Ubiquitously expressed genes participate in cell-specific functions via alternative promoter usage

Guihai Feng^{1,2,†}, Man Tong^{2,†}, Baolong Xia^{1,3,†}, Guan-Zheng Luo², Meng Wang², Dongfang Xie², Haifeng Wan¹, Ying Zhang¹, Qi Zhou^{1,*} & Xiu-Jie Wang^{2,**}

Abstract

How do different cell types acquire their specific identities and functions is a fundamental question of biology. Previously significant efforts have been devoted to search for cell-type-specifically expressed genes, especially transcription factors, yet how do ubiquitously expressed genes participate in the formation or maintenance of cell-type-specific features remains largely unknown. Here, we have identified 110 mouse embryonic stem cell (mESC) specifically expressed transcripts with cell-stage-specific alternative transcription start sites (SATS isoforms) from 104 ubiquitously expressed genes, majority of which have active epigenetic modification- or stem cell-related functions. These SATS isoforms are specifically expressed in mESCs, and tend to be transcriptionally regulated by key pluripotency factors through direct promoter binding. Knocking down the SATS isoforms of *Nmnat2* or *Usp7* leads to differentiation-related phenotype in mESCs. These results demonstrate that cell-type-specific transcription factors are capable to produce cell-type-specific transcripts with alternative transcription start sites from ubiquitously expressed genes, which confer ubiquitously expressed genes novel functions involved in the establishment or maintenance of cell-type-specific features.

Keywords embryonic stem cell; pluripotency factors; stage-specific alternative transcription start sites; ubiquitously expressed genes

Subject Category Stem Cells

DOI 10.15252/embr.201541476 | Received 1 October 2015 | Revised 28 June 2016 | Accepted 28 June 2016 | Published online 27 July 2016

EMBO Reports (2016) 17: 1304–1313

Introduction

How is a multiple cellular organism being regulated to generate different cell types has been an intriguing and fundamental question

for developmental biology. One explanation is that the expression of certain transcription factors at defined developmental stages can drive the formation of new cell types. Yet studies have shown that a great proportion of genes are ubiquitously expressed [1]. How are these ubiquitously expressed genes being regulated to participate in cell-type-specific functions remains obscure.

Alternative splicing has been recognized as a major factor for expanding transcript and protein diversities in different cell types [2,3]. It has been shown that about 95% of human multi-exonic genes undergo alternative splicing [4]. In addition to alternative splicing, alternative promoter usage, which produces transcripts with variable first exons, is also a common phenomenon yet has not been well studied so far [5]. More than 50% of human and mouse genes were reported to have at least two promoters [6,7]. Selective usage of alternative promoters will produce transcripts differ in either 5' UTRs or coding sequences, which may not only affect the abundance and translational efficiency of transcripts, but also alter the subcellular localizations or biological functions of their encoded proteins [8]. It has been shown that isoforms with alternative transcription start sites are frequently observed among transcriptomes derived from different developmental stages of eukaryotes [9,10]. Several lines of evidence have indicated that transcripts with stage-specific alternative transcription start sites (SATS) may be involve in the regulation of cell development in spatial and temporal manners. For example, the oocyte-expressed SATS isoform of Dicer encodes an N-terminal trimmed endoribonuclease with higher enzyme activity, which processes small RNA precursors with higher efficiency than the ubiquitously expressed Dicer isoform [11].

Mouse embryonic stem cells (mESCs) are a class of pluripotent stem cells characterized by their ability of unlimited self-renewal and differentiation into all derivatives of the three primary germ layers [12]. Compared to somatic stem cells or terminally differentiated cells, mESCs have specific or high expression of many key pluripotency factors, such as Oct4 (also known as Pou5f1), Sox2,

1 State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

2 Key Laboratory of Genetic Network Biology, Collaborative Innovation Center of Genetics and Development, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

3 University of Chinese Academy of Sciences, Beijing, China

*Corresponding author. Tel: +86 10 6480 7299; Fax: +86 10 6480 7299; E-mail: qzhou@ioz.ac.cn

**Corresponding author. Tel: +86 10 6480 6590; Fax: +86 10 6480 6595; E-mail: xjwang@genetics.ac.cn

†These authors contributed equally to this work

and Nanog, which are crucial for the pluripotent feature of mESCs [13]. Although significant efforts have been devoted to decipher the functional mechanisms of these key pluripotency factors, whether they could regulate ubiquitously expressed genes to confer cell-type-specific functions remains largely unknown yet.

Here, we systematically searched for SATS isoforms in mESCs and identified 110 mESC-specific SATS isoforms produced by 104 genes expressed in both mESCs and differentiated cells. Distinct histone modification profiles and Pol II binding patterns are detected around the transcription start sites of the SATS isoforms. The transcription of ~50% mESC-specific SATS isoforms is regulated by Oct4, Sox2, or Nanog. Functional studies have shown that some SATS isoforms are required for the maintenance of the pluripotent features of mESCs.

Results and Discussion

Identification of SATS isoforms in mESCs

To systematically discover mESC-specific SATS isoforms, we extracted polyadenylated RNAs from three independent mESC lines, mouse tail tip fibroblasts (TTFs), mouse adipose-derived stem cells (ADSCs), and mouse neural stem cells (NSCs) for transcriptome profiling using the paired-end RNA sequencing method. After assembly, 67,394 potential transcripts derived from 20,539 NCBI annotated RefSeq protein-coding genes were obtained. To identify genes with mESC-specific SATS isoforms, we focused on the 10,612 genes expressed in both mESCs and somatic cells (average FPKM > 1). mESC-specific SATS isoforms were identified among genes with at least one transcript (common isoform) being detected in one or more mESC lines as well as majority of examined somatic cell lines, and with one or more transcripts (SATS isoform) specifically expressed in mESCs but not in somatic cells (Figs EV1 and EV2A). The mESC-specific SATS isoforms and their corresponding common isoforms should have different transcription start sites (Figs 1A and EV2B). In total, we identified 110 mESC-specific SATS isoforms encoded by 104 genes (SATS genes). According to our selection criteria, each of these SATS genes expressed at least one common isoform (107 in total) in both the mESCs and all examined somatic cell lines (Dataset EV1).

To validate the mESC-specific 5'-exon and expression of the SATS transcripts, we used reverse transcription PCR (RT-PCR) to detect the presence of SATS isoforms with different expression abundance in mESCs and somatic cells. A total of 18 SATS isoforms with expression ranked among the top 20% of all SATS transcripts, 8 SATS isoforms with expression ranked among the top 20–50% of all SATS transcripts, and 9 SATS isoforms with expression ranked among the 50–100% of all SATS transcripts were tested. Among them, 30 SATS isoforms were specifically expressed in mESCs, but not in somatic cell lines, and the rest 5 SATS isoforms (of genes *Cdyl2*, *Hmgxb4*, *Mitf*, *Mras*, *Tpd52*) had weak expression in one or more somatic cells, in addition to the expression in mESCs (Figs 1A and EV2B and C). All these detected genes also had one ubiquitously expressed common isoforms (Figs 1A and EV2B and C). We further used qRT-PCR to quantify the expression of 7 SATS isoforms from the above tested list and obtained the same expression patterns (Fig EV3). The H3K4me3 histone modification has been proven to be accumulated around the transcription start sites (TSSs) of active

promoters [14]. By analyzing the public H3K4me3 ChIP-seq data of mESCs and MEFs [15], we observed an enrichment of H3K4me3 modification at the TSS regions of the SATS isoforms in mESCs, but not in MEFs (Fig 1B, left panel). Similar pattern was also observed for Pol II binding (Fig 1C, left panel). In addition, DNase I hypersensitive sites, which mark regulatory DNA regions such as promoters and enhancers [16], were also enriched around the TSS regions of the SATS isoforms among three independent mESC lines (Fig 1D, left panel). On the other hand, enrichment of the above three factors were observed around the TSS regions of the common isoforms in both mESCs and somatic cells with comparable densities (Fig 1B–D, right panel). These observations indicate the transcription potential of the SATS isoforms in mESCs, which is in concert with the mESC specific expression pattern of these SATS isoforms.

Transcription of mESC-specific SATS isoforms are regulated by pluripotency factors

To investigate whether the transcription of mESC-specific SATS isoforms were regulated by mESC-specific pluripotency factors, we reanalyzed previously published ChIP-seq data of 18 regulatory factors in mESCs, including the master pluripotency factors Oct4, Sox2, and Nanog, as well as some signaling or chromatin remodeling proteins with crucial functions for mESCs (Fig EV4A). Consistent with the previous reports [17], Oct4, Sox2, Nanog, together with other factors crucial for mESCs, namely Tcf3, Ctnnb1, and Smad1, exhibited high co-bound frequency, indicating their coordinated functions in mESCs (Fig EV4A). In addition, the ChIP-seq data of mESCs showed a significant enrichment of Oct4, Sox2, and Nanog binding around the TSS regions of the SATS isoforms, but not around the TSS regions of the common isoforms (Figs 2A and EV4B). As the positive control, enrichment of Pol II binding was observed around the TSS regions of both the SATS and common isoforms, as both the SATS and common isoforms were expressed in mESCs. On the other hand, Suz12, a repressive factor for gene expression, showed no binding enrichment around the TSS regions of both the SATS and common isoforms of the SATS genes (Fig 2A). Specifically, we found that the promoters of 51 mESC-specific SATS isoforms were enriched of binding sites for one or more key pluripotency factors, namely Oct4, Sox2, and Nanog (with a threshold of P -value = 10^{-5} , Fisher's exact test), as compared to the promoter regions of the common isoforms of the same genes (Fig 2B). In addition, Tcf3 and Ctnnb1, which are both Wnt signaling pathway proteins with essential functions in mESCs, were also enriched around the TSS regions of SATS isoforms (with a threshold of P -value = 10^{-2} , Fisher's exact test) (Fig 2B). Among the 51 mESC-specific SATS isoforms with Oct4, Sox2, or Nanog binding sites, 45% (23 of 51) could be co-regulated by these three factors, yet the promoter regions of the common isoforms of the same SATS genes lacked binding sites for these factors (Fig 2C and D and Dataset EV2).

To test whether the presence of pluripotency factor binding peaks around the promoter regions of the SATS isoforms indeed contributes to transcriptional regulation, we examined the effects of knocking down selected factors on the expression of the SATS isoforms. As expected, silencing Oct4 expression by short hairpin RNAs (shRNAs) in mESCs dramatically reduced the expression of all examined SATS isoforms with Oct4 binding peaks in their

Figure 1. Identification of mESC-specific SATS isoforms.

- A Expression validation of mESC-specific SATS isoforms and their corresponding common isoforms by RT-PCR. The left panel shows the schematic structures of each isoform, with the first exons of the common isoforms and SATS isoforms named 1 and 1a (and 1b when multiple mESC-specific SATS isoforms were identified), respectively. Other exons are numbered according to their orders in the common isoforms. Due to space limitation, some canonical exons presented in both the common and SATS isoforms are omitted and represented by "/". The sites of forward primers (red) and reverse primers (blue) are marked at corresponding exons. mESC1 and mESC2 represent two different mESC lines; ADSC, adipose-derived stem cells; NSC, neural stem cells; MEF, mouse embryonic fibroblasts; TTF, tail tip fibroblasts; NC represents PCR negative control without template. Numbers in parenthesis represent the abundance rank of SATS isoforms.
- B Distribution of H3K4me3 ChIP-seq reads around the transcription start sites (TSSs) of SATS (left) and common (right) isoforms in mESCs and MEFs.
- C Distribution of Pol II ChIP-seq reads around the transcription start sites (TSSs) of SATS (left) and common (right) isoforms in mESCs and MEFs.
- D Distribution of DNase-seq reads around the transcription start sites (TSSs) of the mESC-specific SATS (left) and common (right) isoforms in three mESC lines (Ese14, Zhbctc4, and Escj7S) and three somatic cell lines (Nih3t3, Mel, and Fibroblast).

Source data are available online for this figure.

promoter regions, yet the expression of common isoforms of the same genes were almost not affected (Fig 2E). Similar results were also obtained in the Sox2 or Nanog knockdown experiments (Fig 2F and G). For SATS isoforms being regulated by multiple factors, such as the *Fkbp5* SATS isoforms with binding peaks of Oct4 and Nanog, the *Kdm3a* SATS isoform with binding peaks of Oct4 and Sox2, and the *Pias2* SATS isoform with binding peaks of Oct4, Nanog, and Sox2, all showed reduced expression in corresponding knockdown experiments (Fig 2E–G). It is worth noting that the 2.5-kb upstream region of the *Fkbp5* common isoform also had a weak Nanog binding peak, which may be the reason for the expression reduction of the *Fkbp5* common isoform in the Nanog knockdown experiment. The SATS isoform of *Phf17*, whose promoter does not contain binding sites for Oct4, Sox2, or Nanog, was included as the negative control. As expected, the expression of the *Phf17* SATS isoform did not show expression changes in the Oct4, Sox2 or Nanog knockdown cells, respectively (Fig 2E–G). These results demonstrate that key pluripotency factors, such as Oct4, Sox2, and Nanog, play essential roles in regulating the transcription of SATS isoforms in

mESCs, which also explains the mESC-specific expression of these SATS isoforms. For SATS isoforms without direct binding sites of these examined pluripotency factors, it is possible that they are regulated by other transcription factors with mESC-specific expression, such as transcription factors or non-coding RNAs, some of which may be the downstream targets of Oct4, Sox2, or Nanog.

Characterization and functional analysis of mESC-specific SATS isoforms

Previous studies have shown that transposable elements could rewire gene regulatory networks and therefore participate in the regulation of cell pluripotency in mouse and human ESCs [18–20]. To investigate whether the presence of transposable elements might be involved in the transcription regulation of SATS isoforms, we examined the distribution of transposable elements within the upstream 200 bp of the transcription start site and the first exon of each SATS isoform. Among the 110 SATS isoforms, 58 contained LTR- or SINE-related sequences in the examined regions, yet the

Figure 2. Regulation of mESC-specific SATS isoforms by key pluripotency factors.

A Distribution of Oct4, Nanog, Sox2, Suz12, and Pol II ChIP-seq reads around the transcription start sites (TSSs) of the mESC-specific SATS (left) and common (right) isoforms in mESCs.

B Occupancy frequency of key regulatory factors on the promoter regions of the mESC-specific SATS and common isoforms. Numbers in boxes represent the number of isoforms bound by each factor. Triple stars, double stars, and single star represent P -values $< 1e-5$, $1e-2$, and 0.05 , respectively (Fisher's exact test).

C, D Venn diagrams representing the presence of Oct4, Sox2, or Nanog binding sites on the promoter regions of the mESC-specific SATS (C) and common (D) isoforms.

E-G Expression changes of the mESC-specific SATS isoforms and their corresponding common isoforms in Oct4 (E), Sox2 (F), and Nanog (G) knockdown mESCs detected by qRT-PCR. Error bars represent the standard error of the mean relative expression levels (SEM) of three independent biological replicates. The double stars and single star represent P -values < 0.01 and 0.05 , respectively (two-tailed Student's t -test).

Source data are available online for this figure.

same regions of the common isoforms and other RefSeq genes did not show such enrichment (Fig 3A). These results suggested that the formation of SATS isoforms might be mediated by the transpositions of the LTR and SINE types of transposons.

We next examined the expression abundance difference of the SATS and common isoforms by comparing the read coverage

(FPKM) of the specific 5' exons of each isoform. In general, the common isoforms of the SATS genes tended to have moderate expression, with a few also had low expression (Fig 3B). Unexpectedly, the overall expression abundance of the SATS isoforms was higher than that of their corresponding common isoforms (Fig 3B). It is worth noting that the expression abundance of some SATS

Figure 3. Characterization and functional analysis of mESC-specific SATS isoforms.

- A** Distribution of repeat elements around the transcription start site (TSS) regions of the SATS isoforms (SATS), common isoforms (common), and RefSeq genes (RefSeq). The y-axis represents the proportion of isoforms (or genes) with each type of repeat element around their transcription start site regions. LTR, long terminal repeat; SINE, short interspersed nuclear element; LINE, long interspersed element; S_R, simple repeat; L_C, low complexity region; others, other types of repeats; NA, without repeat sequence.
- B** Expression comparison of the SATS isoforms (SATS), common isoforms (common), and RefSeq genes (RefSeq) in mESCs. The y-axis represents the log₂-transformed average FPKM values of each isoform or gene in three mESC lines (F145, F146, and F147). The FPKM values in each cell line were calculated by the reads mapped to the first exon and normalized by the exon length. Red dots represent the mean expression level of isoforms or gene in each group. The SATS and common isoforms from the same gene are connected by gray lines.
- C** Expression comparison of SATS genes in mESCs and somatic cells. The x-axis and y-axis represent the log₂-transformed average FPKM values of each gene of three mESC lines (F145, F146, and F147) and three somatic cell lines (ADSC, NSC, TTF), respectively. Big red dots, blue dots, and small yellow dots represent SATS genes with $\geq$ twofold expression difference and P -value ≤ 0.05 (two-tailed Student's t -test), SATS genes with $<$ twofold expression changes, and other genes with $\geq$ twofold expression difference and P -value ≤ 0.05 (two-tailed Student's t -test) between mESCs and somatic cells, respectively. R represents the Pearson correlation coefficient of gene expression between mESCs and somatic cells.
- D** Relationship between nucleotide and amino acid changes of the mESC-specific SATS isoforms as compared to corresponding common isoforms. Plus values represent nucleotide/amino acid expansions in the SATS isoforms, minus values represent nucleotide/amino acid deletions. The top and right boxplots are the quantile plots of nucleotide changes or amino acid changes, respectively. Black bars within the quantile box represent the median of changes. The lower and upper error bars represent the first quartile (Q1) minus 1.5 interquartile range (IQR) and the third quartile (Q3) plus 1.5 IQR, respectively. Points outside of these bars are defined as outliers.
- E** Examples of domain changes in proteins encoded by the SATS isoforms. Regions encoded by the common isoforms but not by the SATS isoforms are marked by red-dashed boxes. Numbers represent the orders of amino acids in the protein.
- F** Gene Ontology enrichment analysis of mESC-specific SATS genes. Circles are the enriched biological processes (P -value < 0.01 , two-sided hypergeometric test with Benjamini–Hochberg correction) among SATS genes identified by ClueGO. The names of processes and their related GO terms are shown in the same colors. Circles are connected according to the hierarchical relationships of GO terms. The sizes of circles are negatively correlated with the enrichment P -values of GO terms.

isoforms was much higher than that of their corresponding common isoforms, which further proved the unique transcription regulation of the SATS isoforms. When using twofold expression difference and P -value = 0.05 (two-tailed Student's t -test) as the cutoff, 58 of the 104 SATS genes were identified with increased expression

in mESCs as compared to somatic cell lines (Fig 3C), a proportion significantly higher than that of other expressed genes (P -value = 3.7×10^{-7} , Fisher's exact test), indicating that the production of the SATS isoforms tended to increase gene expression in mESCs.

Figure 4. The SATS isoform of *Nmnat2* participates in mESC pluripotent state maintenance.

A Expression pattern of *Nmnat2* common and SATS isoforms in mESCs and somatic cells detected by qRT-PCR. Gray triangles represent independent biological replicates; red dots and blue intervals represent the mean values and SEM, respectively.

B Regulation of *Nmnat2* SATS isoform by Oct4, Sox2, and Nanog. Left, distribution of Oct4, Sox2, and Nanog ChIP-seq reads around the TSS region of the *Nmnat2* SATS isoform in mESCs; right, reduced expression of the *Nmnat2* SATS isoform in Oct4, Sox2, or Nanog knockdown mESCs. Gray triangles represent independent biological replicates; red dots and blue intervals represent the mean values and SEM, respectively.

C Expression changes of *Oct4*, *Nanog*, and the common isoform of *Nmnat2* in *Nmnat2* SATS isoform knockdown mESCs. Double stars and single star represent *P*-values ≤ 0.01 and 0.05 (two-tailed Student's *t*-test), respectively. Error bars represent the SEM of three independent biological repeats.

D Abundance changes of PE-labeled Oct4 and APC-labeled Nanog proteins in *Nmnat2* SATS isoform knockdown mESCs by flow cytometry analysis. Knocking down *Nmnat2* SATS isoform by two independent siRNAs is included. The siCtrl and siOct4 mESCs are used as negative and positive controls, respectively. MFI, median fluorescence intensity.

E Morphology and alkaline phosphatase (AP) staining of *Nmnat2* SATS isoform knockdown ESCs. BF, bright field. Scale bars are 100 μ m.

Source data are available online for this figure.

Compared to their corresponding common isoforms, 61.8% (68 out of 110) of the mESC-specific SATS isoforms encoded proteins with altered amino acids, and the rest produced transcripts with

variable 5' UTR (Fig 3D and Dataset EV2). Among the mESC-specific SATS isoforms with changes in coding sequences, only 13 resulted in extended ORF lengths, all others encoded shortened proteins as

compared to the common isoforms (Fig 3D). Several key genes for major epigenetic modifications, such as histone modification-related *Ash2l*, *Hmgxb4*, and *Kdm2b*, were identified to have mESC-specific SATS isoforms with altered coding sequences, and some resulted in the loss of functional domains (Fig 3E).

Functional analysis of the SATS encoding genes using gene ontology annotation showed strong preference for histone modification-related functions, including histone modification related to stem cell population maintenance (Fig 3F). Similar functions were also identified to be enriched among the SATS encoding genes using the gene set enrichment analysis (GSEA) (Fig EV5).

SATS isoforms of *Nmnat2* and *Usp7* participated in mESC pluripotency maintenance

As the SATS isoforms in this study are specifically expressed in mESCs, it is very likely that they possess functions related to the specific features of mESCs. To verify this hypothesis, we knocked down the SATS isoforms of *Nmnat2* and *Usp7* genes, respectively, to study their functions in mESC pluripotency maintenance. Both of the selected SATS isoforms encoded proteins with altered amino acids as compared to those encoded by the corresponding common isoforms. *Nmnat2* belongs to the nicotinamide mononucleotide adenylyltransferase (NMNAT) enzyme family, which catalyzes NAD biosynthesis. Previous studies have shown that the complete deficiency of *Nmnat2* common isoform in mouse lead to perinatal lethality due to respiration failure [21]. The SATS and common isoforms of *Nmnat2* in mESCs differ in the first 7 exons of the common isoform (Fig 1A and Dataset EV2). The common isoform of *Nmnat2* (encoding 307 amino acids) was detected in most examined cell types, with higher expression in NSCs (Figs 1A and 4A), which is in agreement with the previous report [22]. On the contrary, the expression of the SATS isoform (encoding 90 amino acids) was only specifically detected in mESCs (Figs 1A and 4A). Potential binding sites of Oct4, Sox2, and Nanog (yet the binding affinity did not pass the statistical test of the MACS software) were detected in the promoter region of the *Nmnat2* SATS isoform. To examine whether the expression of *Nmnat2* SATS isoform is regulated by Oct4, Sox2, and Nanog, we used shRNAs to knock down each of these factors in mESCs, respectively, and observed remarkable reduction of *Nmnat2* SATS isoform expression in each knockdown experiment (Fig 4B).

We next examined the potential function of *Nmnat2* SATS isoform in regulating mESC pluripotency. Knocking down *Nmnat2* SATS isoform by two independent siRNAs both reduced the expression of key pluripotency factors, namely Oct4 and Nanog, at both the mRNA and protein levels (Fig 4C and D). In concert with the reduced pluripotency factor expression, the *Nmnat2* SATS isoform knockdown mESCs exhibited differentiated morphology and failed to be stained by alkaline phosphatase (AP) (Fig 4E), which specifically stains pluripotent cells. These results indicated that the *Nmnat2* SATS isoform is required to maintain the pluripotent feature of mESCs. As the control, the abundance of the common isoform of *Nmnat2* did not change in these experiments (Fig 4C), indicating that the pluripotency regulatory function is specific for the *Nmnat2* SATS isoform.

We also observed the similar effect in maintaining mESC pluripotency for the SATS isoform of *Usp7*, which encodes an ubiquitin

carboxy-terminal hydrolase. Knocking down *Usp7* SATS isoform resulted in reduced Oct4 and Nanog expression, and leads to mESC differentiation (Fig EV6). As the SATS isoforms of both *Nmnat2* and *Usp7* genes are required for the maintenance of mESC pluripotency, we speculate that many other SATS isoforms may have similar functions in mESCs.

One intriguing question in cell biology is how cells are regulated to differentiate into different types. As the majority of genes are expressed in multiple organs, although the expression of cell-type-specific transcription factors could initiate cell fate conversion, how are those ubiquitously expressed genes being regulated to participate in gene networks determining the specific features of each cell type remains largely unknown. Here, we have identified that the specific expression of SATS isoforms, which tends to be regulated by cell-type-specific factors, as one solution. As genes encoding mESC-specific SATS isoforms also have at least one common isoform ubiquitously expressed in both mESCs and differentiated cells, the additional expression of SATS isoforms could confer functional diversity of these genes, thus enabling them to execute functions specifically needed in mESCs. The specific functions of the SATS isoforms of the *Nmnat2* and *Usp7* genes in regulating mESC pluripotency further proved this notion.

It is worth noting that many key epigenetic modification-related genes in mESCs, such as DNA demethylase *Tet2*, histone methyltransferase complex subunit *Ash2l*, and histone demethylases *Kdm2b* and *Kdm3a*, are all among the list of genes with mESC-specific SATS isoforms. This is in concert with the previous observation on the dynamic changes of DNA methylation and histone modifications among different cell types. The production of SATS isoforms of these genes may be responsible for the formation of cell-type-specific epigenetic modifications, and other SATS isoforms of these genes are expected in other cell types.

Compared to the progresses in functional gene identification and epigenetic regulation studies, little attention has been paid on the transcriptome complexity and dynamics in mESCs. The identification of mESC-specific SATS isoforms in this work revealed the wide presence of SATS isoforms in mESCs, even for genes commonly expressed in pluripotent and somatic cells. Results of this work can serve as a resource for further functional studies of the SATS isoforms in mESCs. The discovery of cell-type-specific key pluripotency factors in regulating the transcription of SATS isoforms also revealed a new layer of gene expression regulation, and will shed new light on understanding the formation of cell-type-specific features.

Materials and Methods

Cell culture

Mouse ESCs (F145, F146, F147) with B6D2F1 (C57BL/6 × DBA/2) genetic background were cultured with 2i medium using mitomycin C-treated mouse embryonic fibroblast (MEF) cells as feeder cells as previously described [23]. The 2i medium consists of N2, B27, 1 μ M MEK inhibitor PD0325901 (Stemgent), 3 μ M GSK3 β inhibitor CHIR99021 (Stemgent), and 10³ units mouse recombinant LIF (Millipore). The ADSCs were derived from the adipose tissues of mouse with CF1 genetic background. Briefly, the adipose tissues were

freshly collected and washed with sterile 1× PBS (Invitrogen), and then minced and incubated with 0.1% type I collagenase (Invitrogen) in DMEM/F-12 (1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F-12 medium; Invitrogen) for 1 h at 37°C. After inactivation of collagenase by diluted DMEM/F-12 plus 10% FBS and vigorously agitation, cells were centrifuged at 300 g for 5 min. The pellet was resuspended and cultured with DMEM/F-12 containing 10% FBS. Neural stem cells (NSCs) were isolated from brains of newborn pups (B6D2F1 genetic background) as previously described [24]. NSCs were cultured in N2B27 medium (Invitrogen) supplemented with 20 ng/ml bFGF and 20 ng/ml EGF (PeproTech). MEF cells (CF1 genetic background) and tail tip fibroblast (TTF) cells (B6D2F1 genetic background) were isolated from E13.5 fetuses and adult mouse tail tip tissues, respectively, and cultivated with DMEM plus 10% FBS (Invitrogen). Sertoli cells were isolated from decapsulated testis of postnatal day 5 (P5) mice (B6D2F1 genetic background) as previously described [24]. All animal experiments were performed following the Guidelines for the Care and Use of Laboratory Animals established by the Beijing Association for Laboratory Animal Science.

Public data sources

Mouse genome assembly mm9 was used in this study. The NCBI RefSeq, Ensembl, and UCSC transcript annotations for mm9 were downloaded from the UCSC database using the Table Browser tool [25] and processed by the Cuffmerge script of the Cufflinks software package [26–29], and then used in RNA-seq data mapping and transcript assembly. All public high-throughput sequencing data used in this study were obtained from the NCBI Gene Expression Omnibus (GEO) database [15,17,30–33], and the detailed information is shown in Table EV1.

RNA-seq library preparation

Total RNA was extracted from cultured cells by TRIzol reagent (Invitrogen), after which 1 µg total RNAs was used for each reverse transcription reaction (Invitrogen). For RNA-seq library construction, two rounds of PolyA+ tailed RNA purification were performed for each sample using Oligotex mRNA Mini Kit (Qiagen). Sequencing was performed on an Illumina HiSeq 2000 or HiSeq 2500 sequencer with 100-bp or 125-bp paired-end sequencing reactions.

RNA-seq data processing

The RNA-seq reads of each sample were mapped to the mouse mm9 genome assembly independently by the TopHat2 software (version 2.0.4) [34] using the annotated gene structures as templates. Default parameters of TopHat2 were used except that the maximum mismatch number allowed in read alignment was set to 2, and the segment length was set to one-third of the query length. Reads with unique genome location were reserved for transcriptome assembling using Cufflinks (version 2.0.2) with default parameters. Only transcripts with lengths longer than 200 nt were retained as input for Cuffcompare for splicing isoform identification. To ensure the accuracy of the results, two replicates of one mESC line and the TTFs were prepared and sequenced separately.

Identification of mESC-specific SATS isoforms

To search for mESC-specific SATS isoforms, the transcriptome data of mESCs, TTFs, ADSCs, and NSCs produced in this work were analyzed using the following criteria: (i) The read coverage of the specific region of the SATS isoforms should be higher than 1 FPKM in each mESC line and lower than 0.5 FPKM in each somatic cell line; (ii) at least two transcripts with different first exons should be detected for each gene; and (iii) the raw counts of reads spanning the junction between the isoform-specific and common exons should be higher than 5 as well as 1 PTM (per ten million mapped reads) in at least one mESC line, and be lower than 2 as well as 0.2 PTM in every somatic cell line.

Transposable element identification in the TSS regions of the mESC-specific SATS isoforms

The annotations of transposable elements were downloaded from RepeatMasker database (<http://repeatmasker.org>). The upstream 200 bp and the first exon region of each transcript in RefSeq and mESC-specific SATS genes were used to examine the transposable element distribution (with at least 1 nt overlap). Transposable elements were classified into six types, namely long terminal repeat (LTR), short interspersed nuclear element (SINE), long interspersed element (LINE), simple repeat (S_R), low complexity (L_C), and the others.

Reverse transcription PCR (RT–PCR) and quantitative real-time PCR (qRT–PCR) analysis

Total RNA was extracted from each sample by TRIzol reagent (Invitrogen), and 0.5 µg total RNA was used for each reverse transcription reaction (Invitrogen). RT–PCR and qRT–PCR were performed using GoTaq Green master mix (Promega) and SYBR Green master mix (Toyobo), respectively. Primers were designed to target the mutually exclusive exons of the mESC-specific SATS and common isoforms, and are listed in Dataset EV3. Samples from 3 different mESC lines each with 2 parallel loading were included in all RT–PCR and qRT–PCR experiment as biological and technical replicates, respectively.

RNAi experiments

To generate short hairpin RNA (shRNA)-mediated stable RNAi ESCs, we designed shRNAs for *Oct4*, *Nanog*, and *Sox2* according to a previous report [35], and primers used for shRNA vector construction are listed in Dataset EV3. To generate lentivirus, pLL3.7 (Addgene #11795) vectors containing short hairpin RNAs (shRNAs) were transferred into HEK293T cells with packing vectors psPAX.2 and pMD2.G (Addgene #12260 and #12259), harvested at 72 h post-transfection, filtered through 0.45-µm filter (Millipore), pooled, and centrifuged at 2,475 g for 30 min at 4°C. Virus suspensions were titrated to $\sim 5 \times 10^8$ infection units/ml using the rapid titer kit according to the protocol (Clontech) and then added to the culture medium of mESCs to achieve a multiplicity of infection (MOI, average number of lentiviral particles per cell) of 1. Lentivirus-transfected mESCs were cultured in mESC medium for 72 h, and

then washed and resuspended in FACS buffer for analysis using a FACSCalibur cell sorter (BD Biosciences), GFP (expressed by the pLL3.7 vector)-positive mESCs were collected for further studies.

To screen for functional SATS isoforms in mESCs, 21-nt siRNAs including the negative control siRNAs, without homology to known mouse genes, were designed using online software (Invitrogen) and synthesized (GenePharma). Two or three different siRNAs (20 pmol each) targeting *Nmnat2* or *Usp7* were independently transfected into 1×10^5 mESCs with RNAiMAX reagent following standard procedures (Invitrogen). Independent experiments using three different cell lines were performed for each set of siRNAs. mESCs were harvested for molecular and phenotypic examinations after 48 h.

Alkaline phosphatase staining and flow cytometry analysis

SiCtrl, siNmnat2, and siUSP7 mESCs were fixed and stained using the alkaline phosphatase detection kit according to a standard protocol (Millipore). Cells were trypsinized, harvested, and fixed in 2% paraformaldehyde (PFA) for 10 min at room temperature. After washing by $1 \times$ PBS, cells were incubated in blocking buffer (0.3% Triton X-100, 5% BSA; Sigma) in $1 \times$ PBS for 30 min at room temperature and then incubated with PE-conjugated OCT4 antibody (BD, 560186), APC-conjugated NANOG antibody (BD, 560279), or isotype controls (BD, 559320 and 557732), respectively, for 30 min at 37°C. Nuclear staining was performed with DAPI (Invitrogen) following the standard protocol. The fluorescence levels of OCT4 and NANOG in cells were analyzed by using FACSCalibur cell sorter.

ChIP-seq data analysis

The raw ChIP-seq data were mapped to the mouse mm9 genome assembly using Bowtie1 software (version 0.12.7) [36] with the parameters of seed length set to 12 and the maximum number of mismatches allowed in each read set to 2. Duplicated reads were merged, and reads with multiple genomic loci were discarded. The remaining reads were analyzed by the MACS software to identify peaks (transcription factor binding sites) with a threshold of P -value $< 1e-5$ (version 1.4.2) [37].

ChIP-seq and RNA-seq data visualization

Distribution of high-throughput sequencing data on genes was plotted and visualized by the Integrative Genomics Viewer (IGV) software [38]. Distributions of the H3K4me3, Pol II, and DNase I ChIP-seq read density around the transcription start site (TSSs) regions (the upstream and downstream 2 kb regions, respectively) were plotted using sitepro from the CEAS package with a bin size of 20 bp [39]. The hierarchical clustering and heat map for genomic co-occupancy of the studied regulatory factors were generated using R. Genomic regions co-occupied by two factors were identified by searching for their binding peaks with at least 1 nt overlap on the genome. The asymmetric color distribution on the matrix is caused by the unequal amount of total peak numbers among samples.

Others tools and statistical analysis

Gene ontology analysis of SATS genes was performed by the ClueGO [40] and GSEA software [41]. Point plots were produced by

the smoothScatter function of R. The heat maps were produced by the heatmap.2 function of R. Two-tailed Student's t -test was used to calculate the P -values of the qRT-PCR results.

Accession numbers

All sequencing data files generated by this study have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE56529.

Expanded View for this article is available online.

Acknowledgements

We sincerely thank Zhi-min Wang for the help with primer design and Ting Li at Center for Developmental Biology of Institute of Genetics & Developmental Biology for FACS assay. This work was supported by China 973 Program (2014CB964901 to X.-J.W.), CAS Strategic Priority Research Program grants XDA01020105 (to X.-J.W.) and XDA01020101 (to Q.Z.), and the National Natural Science Foundation grant 91319308 (to Q.Z. and X.-J.W.).

Author contributions

XW and QZ conceived this project, supervised the experiments, analyzed the data, and wrote the manuscript. GF and XW performed bioinformatic analysis and experimental designs. MT, BX, MW, DX, HW, and YZ were involved in the cell culture and molecular experiments. GL helped with bioinformatics analysis. GF, MT, and BX contributed equally to this work.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Ramskold D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 5: e1000598
2. Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11: 345–355
3. Modrek B, Lee C (2002) A genomic view of alternative splicing. *Nat Genet* 30: 13–19
4. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413–1415
5. Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, Dahmane N, Davuluri RV (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res* 21: 1260–1272
6. Baek D, Davis C, Ewing B, Gordon D, Green P (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res* 17: 145–155
7. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C et al (2005) The transcriptional landscape of the mammalian genome. *Science* 309: 1559–1563
8. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* 24: 167–177
9. The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507: 462–470

10. Haberland V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM *et al* (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* 507: 381–385
11. Flemr M, Malik R, Franke V, Nejepinska J, Sedlacek R, Vlahovicek K, Svoboda P (2013) A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* 155: 807–816
12. Smith AG (2001) Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol* 17: 435–462
13. Jaenisch R, Young R (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* 132: 567–582
14. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. *Cell* 128: 707–719
15. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV *et al* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* 488: 116–120
16. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B *et al* (2012) The accessible chromatin landscape of the human genome. *Nature* 489: 75–82
17. Zhang X, Peterson KA, Liu XS, McMahon AP, Ohba S (2013) Gene regulatory networks mediating canonical Wnt signal-directed control of pluripotency and differentiation in embryo stem cells. *Stem Cells* 31: 2667–2679
18. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487: 57–63
19. Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S (2010) Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res* 20: 804–815
20. 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: 631–634
21. Hicks AN, Lorenzetti D, Gilley J, Lu B, Andersson KE, Miligan C, Overbeek PA, Oppenheim R, Bishop CE (2012) Nicotinamide mononucleotide adenylyltransferase 2 (Nmnat2) regulates axon integrity in the mouse embryo. *PLoS One* 7: e47869
22. Mayer PR, Huang N, Dewey CM, Dries DR, Zhang H, Yu G (2010) Expression, localization, and biochemical characterization of nicotinamide mononucleotide adenylyltransferase 2. *J Biol Chem* 285: 40387–40396
23. Tong M, Lv Z, Liu L, Zhu H, Zheng QY, Zhao XY, Li W, Wu YB, Zhang HJ, Wu HJ *et al* (2011) Mice generated from tetraploid complementation competent iPS cells show similar developmental features as those from ES cells but are prone to tumorigenesis. *Cell Res* 21: 1634–1637
24. Sheng C, Zheng Q, Wu J, Xu Z, Wang L, Li W, Zhang H, Zhao XY, Liu L, Wang Z *et al* (2012) Direct reprogramming of Sertoli cells into multipotent neural stem cells by defined factors. *Cell Res* 22: 208–218
25. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32: D493–D496
26. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 12: R22
27. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28: 511–515
28. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31: 46–53
29. Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27: 2325–2329
30. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J *et al* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133: 1106–1117
31. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J *et al* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134: 521–533
32. Chen Q, Chen Y, Bian C, Fujiki R, Yu X (2013) TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature* 493: 561–564
33. Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* 473: 389–393
34. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14: R36
35. Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L *et al* (2006) Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat Cell Biol* 8: 1114–1123
36. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25
37. Zhang Y, Liu T, Meyer CA, Eickhout J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9: R137
38. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. *Nat Biotechnol* 29: 24–26
39. Shin H, Liu T, Manrai AK, Liu XS (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics* 25: 2605–2606
40. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25: 1091–1093
41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550