation value, relative to 2°MEF, per annotated gene locus. Analysis was confined to -1kb to +1kb region of TSS as we found this to be the region frequently spanning hypo-methylation for key ESC genes.

<u>Methylome sequencing analysis pipeline</u>:

- Trim Sequence (filter out 3' adaptor, and remove last 2 bases and 3 extra bases if it matches with adaptor sequence).

- Mapping sequences to mouse genome (mm9/NCBI37) using Bismark/Bowtie using the following parameters:
  - Command: -e 90 -n 2 -l 32 -X 550
  - Sequence reads are first transformed into fully bisulfite-converted forward (C->T) and reverse read (G->A conversion of the forward strand) versions. They are then aligned to similarly converted versions of the genome (also C->T and G->A converted). Sequence reads that produce a unique best alignment from the four alignment processes against the bisulfite genomes (which are running in parallel) are then compared to the normal genomic sequence and the methylation state of all cytosine positions in the read is inferred. A read is considered to align uniquely if one alignment exists that has fewer mismatches to the genome than any other alignment (or if there is no other alignment).
    - Refer to **Supplementary Table 6** for read counts and methylated Cytosine distribution.
- Remove duplicates.
- Calculate base-by-base methylation level and final CpG methylation counts.
    - Refer to **Supplementary Table 6** for methylated CpG counts.
- Integrate CpG methylation level of positive and negative strand.
- Adjust methylation level using bisulfite conversion rate from unmethylated Lambda control.

Global Proteomics

Sample preparation for MS analysis. Cells were harvested by centrifugation and lysed in 8 M urea (100mM triethyl ammonium bicarbonate, pH 8.2, with protease and phosphatase inhibitors). Proteins (~1 mg) were first reduced/alkylated and digested for 4 h with Lys-C. The mixture was then diluted 4-fold to 2M urea and digested overnight with sequencing grade trypsin (Promega) in substrate/enzyme ratio of 50:1 (w/w). Digestion was quenched by acidification with formic acid (FA) (final concentration 10%). Resulting peptides were subsequently desalted by solid phase extraction (Sep-pack Vac C18 cartridges, Waters), dried down and then re-suspended in TEAB buffer 100mM to a final concentration of ~1 mg/ml. An aliquot of 100µg of each sample was chemically labeled with Tandem Mass Tag (TMT) reagents (Thermo Fisher) according to the manufacturer instructions. Data for all samples were normalized to an internal standard (ISTD) made up of equal proportions of the samples (refer to **Global Proteome analysis pipeline** below for details). Before the mass spectrometric analysis, both the TMT labeled peptides mixtures were fractionated as described elsewhere [57]. The SCX system consisted of an Agilent 1200 HPLC system (Agilent Technologies, Waldbronn, Germany) with one C18 Opti-Lynx (Optimized Technologies, OR) trapping cartridges and a Zorbax BioSCX-Series II column (0.8mm inner diameter 50mm length, 3.5mm). The labeled peptides were dissolved in 10% FA and loaded onto the trap columns at 100µl/min and subsequently eluted onto the SCX column with 80% acetonitrile (ACN; Biosolve, The Netherlands) and 0.05% FA. A total of 50 SCX fractions (1min each, i.e. 40µl elution volume) were collected and used for subsequent LC-MS/MS analysis.

Mass spectrometric analysis. We performed nanoflow LC-MS/MS using an LTQ-Orbitrap Velos mass spectrometer (Thermo Electron, Bremen, Germany) coupled to an Agilent 1200 HPLC system (Agilent Technologies). SCX fractions were dried, reconstituted in 10% FA and delivered to a trap column (ReproSil C18, (Dr Maisch GmbH, Ammerbuch, Germany); 20mm x 100µm inner diameter, packed in-house) at 5µl/min in 100% solvent A (0.1M acetic acid in water). Next, peptides eluted from the trap column onto an analytical column (ReproSil-Pur C18-AQ (Dr Maisch GmbH, Ammerbuch, Germany); 40cm length, 50µm inner diameter, packed in-house) at approximately 100nl/min in a 90min or 3h gradient from 0 to 40% solvent B (0.1M acetic acid in 8:2 (v/v) ACN/water). The eluent was sprayed via distal coated emitter tips butt-connected to the analytical column. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS. Full-scan MS spectra (from m/z 350 to 1500) were acquired in the Orbitrap with a resolution of 30,000 FHMW at 400 m/z after accumulation to target value of 500,000 in the linear ion trap (maximum injection time was 250ms). After the survey scans, the ten most intense precursor ions at a threshold above 5000 were selected for MS/MS with an isolation width of 1.2 Da after accumulation to a target value of 30,000 (maximum injection time was 50ms). Peptide fragmentation was carried out by using higher-energy collisional dissociation (HCD) with an activation time of 0.1ms and a normalized collision energy of 45%. Fragment ions analysis was performed in the Orbitrap with a resolution of 7,500 FHMW and a low mass cut-off setting of 100 m/z.

Data processing. MS raw data were processed with Proteome Discoverer (version 1.3, Thermo Electron). Peptide identification was performed with Mascot 2.3 (Matrix Science) against a concatenated forward-decoy UniPROT database supplemented with all the frequently observed contaminants in MS (version 5.62). The following parameters were used: 50p.p.m. precursor mass tolerance, 0.02Da fragment ion tolerance, up to 2 missed cleavages, carbamidomethyl cysteine as fixed modification, oxidized methionine and TMT modification on N-Term and Lysine as variable modifications. Finally, we performed a deconvolution of the high resolution MS2 spectra, by which all the fragment ions isotopic distributions were converted to an m/z value corresponding to the monoisotopic single charge. Reporter ion based quantification method was chosen in Proteome Discoverer, with the following requirements for reporter ion integration in the MS2 spectra; mass accuracy of maximum 20ppm, peptide ratio maximum limit 100. In order to minimize ratio distortion due to the presence of more than one peptide species within the precursor ion isolation width, we also reject the quantification of MS/MS spectra having a co-isolation higher than 30%. Finally, results were filtered using the following criteria: (i) mass deviations of ±5p.p.m. (ii) Mascot Ion Score of at least 25, (iii) a minimum of 7 amino-acid residues per peptide and (iv) position rank 1 in Mascot search. As a result, we obtained peptide FDRs of 0.3% for the mix 1 and 0.5% for the mix 2, which corresponded to a protein FDR of 1% for the overlapping protein identification of the two 6-plex analyses**.** Finally, peptide ratios were Log2 transformed and normalized by median subtraction (refer to **Global Proteome analysis pipeline** below for details).

Global Proteome analysis pipeline:

- Sample mix composition:
  - ISTD: mixture in 1:1 ratio of 2°MEF, D2H, D5H, D8H, D11H, D16H, D18H, 2°iPSC, ESC, 1°iPSC
  - Mix1: 2°MEF, D2H, D5H, D8H, D11H, and ISTD
  - Mix2: D16H, D18H, 2°iPSC, ESC, 1°iPSC, and ISTD
  - Mix3: D16L, D21L, D21ø, and ISTD
- Raw data processing (e.g. noise filtering, deisotoping, deconvolution) by using Proteome Discoverer 1.3 Software (Thermo):
  - Mix1: 731,645 MS2 events
  - Mix2: 725,642 MS2 events
  - Mix3: 908,982 MS2 events
- Mapping MS2 spectra to peptide sequences by using Mascot 2.3 search engine (Matrix Science), and the following parameters were used for database search:
  - Mass tolerance of 50ppm and 0.02Da for precursor
  - Up to two missed cleavages
  - Cysteine carbamidomethylation as fixed modification
  - Methiodine oxidation, TMT modification on Lysine and peptide N-Termini as variable modificaitions
  - Concatenated forward-decoy database supplemented with all the frequently observed contaminants in MS (Uniprot v_2011_07_Mus musculus) was used.
- Filtered identification at a false discovery rate (FDR) lower than 1%
- Peptide spectrum matches (PSMs) are filtered based on the following criteria in order to obtain an FDR < 1%:
  - Peptide length > 6 aminoacids
  - Peptide rank =1
  - Ion score > 25
  - Delta mass < 5ppm
  - The filtered identifications are summarized as follows:
    - Mix1: 199,373 PSMs; 39,518 peptides; 5,943 proteins
    - Mix2: 220,729 PSMs; 46,206 peptides; 6,408 proteins
    - Mix3: 153,869 PSMs; 40,838 peptides; 6,136 proteins
- Reporter ion based quantification is performed by using Proteome Discoverer 1.3 Software (Thermo):
  - Relative quantification is performed dividing the MS2 intensities of the reporter ions of a given sample by the internal standard mixture (sample x/ISTD). Protein ratios are then calculated as the median of the peptides ratios within the same protein, and peptide quantification is accepted only if:
    - The reporter ions mass deviation is <20ppm
    - The peptide is labeled both at the N-terminal and at the Lysine residues (when present)
    - The precusrsor ion shows a co-isolation interference <30%
    - Refer to **Supplementary Table 6** for number of quantified proteins that passed the filters

Cell Surface Proteomics

Sample preparation for MS analysis. A simplified version of the cell surface capture (CSC) protocol introduced by Wollscheid *et al.*[58] was applied to identify N-glycosylated surface proteins over the project time course. Fixed quantities of protein (5mg), as determined by a duplicate DC protein assay (Bio-Rad), were used in place of cell counts to determine the volumes of cell lysate to process further.

Mass spectrometric analysis. Vacuum centrifugation was performed on a volume of glycopeptide mixture calculated to be derived from 2mg of total protein. After the volume was concentrated to several mL, it was then adjusted to 11mL with 0.1% formic acid and transferred to a well of a 96-well plate, which was subsequently placed in an EASY-nLC nano LC pump (Proxeon) connected to a microcolumn. Microcolumns were created from sections of capillary-scale nanoflow 75μm I.D. fused silica tubing (Polymicro Technologies) pulled to a fine tip using a P-2000 laser puller (Sutter Instruments). They each were packed to a length of 10cm with 5μm Luna C18 resin (Phenomenix) using a pressure vessel, then flushed for 15min with methanol. Microcolumns were regenerated with buffer 'A' (5% acetonitrile and 0.1% formic acid in HPLC-grade water from Fisher) before loading of sample by the nano LC pump. Each chromatography session began with a linear gradient elution of 5% to 25% buffer 'B' (95% acetonitrile and 0.1% formic acid in HPLC-grade water from Fisher) over 45min followed by a linear gradient of 25% to 80% buffer 'B' over 9min. A flow rate of 300nL/min was maintained. Peptides were analyzed using nanospray ionization on an Orbitrap-Velos mass spectrometer (Thermo). MS1 and MS2 spectra were acquired with the instrument operating in the data-dependent mode of one MS scan (on the Orbitrap) followed by up to ten MS2 scans (on the LTQ-Velos) when triggered by ion signals above a threshold of 500. Fragmentation was accomplished using collision-induced association. Three LC-MS replicates were performed for each of the selected time points.

Database searching and analysis. All MS2 spectra were searched using the SEQUEST algorithm and the International Protein Index (IPI) mouse database (Version 3.84) with the reversed protein sequences appended as decoys. Confidences in identifications of peptides (of at least seven amino acids in length) were evaluated using the Statquest probabilistic model[59] and further filtered to within a mass tolerance of 20p.p.m using the accurate ion masses generated by the Orbitrap, thereby attaining an estimated false positive rate of 2%. Any identified peptides were then excluded if they did not include the N-glycosylation consensus sequon NxS/T or did not demonstrate the asparagine to aspartic acid deamidation of 0.986Da resulting from the treatment with PNGaseF. Relative quantities of cell surface proteins were assessed by spectral counting or through use of matching global proteomic quantitative data where possible (refer to **Cell Surface Proteome analysis pipeline** below for details).

Cell Surface Proteome analysis pipeline:

- Samples and controls:
    - Samples: 2°MEF, D2H, D5H, D8H, D11H, D16H, D18H, 2°iPSC
    - Controls: ESC, 1°iPSC
    - Three replicates per sample
- Raw data processing (charge state assignment) using 'extractms' v.2 (rev.11):
    - Replicate set 1: 219,736 MS2 events
    - Replicate set 2: 219,467 MS2 events
    - Replicate set 3: 238,273 MS2 events
- Mapping MS2 spectra to peptide sequences by using Sequest v.27 (rev.9) and in-house supporting programs seach engine (Matrix Science), and the following tolerances and parameters were used for database search:
    - Peptide mass tolerance of 3.0DA
    - Up to one missed cleavage
    - Cysteine carbamidomethylation as fixed modification
    - Asparagine deamidation as a variable modification
    - International Protein Index (IPI) mouse database (Version 3.84), with appended reversed (decoy) database was used.
- Filtered identification at a false discovery rate (FDR) lower than 2%.
- Peptide identifications were filtered as follows to obtain an FDR score < 2%:
    - Initial confidence estimation using STATQUEST methodology
    - Precursor delta mass < 20 ppm
    - Peptide sequences contain N-glycosylation 'sequon' (NxS or NxT, x being any amino acid save proline)
    - The filtered identifications are summarized as follows:
        - 14917 spectral counts
        - 896 identified glycopeptides
        - 432 identified glycoproteins
- 432 cell surface proteins passed the filters.
- 185 overlapped with the global proteomics quantitative data set.

Supplementary Tables

Supplementary Table 1 Gene list referring to Figure 2c-e Reprogramming stage specific gene expression

Supplementary Table 2 (refers to Figure 3c,d,f,g and Extended Data Figure 8e)

Supplementary Table 3 (refers to Figure 4a,b)

Supplementary Table 4 (refers Figure 5a-e; Extended Data Fig. 10c)

Supplementary Table 5 (refers to Figure 6 and Extended Data Figure 10e,f)

Supplementary Table 6 (refers to read/tag counts and other information pertaining to data analysis described above)

Supplementary References

43. Behringer, R. R., Gertsenstein, M., Nagy-Vintersten, K. & Nagy, A. *Manipulating the Mouse Embryo: A Laboratory Manual*. (2013). at <http://books.google.ca/books/about/Manipulating_the_Mouse_Embryo.html?id=4juoa5xMs8oC&redir_esc=y>
44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
45. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35,** W345–9 (2007).
46. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22,** 2008–2017 (2012).
47. Xie, W. *et al.* Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* **153,** 1134–1148 (2013).
48. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25 (2010).
49. O'Geen, H., Echipare, L. & Farnham, P. J. in *Epigenetics protocols* **791,** 265–286 (Humana Press, 2011).
50. Gaspar-Maia, A. *et al.* Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460,** 863–868 (2009).
51. Wang, T. *et al.* The Histone Demethylases Jhdm1a/1b Enhance Somatic Cell Reprogramming in a Vitamin-C-Dependent Manner. *Cell Stem Cell* **9,** 575–587 (2011).
52. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **12,** R22 (2011).
53. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7,** 1728–1740 (2012).
54. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
55. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC*

Genomics **15,** 284 (2014).

56. Krueger, F., Krueger, F., Andrews, S. R. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–1572 (2011).
57. Gauci, S. *et al.* Lys-N and Trypsin Cover Complementary Parts of the Phosphoproteome in a Refined SCX-Based Approach. *Anal. Chem.* **81,** 4493–4501 (2009).
58. Wollscheid, B. *et al.* Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* **27,** 378–386 (2009).
59. Kislinger, T. *et al.* PRISM, a Generic Large Scale Proteomic Investigation Strategy for Mammals. *Molecular & Cellular Proteomics* **2,** 96–106 (2003).