## Supplementary Methods

### *Culturing of hESC*

HES-2 cells were maintained on mouse embryonic fibroblast feeder cells in serum containing medium and passaged by mechanical dissection as described previously (Laslett et al. 2007). Before harvest, HES-2 were grown for 3-4 passages on mouse embryonic fibroblast (MEF) feeder cells with media containing 20% KnockOut serum (KOSR)/DMEM F12, 1% L-Glutamine, 1% non-essential amino acids and 90 µM 2B-ßMercaptoethanol (Invitrogen)(Amit et al. 2000).

### *Membrane-polysome fractionation*

Membrane-associated polysomes were fractionated from cytosolic and other RNA by sucrose density gradient centrifugation as described previously (Kolle et al. 2009). After treatment of hESC with 10 µg/ml of Cyclohexamide (Sigma-Aldrich, St. Louis, http://www.sigmaaldrich.com) for 10 mins, approximately 5 x $10^7$ cells were collected and frozen in liquid nitrogen. Cells were resuspended in 500uL of cold isotonic lysis buffer (10 mmol/L KCl, 1.5 mmol/L $MgCl_2$, and 10 mmol/L Tris-Cl (pH 7.4)) and lysed by multiple passage through a 21 gauge needle and a sample of RNA (50uL) was retained to measure total cellular mRNA. Nuclei and cellular debris were removed by centrifugation at 2,000 x *g* (4°C) for 2.5 minutes. The supernatant was applied to a discontinuous-step sucrose gradient (2.5 mol/L, 2.1 mol/L, 1.95 mol/L, and 1.3 mol/L sucrose) as described previously (Diehn et al. 2000) in a total volume of 2mL. RNA was separated by ultracentrifugation at 26,000 x g for 16 hours. 60µl fractions were obtained from the top of the gradient, and the concentration of RNA was determined by measurement of $A_{260\ nm}$. RNA from fractions 1-16 (cytosol) and 28-37 (membrane) were pooled and purified through RNeasy columns as per manufacturers instructions (Qiagen).

### *RNA isolation, mRNA extraction and ribosome depletion*

RNA was isolated from HES-2 hESC using the RNeasy mini kit (Qiagen). PolyA enriched mRNA was isolated from the 100ug of total RNA using OligoTex (Qiagen). Ribosomal RNA was depleted by using ribosome specific biotinylated probes. For M/S RNA and C/N RNA, two rounds of ribosome depletion using 5ug of starting material were performed.

*Gene expression analysis*

The normalised expression level of known genes, predicted genes, or gene regions was calculated by determining the number of reads mapping within a given set of coordinates and normalising the value per kilobase and per one million mappable tags (RPKM).

*Comparison to existing sequencing data*

The Pearson correlation of the log(2) expression values were calculated in R using the "cor" function, and the heirachical cluster diagram was produced using the "heatmap" function.

*Quantification and comparison to chip-seq histone modification data*

Raw Chip-seq data was sourced from Ku et al (Ku et al. 2008). To determine expression values, we calculated absolte tag counts for H3K4me3 (-1kb - +4kb), H3K27me3 (0-+4kb) and H3K36me3 (across entire gene length) and normalised these as tags per 100bp.

*Measurement of alternative splicing*

A given junction was considered as active if there were 10 or more tags spanning that junction sequence. A unique exon was considered active if the level of expression was greater than 1 RPKM, which is based on previous estimates that this corresponds to approximately 1 transcript per cell and can be robustly detected by mRNA sequencing (Mortazavi et al. 2008). Alternative splicing events were considered as the expression of common donor or acceptor sequence categorised into the following types: 1.

Alternative 5’ Splice Site (A5SS); 2. Alternative 3’ Splice Site (A3SS); 3. Alternative First Exon (AFE); 4. Alternative Last Exon (ALE); 5. mutually exclusive exon usage (MXE); 6. Exon Skipping (SE) or 7. complex splicing combining two or more of the above events (other). Alternatively, ALE events were also determined by the expression of two or more unique 3’ UTR containing exons.

*Support for alternative splicing from paired-end data*

We employed the information from paired-end reads to infer exon junctions. To this end, we first identified for each tag position the closest up- and downstream exon junction using the exon annotations from Ensembl v55. Tag pairs whose forward and reverse tags showed genomic proximity to the same up- and downstream exon junctions were classified as being derived from the same exon. These tag pairs were used to examine the average insert size of the library. ext, tag pairs that did not map to the same exon were used to infer exon junctions. Given the small insert size of the libraries (70-79bp), we ignored any additional exons that may be annotated to a genomic interval between two exons. This strategy thus inferred the exon combination resulting in the smallest possible transcript. This junction information was compared to the data from individual reads spanning an exon-exon junction to provide evidence for alternative splicing.

*Identifying domain changes due to alternative splicing*

Where we found evidence for an alternative transcript isoform, we determined whether the splicing event affects the protein domain architecture of the resulting transcripts. InterPro protein domain annotations were taken from Ensembl v55 and their protein coordinates were mapped to genomic coordinates of the human reference genome (hg19). For each alternative splice event, we determined whether the spliced region is annotated as encoding a protein domain. For the purpose of reporting and given that some splice events are shared by multiple transcript isoforms, one isoform was randomly chosen to represent alternative transcript isoforms that have or do not have the splice junction.

*Clustering of novel sequences into gene regions*

Tags that did not match to known gene models were clustered into regions, by using a sliding window approach across the genome. A significant window had the following properties – at least 40 total tags, and 20 unique start positions per 500bp genome window. Adjacent windows were then combined into putative novel regions. These novel regions were compared to known genes to determine the distribution of 5', 3', intron and intergenic putative novel regions.

*Identification of long 5' and 3'UTR sequences*

Any tags that overlap a coding region was first removed from the analysis. Next, tags mapping directly upstream of the 5'UTR boundary or downstream of the 3'UTR boundary were analysed in 30bp windows. Extension of the 5' UTR or 3' UTR was confirmed if the expression level in any 10 adjacent 30bp window was not less than 50% of the expression level of the canonical transcript (based on Ensembl V55). The most 5' end or 3' end of the gene was considered as the most proximal 30bp window were the signal was still >50% of the expression level of the gene. Only novel 5' or 3'UTRs that extended the gene more than 500bp were considered as truly extended UTR sequences in our analysis.

*Intron retention events*

To determine novel intron retention events, we first stripped all tags that map within known exonic regions (Ensembl v55) and we determined the tags that map within the bounds of introns for the mRNA sample. We then added the following filters: 1. The raw expression level of the intron must be at least 1 RPKM. 2. The expression level of the intron must be at least 50% of the expression level of the Ensembl canonical gene (i.e. RPKM_intron/RPKM_gene > 0.3). This was done to remove any low level intronic noise, which is has a median of ~2%. 3. The ratio of raw tags / unique tags must be less than 5 and the Length/unique tags must be less than 10. These filter steps were used to remove highly expressed small elements that occur within the introns of genes and to ensure continuous signal across the entire length of the intron. To determine exon-extension events, we used a similar algorithm to extended 3'UTRs (above).

*Comparison of M/S RNA:C/N RNA ratios between Illumina microarray and SOLiD sequencing*

We directly compared SOLiD sequencing *M/S RNA:C/N RNA* ratios with our previous Illumina microarray data ((Kolle et al. 2009), GEO accession number GSE14037). Illumina probes were first mapped to Ensembl canonical transcripts using Vmatch. Only probes that uniquely matched a single refseq transcript were then chosen. Only Illumina probes with a precision score > 0.99 and refseq genes with an expression of 2 RPKM or greater in the *M/S RNA* or *C/N RNA* fraction were considered. Pearson correlation coefficient was determined in R using the "Cor" function in R.

*Comparison with ChIP-seq* POUF51 *and* NANOG *data*

POUF51 and NANOG ChIP-seq data was sourced from a study by Kunarso and colleagues (GEO: GSE20650) (Kunarso et al. 2010). The authors used MACS for peak detection and defined binding peaks as those for which $p \leq 1.00e\text{-}05$. Chromosomal coordinates for POUF51 or NANOG binding regions were converted to HG19 using Galaxy (http://galaxy.psu.edu/). To determine which genes expressing alternate 5' exons were POUF51 or NANOG targets, transcription factor binding regions were aligned to regions 8kb upstream from 5' end of alternate 5' exons or directly to alternate 5' exons themselves.

To determine which genes from the Plurinet were POUF51 or NANOG targets, transcription factor binding regions were first annotated with Ensembl gene IDs using

Galaxy and BioMart (http://www.biomart.org/). Simple text matching between these and the Plurinet on Ensembl gene IDs was used to determine which target genes were included in the Plurinet.

Determination of which Plurinet genes had active alternate 5' exons was done using simple text matching on Ensembl gene ID. POUF51 or NANOG targeting of the alternate 5' exons was annotated manually by data visualization in IGV. Alternate 5' exons were annotated as POUF51 or NANOG targets if the transcription factor binding region was within -8 to +2 kb from the alternate 5' exon.

*Re-annotation of 'inactive' genes*

A list of 2990 genes defined by a previous study as inactive in hESCs was sourced from Guenther et al, 2007. The list, initially consisting of gene symbols only, was annotated with Ensembl gene IDs by text matching gene symbols with HUGO Gene Nomenclature Committee (HGNC) 'approved' (2629 matched), 'previous' (151 matched), and 'alias' (70 matched) gene symbols and assigning corresponding Ensembl gene IDs (Galaxy; BioMart). 95% of gene symbols from the original list were successfully annotated with Ensembl gene IDs. Plurinet genes were identified by simple text matching on Ensembl gene ID, and filtered according to RPKM values. Chromatin modification alignments were manually annotated by visualizing alignments in IGV with data from Ku et al as described previously by this study.

Amit M, Carpenter MK, Inokuma MS, Chiu CP, Harris CP, Waknitz MA, Itskovitz-Eldor J, Thomson JA. 2000. Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Dev Biol* **227**(2): 271-278.

Diehn M, Eisen MB, Botstein D, Brown PO. 2000. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nat Genet* **25**(1): 58-62.

Kolle G, Ho M, Zhou Q, Chy HS, Krishnan K, Cloonan N, Bertoncello I, Laslett AL, Grimmond SM. 2009. Identification of human embryonic stem cell surface markers by combined membrane-polysome translation state array analysis and immunotranscriptional profiling. *Stem Cells* **27**(10): 2446-2456.

Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**(10): e1000242.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**(7): 631-634.

Laslett AL, Grimmond S, Gardiner B, Stamp L, Lin A, Hawes SM, Wormald S, Nikolic-Paterson D, Haylock D, Pera MF. 2007. Transcriptional analysis of early lineage commitment in human embryonic stem cells. *BMC Dev Biol* **7**: 12.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.