Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency

Chan-Shuo Wu, Chun-Ying Yu, Ching-Yu Chuang, Michael Hsiao, Cheng-Fu Kao, Hung-Chih Kuo, and Trees-Juen Chuang

Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan; Institute of Cellular and Organismic Biology, Academia Sinica, Taipei 11529, Taiwan

Trans-splicing is a post-transcriptional event that joins exons from separate pre-mRNAs. Detection of trans-splicing is usually severely hampered by experimental artifacts and genetic rearrangements. Here, we develop a new computational pipeline, TSscan, which integrates different types of high-throughput long-/short-read transcriptome sequencing of different human embryonic stem cell (hESC) lines to effectively minimize false positives while detecting trans-splicing. Combining TSscan screening with multiple experimental validation steps revealed that most chimeric RNA products were platform-dependent experimental artifacts of RNA sequencing. We successfully identified and confirmed four trans-spliced RNAs, including the first reported trans-spliced large intergenic noncoding RNA ("tsRMST"). We showed that these trans-spliced RNAs were all highly expressed in human pluripotent stem cells and differentially expressed during hESC differentiation. Our results further indicated that tsRMST can contribute to pluripotency maintenance of hESCs by suppressing lineage-specific gene expression through the recruitment of NANOG and the PRC2 complex factor, SUZ12. Taken together, our findings provide important insights into the role of trans-splicing in pluripotency maintenance of hESCs and help to facilitate future studies into trans-splicing, opening up this important but understudied class of post-transcriptional events for comprehensive characterization.

[Supplemental material is available for this article.]

Alternative splicing, which arises from post-transcriptional events, can lead to the generation of multiple transcript isoforms from a single gene, thus providing an essential source of diversity for the transcriptome and proteome (Graveley 2001; Maniatis and Tasic 2002; Black and Grabowski 2003; Bracco and Kearsey 2003; Blencowe 2006; Chen et al. 2006; Ben-Dov et al. 2008; Huang et al. 2008; Jin et al. 2008; Mudge et al. 2011). Splicing can occur either in cis or in trans (Horiiuchi and Aigaki 2006; Gingeras 2009). Cis-splicing joins exons within a single precursor mRNA (pre-mRNA), whereas trans-splicing joins exons from two or more separate pre-mRNAs originating from the same gene (intragenic trans-splicing) or two or more different genes (intergenic trans-splicing). The best-characterized form of trans-splicing is spliced- leader (SL) trans-splicing, which provides mRNAs with a new 5’ cap and leader sequence, and commonly occurs in unicellular organisms, nematodes, and trypanosomes (Sutton and Boothroyd 1986; Krause and Hirsh 1987; Nilsson 2001; Hastings 2005). However, the mechanisms underlying non-SL trans-splicing remain largely unknown (Lasda and Blumenthal 2011). To date, only a few non-SL trans-splicing events have been well-documented. In higher eukaryotes, the best-known trans-splicing examples are two Drosophila genes, mod(mdg4) and lola, which are involved in apoptosis and axon guidance decisions, respectively (Dorn and Krauss 2003; Goeke et al. 2003). The most prominent examples of human genes that undergo trans-splicing reported so far are JAZF1-SUZ12 (also known as JJAZ1) and SLC45A3-ELK4; the former is translated into a chimeric protein with anti-apoptotic function and is believed to be a prerequisite for chromosomal exchange (Li et al. 2008; Gingeras 2009; Schoenfelder et al. 2010), and the latter is related to prostate cancer (Gingeras 2009; Rickman et al. 2009). Although trans-splicing remains an understudied class of post-transcriptional events in higher eukaryotes, accumulating evidence suggests it is of biological significance (Gingeras 2009).

Generally, trans-splicing is detected by comparing the reference genomes with ESTs/mRNAs (Shao et al. 2006; Li et al. 2009; Herai and Yamagishi 2010; Kim et al. 2010) or by next-generation sequencing (NGS) of mRNAs (RNA-seq) (McManus et al. 2010; Zhang et al. 2010; Al-Balool et al. 2011; Fang et al. 2012). Transsplicing events detected by such means may, however, include a considerable number of false positives that arise from experimental artifacts, such as template switching (McManus et al. 2010; Ozsolak and Milos 2011). Template switching is generated during RT-PCR and frequently emerges in cDNA products (Cocquet et al. 2006; Houseley and Tollervey 2010). A prominent study using hybrid mRNAs (i.e., Drosophila melanogaster females vs. Drosophila sechellia males) demonstrated that experimental artifacts are the predominant source of apparent trans-spliced RNA products observed in mRNA (McManus et al. 2010). It would, however, be impossible to apply such a system to humans. Furthermore, genetic rearrangements can form noncolinear (or chimeric) RNAs (Shao et al. 2006; Gingeras 2009; Frenkel-Morgenstern et al. 2012),

2These authors contributed equally to this work.
4Corresponding authors
E-mail kuohuch@gate.sinica.edu.tw
E-mail trees@gate.sinica.edu.tw

© 2014 Wu et al. This article, published in Genome Research, is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.
which are not easily distinguished from trans-spliced RNAs, presenting another challenge to the accurate detection of trans-splicing. To our knowledge, there is currently no available method of systematically analyzing trans-splicing that can simultaneously account for experimental artifacts and genetic rearrangements in humans. To address these issues, we developed TSscan, which utilizes transcriptome sequencing data from different NGS platforms and different undifferentiated human embryonic stem cell (hESC) lines. It is important to note that hESCs have been reported to have a very high level of transcriptome complexity, and high transcript diversity has been suggested to contribute to hESC pluripotency (Wu et al. 2010). As such, it is worth investigating whether trans-splicing exists and is biologically significant in hESCs. By performing TSscan screening combined with multiple experimental validation steps, we successfully confirmed that four trans-splicing events occur in hESCs. We found that these trans-spliced RNAs were all highly expressed in human pluripotent stem cells and differentially expressed during hESC differentiation. We also identified the first trans-spliced large intergenic noncoding RNA, showed that it tends to be specifically transcribed in human pluripotent stem cells, and significantly affects pluripotency maintenance of hESCs. Therefore, this study not only describes a new approach to systematically detect trans-splicing, but also provides further insight into the potential roles of trans-splicing in hESC pluripotency and early human embryonic development.

Results

Identification of trans-spliced RNAs in hESCs using TSscan

To search for trans-splicing events in hESCs, we generated 0.83 million long reads (averaging 353.7 bp) and 230.63 million short reads (50 bp) from H9 hESCs by performing Roche 454 and SOLID whole-transcriptome sequencing, respectively, and directly downloaded the 454/Illumina RNA-seq data of H1 hESCs from a publicly available database (Table 1; Wu et al. 2010). Since trans-splicing may also join colinear exons such as modmdk4 and lola (Dorn and Krauss 2003; Goeke et al. 2003), in this study we only consider trans-spliced noncolinear (or chimeric) RNA candidates for simplicity. TSscan involved four main screening steps (Fig. 1A). First, we searched for all possible chimeric RNA candidates by aligning the long 454 reads of hESC H1/H9 against the human reference genome (see Methods) and extracted 8822 candidates (Fig. 1B). The junction sites between two topologically distinct genomic loci were designated as “chimeric junction sites.” Of note, since both 454 RNA-seq libraries were prepared by oligo-DT selection, the retrieved chimeric RNA candidates are unlikely to be noncolinearly encoded RNAs that were cis-spliced to form circular RNAs (i.e., RNAs in which the exon order is a circular permutation of that encoded by the corresponding genomic sequence [Hsu and Coca-Prados 1979; Nigro et al. 1991]). Second, to minimize the possibility of false positives resulting from lack of depth in long reads or NGS platform specificity, we aligned the short reads (i.e., SOLID/Illumina reads) against the 8822 long-read-nominated candidates and discarded the candidates that were not supported by short reads. The remaining candidates were then categorized into four subsets according to the types of supporting RNA-seq data (i.e., S1–S4; see Fig. 1B). Third, to further eliminate potential experimental artifacts, candidates that satisfied any of the following in silico filters were eliminated: (1) chimeric junction sites containing short homologous sequences (SHSS) or gaps (as these tend to arise from template switching) (McManus et al. 2010); (2) sense–antisense fusion candidates containing noncanonical splicing signals at the chimeric junction sites (Houseley and Tollervey 2010); and (3) candidates containing sequences from the mitochondrial genome (McManus et al. 2010). Finally, to eliminate potential genetic rearrangements, only the nine candidates that were supported by RNA-seq data from both H1 and H9 hESCs were retained (Fig. 1B; Table 2). Upon completion of the TSscan screening process, ~99.9% of the 454-nominated candidates had been discarded. We reason that the presence of experimental artifacts is the most likely explanation for the TSscan-excluded cases (see Discussion).

Experimental validation of trans-splicing events identified by TSscan

To confirm that the nine candidates identified by TSscan were indeed examples of trans-splicing, we designed a series of experimental validations (Fig. 1A). We first performed RT-PCR with Moloney Murine Leukemia Virus (MMLV)-derived reverse transcriptase (RTase) (the NGS cDNA library was generated using the same RTase), and found that the transcripts of five trans-splicing candidates were readily detected in multiple hESC lines (H1, H9, and NTU1) (Fig. 1C). These transcripts were not detected in the RT-free control, establishing that they did not arise from genomic contamination (Fig. 1C). As the chimeric junction sites of these five cases are all intragenic (originating from CSNK1G3, ARHGAP5, FAT1, RMST, and SOBP), they are designated as tsCSNK1G3, tsaRHGAP5, tsaFAT1, tsRMST, and tsSOBP, respectively. False positive splicing events that arise from template switching tend to be RTase-dependent and can therefore be detected by comparing the PCR products that arise from the products of different RTases (Houseley and Tollervey 2010). As such, we further validated the five cases by RT-PCR using Avian Myeloblastosis Virus (AMV)-derived RTase. This revealed that tsSOBP is MMLV-RTase-dependent, and thus an artifact (Fig. 1D). Increasing the primer annealing temperature of MMLV-based RTase experiments has been previously shown to suppress the occurrence of template switching (Ouhammouch and Brody 1992; Cocquet et al. 2006); however, we found that increasing the temperature did not eliminate the tsSOBP artifact (Supplemental Fig. 1). We further examined these five cases by performing RNase protection assay (RPA; Supplemental Material), a non-RTase-dependent validation of RNA (Ghislain et al. 2012), on total RNA of hESC H9, and found that only the probes for the RTase dependent-case tsSOBP were degraded (Supplemental Fig. 2). These results indicated that tsSOBP was indeed an experimental artifact. We thus emphasize the necessity of comparing the products of different RTases in confirming trans-splicing. Finally, we used qRT-PCR to show that trans-splicing was not a rare event in hESCs (Fig. 1E), and sequenced the RT-PCR amplicons to validate the identity of the chimeric junction sites.

| Table 1. NGS data sets used in this study |
|------------------------------------------|
| **NGS platform** | **hESC lines** | **Number of reads** | **Length (bp)** |
|------------------|---------------|---------------------|----------------|
| Long reads       | Roche 454     | H1*                 | 1,545,096       | 235.9 (average) |
|                  |               | H9                  | 832,438         | 353.7 (average) |
| Short reads      | Illumina GA II| H1*                 | 132,455,091     | 27–36           |
|                  | SOLID         | H9                  | 230,632,477     | 50              |

*The NGS data were downloaded from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) (accession number GSE20301).
Figure 1. Identification and experimental validation of trans-splicing events in the transcriptome of hESCs. (A) TSscan identification and subsequent experimental validation. TSscan identification involved four steps. The first three steps identified trans-spliced RNA candidates and removed potential in vitro artifacts, and the last step removed potential genetic rearrangements. (SHS) Short homologous sequences. (B) The number of candidates remaining after each TSscan filter step. Note that one candidate may simultaneously belong to different data sets. For example, in Step 2, one candidate belongs to both S1 and S2, and one candidate belongs to both S3 and S4. (C) MMLV-RTase-based and (D) AMV-RTase-based RT-PCR products of tsCSNK1G3, tsARHGAP5, tsFAT1, tsRMST, and tsSOBP in three types of hESC line (H1, H9, and NTU1). (E) qRT-PCR analysis of tsCSNK1G3, tsARHGAP5, tsFAT1, and tsRMST in multiple hESC lines (H1, H9, and NTU1). (F) Schematic representations (top) and sequence chromatograms (bottom) of the corresponding trans-spliced RNAs are shown.
Table 2. Nine trans-spliced RNA candidates identified by TSscan

| Loci          | Supported NGS platforms and hESC lines | Number of supported reads (long/short) | Splicing signals at the junction sites | Fusion type | AMV-based validation |
|---------------|---------------------------------------|---------------------------------------|---------------------------------------|-------------|----------------------|
| CSNK1G3 (e4–e2) | H1 454, H9 SOLID                     | 2 (1/1)                               | Canonical                              | Intragenic  | Pass                 |
| ARHGAP5 (e3–e2) | H1 454, H9 SOLID                     | 2 (1/1)                               | Canonical                              | Intragenic  | Pass                 |
| FAT1 (e3–e2)   | H1 454, H9 SOLID                     | 4 (1/3)                               | Canonical                              | Intragenic  | Pass                 |
| RMST (e1–e3)   | H9 454, H9 SOLID, H1 Illumina        | 5 (1/4)                               | Canonical                              | Intragenic  | Pass                 |
| SOBP (e3–e2)   | H1 454, H9 SOLID                     | 10 (1/9)                              | Canonical                              | Intragenic  | Fail                 |
| NCL (e3–e3)    | H9 454, H1 Illumina                 | 2 (1/1)                               | Noncanonical                           | Intragenic  | Fail                 |
| USMG5–NBB1     | H9 454, H1 Illumina                 | 2 (1/1)                               | Noncanonical                           | Intragenic  | Fail                 |
| 18q23.3–12q14.2 | H9 454, H1 Illumina                 | 4 (1/3)                               | Noncanonical                           | Intragenic  | Fail                 |
| 1q21.2–1q21.1   | H9 454, H1 Illumina                 | 5 (1/4)                               | Noncanonical                           | Intragenic  | Fail                 |

(e) Exon.
*Both chimeric regions of the trans-splicing candidate are in physical proximity on the same chromosome and are located in unannotated regions.

(Fig. 1F). Collectively, the preceding results provide multiple lines of evidence in support of the four candidates being genuine trans-spliced transcripts.

Although tsARHGAP5 was previously detected in tumors (Al-Balool et al. 2011), these four RNAs have not been reported to be expressed in hESCs. For tsCSNK1G3, tsARHGAP5, and tsFAT1, trans-splicing was found to occur in the S' UTR upstream of the start codon (Fig. 1F). Special significance is the identity of RMST as a large intergenic noncoding RNA (lincRNA) (Chen et al. 2002). As examined by the coding potential calculator (score < 53) (Kong et al. 2007), tsRMST is the first trans-spliced lincRNA to be identified through multiple experimental validations.

Trans-spliced RNAs are differentially expressed during hESC differentiation

We proceeded to examine whether the four transcripts identified in hESCs are also expressed in human induced pluripotent stem cells (iPSCs) reprogrammed from various somatic cell types, including skin fibroblast, dermal papilla cells, and granulose cells (Huang et al. 2010). As shown in Figure 2A, all four trans-spliced RNAs were expressed in each tested human iPSC clone, suggesting that these events tend to occur in human pluripotent stem cells. We then examined whether expression of these trans-spliced transcripts is associated with differentiation status by comparing their expression levels between undifferentiated and differentiated hESCs. We observed that tsCSNK1G3, tsARHGAP5, and tsFAT1 exhibited elevated expression levels upon in vitro differentiation, whereas the expression level of tsRMST was significantly decreased after in vitro differentiation (Fig. 2B). These results revealed that these trans-spliced transcripts were differentially expressed during hESC in vitro differentiation, indicating that they may play significant roles in pluripotency-related regulation or pathways regulating early lineage differentiation.

Furthermore, we compared the expression of each trans-spliced isoform with that of its corresponding colinear isoform in pluripotent stem cells. We first performed qRT-PCR analysis to examine the expression levels of each type of isoform in multiple hESC lines (H1, H9, and NTU1) and iPSCs (iGra2, iCFB50, and iCD3). In pluripotent stem cells, the expression level of tsRMST was remarkably higher than RMST; whereas similar or lower expression levels were observed between the other three trans-spliced transcripts and their respective colinear counterparts (Supplemental Fig. 3). We then examined the expression profiles of these two types of isoforms by RT-PCR (Fig. 2C) and qRT-PCR (Supplemental Fig. 4) in ten human normal tissues. We found that although tsCSNK1G3, tsARHGAP5, and tsFAT1 were also expressed in somatic cells, they were expressed in fewer somatic cell types than their corresponding colinear isoforms among the ten tissues examined. These results suggest that the expression patterns of trans-spliced isoforms do not correlate exactly with those of their corresponding colinear isoforms, despite the latter being a source for the former. Intriguingly, although RMST was found to be broadly expressed in the ten tissues examined, expression of tsRMST was not detected in these tissues (Fig. 2C; Supplemental Fig. 4). As tsRMST is highly expressed in both hESCs and iPSCs and exhibits a step-down in expression after in vitro hESC differentiation, we hypothesize that tsRMST may be specifically expressed in pluripotent stem cells and may thus play a role in pluripotency maintenance.

Disruption of tsRMST expression impairs pluripotency maintenance

To explore the functional role of tsRMST in pluripotency maintenance of hESCs, we disrupted tsRMST expression using a small hairpin RNA (shRNA), shTS2, designed to target the chimeric junction site of tsRMST into hESCs (Supplemental Fig. 5). We first showed that alkaline phosphatase staining (Supplemental Material) was reduced in these hESCs as compared to hESCs transfected with control virus (shLuc) (Fig. 3A). Microarray-based global gene expression profiling further revealed that the expression levels of pluripotent genes, such as NANOG, POU5F1, SOX2, and TCF7L1, were significantly decreased in tsRMST knockdown hESCs, whereas key lineage-specific transcription factors, such as GATA6 (endoderm) and PAX6 (neuroectoderm), were increased (Fig. 3B). We reexamined mRNAs of hESCs at various time points after shTS2 transduction by qRT-PCR; this revealed that tsRMST knockdown did indeed result in a significant decrease in pluripotent gene expression (NANOG, POU5F1, SOX2, and TCF7L1) but an increase in the expression of mesodermal genes (T, MIXL1, and GSC), endodermal genes (GATA4, GATA6, SOX7 and SOX17), and neuroectodermal genes (PAX6 and SOX1) (Fig. 3C). To further validate the effect of tsRMST knockdown on pluripotency maintenance, we performed fluorescence activated cell sorting (FACS) and immunocytochemical (ICC) analyses (Supplemental Material), which revealed that the expression of the pluripotent markers NANOG and POU5F1 was significantly decreased by transfection of hESCs with shTS2 (Fig. 3D,E). By contrast, the numbers of T+ (mesoderm) and SOX17+ (endoderm) cells were increased by day 4 after transfection, and the number of PAX6+ (neuroectoderm) cells...
were increased by day 7 (Fig. 3E). To control for possible off-target effects, we proceeded to rescue shTS2 knockdown by expressing the tsRMST transcript in shTS2 virus-infected hESCs (shTS2-rescue) (Supplemental Fig. 5). We showed that the numbers of NANOG+ and POU5F1+ cells were significantly greater in shTS2-rescued hESCs than in knockdown cells, as revealed by FACS (Fig. 3F) and ICC analysis (Fig. 3G). Furthermore, expression of NANOG, POU5F1, TCF7L1, and SOX2 was remarkably increased in the shTS2-rescued hESCs, as shown by qRT-PCR (Fig. 3H). The shTS2-rescued hESCs also possessed the typical morphological traits of hESCs and stained strongly for alkaline phosphatase (Fig. 3I). These results thus demonstrate that tsRMST indeed plays a functional role in pluripotency maintenance in hESCs.

The tsRMST transcript interacts with the pluripotency transcription factor NANOG and the PRC2 complex factor SUZ12

We further investigated the mechanism by which tsRMST regulates pluripotency maintenance. Relative tsRMST expression in the cytoplasm and nucleus of hESCs were examined by qRT-PCR. We found that tsRMST transcripts were highly enriched in the nuclei of hESCs (Fig. 4A), similar to lncRNA-ES1, another lincRNA previously reported in hESCs (Ng et al. 2012). As nuclear lincRNAs may act in cis to activate gene expression or in trans to suppress transcription (Guttman and Rinn 2012), the effects of tsRMST on expression of its neighboring genes were investigated (Fig. 4B).
Figure 3. Effect of tsRMST knockdown on pluripotency maintenance of hESCs. (A) Alkaline phosphatase staining and quantification of hESCs transfected with control (shLuc) or shTS2 virus at 4 and 7 d post-viral transduction. (Scale bar) 200 μm. (B) Heat map clustering analysis of genes related to pluripotency, neuroectoderm, mesoderm, and endoderm in shLuc- and shTS2-transduced hESCs. Relative fold changes are listed. Green and red values represent fold change for down- and up-regulation, respectively. (C) qRT-PCR analysis of RNA isolated from hESCs transfected with control or shTS2 virus at 4 and 7 d post-viral transduction, to detect pluripotency-related genes (NANOG, POU5F1, SOX2, TCF7L1), and lineage-specific genes, including mesodermal markers (T, GSC, MIXL1), endodermal markers (SOX17, GATA4, GATA6, SOX7), and neuroectodermal markers (PAX6, SOX1). (D) Fluorescence-activated cell sorting (FACS) analysis of NANOG+ and POU5F1+ cell populations in hESCs transfected with shLuc or shTS2 virus at 4 and 7 d post-viral transduction. Three independent transfections were performed to determine the mean. (E) Immunocytochemistry analysis and quantification of the expression of NANOG, POU5F1 (pluripotency markers), T (mesodermal marker), SOX17 (endodermal marker), and PAX6 (ectodermal marker) in shTS2-transduced hESCs at 4 and 7 d post-viral transduction. (Scale bar) 20 μm. (F) FACS analysis of NANOG+ and POU5F1+ cell populations in shTS2 virus-transfected hESCs and shTS2-rescue hESCs. Three independent transfections were performed to determine the mean. (G) Immunocytochemistry analysis and quantification of the expression of NANOG and POU5F1 (pluripotency markers) in shTS2 virus-transfected and shTS2-rescue hESCs. (H) qRT-PCR analysis of pluripotency-related genes (POU5F1, NANOG, SOX2, TCF7L1) in RNA isolated from shTS2 virus-transfected and shTS2-rescue hESCs. (I) Alkaline phosphatase staining and quantification of shTS2 virus-transfected hESCs with or without tsRMST coexpression (shTS2-rescue). (Scale bar) 200 μm. All indicated P-values were estimated by the two-tailed two-sample t-test. Significance: (*) P < 0.05; (**) P < 0.01; and (***)) P < 0.001.
Knockdown of tsRMST did not affect expression of the genes and microRNAs located within a 1-Mb range of tsRMST (which include NEDD1, MIR1251, and MIR135A2), suggesting that tsRMST does not act in cis to regulate expression of its neighbors (Fig. 4C). Next, we examined whether tsRMST controls pluripotency by acting in trans. As pluripotency-associated lncRNAs have been shown to bind pluripotency-related transcription factors and recruit the chromatin modifier (i.e., the PRC2 complex) to suppress target gene expression, we proceeded to use ChIP-qPCR to confirm that tsRMST knockdown in hESCs reduced NANOG and SUZ12 occupancy and the H3K27me3 modification on the target genes of the PRC2 complex (which mediates the H3K27me3 modification) to silence a specific set of NANOG-targeted genes in hESCs. Intriguingly, we found that the tsRMST repressed lineage-specific genes, GATA4, GATA6, and PAX6, (Fig. 3B,C), were also bound by both NANOG and SUZ12 (Supplemental Table 1). Thus, we proceeded to use ChIP-qPCR to confirm that tsRMST knockdown in hESCs reduced NANOG and SUZ12 occupancy and the H3K27me3 modification on the GATA4, GATA6, and PAX6 promoters (Fig. 3C–E). Together, these results indicate that tsRMST may contribute to pluripotency maintenance of hESCs by suppressing lineage-specific gene expression via the recruitment of NANOG and the PRC2 complex.

**Discussion**

To the best of our knowledge, this is the first study to investigate trans-splicing in hESCs. The integrative transcriptome sequencing approach used was found to be a powerful approach for minimizing potential false positives while detecting trans-splicing. With the application of Tscan screening, we observed that only a small number of trans-splicing candidates were simultaneously supported by different NGS data sets; thus the events ultimately identified by Tscan only represent ~0.1% (9/8,822) of the 454-nominated candidates generated (Fig. 1A). Three possible scenarios may account for this result: (1) There is considerable sequence diversity (or individual polymorphism) between H1 and H9 hESC lines; (2) trans-spliced RNAs tend to be expressed at a very low level in hESCs and are therefore not easily detected between multiple NGS data sets; or (3) most of the Tscan-excluded cases represent experimental artifacts. The first scenario is unlikely because if we consider the candidates inferred from the NGS of the same hESC line, <0.1% of the 454-nominated candidates were also supported by Illumina reads (20/2511 in S1) or SOLiD reads (4/6312 in S4) (see Fig. 1B). On the other hand, we found that a considerable number of candidates (681/8822) were supported by at least two 454-reads (31 were even supported by five or more reads), which were not detected by any Illumina/SOLiD reads. Of note, the Illumina/SOLiD platforms normally generate RNA-seq reads in much greater quantities than the 454 platform (Metzker 2010; see also Table 1). Thus, the second scenario is also unlikely because a considerable number of candidates were repeatedly detected by the 454 platform but not the Illumina/SOLiD platforms. The presence of experimental artifacts (i.e., the third scenario) is, therefore, the most likely explanation for these platform-dependent cases, consonant with earlier suggestions that experimental artifacts are the most critical issue in trans-splicing detection (McManus et al. 2010; Oszolak and Milos 2011). This result also highlights the power of Tscan for removing experimental artifacts.

We also observed a considerable number of 454-nominated trans-splicing candidates involving sense-antisense (4034 cases; 47.1%) or mitochondrial-nuclear gene (2935 cases; 33.2%) fusions. Regarding the former, it has been shown that mRNA and cDNA can become "template partners" and form an artificial sense-antisense in vitro RNA fusion during RT-PCR (Houseley and Tollervey 2010). To examine this possibility, we took a closer look at the 454-nominated candidates with sense-antisense fusion (187 cases) (Supplemental Table 2), in which both the sense and antisense parts came from well-annotated transcripts, and included at least

**Figure 4.** Investigation of the mechanism by which tsRMST regulates pluripotency maintenance and early lineage differentiation. (A) The nuclear to cytoplasmic expression ratio of RMST, tsRMST, IncRNA-ES1, and GAPDH in hESCs. Error bars represent the mean values ± one standard deviation. (B) Neighboring genes (NEDD1) and miRNAs (MIR1251 and MIR135A2) of tsRMST within a 1-Mb window on chromosome 12q based on the UCSC annotation. Arrowheads indicate the transcriptional orientations of genes or miRNAs. (C) qRT-PCR analysis of NEDD1, MIR1251, and MIR135A2 on hESCs transfected with control shLuc and shTS2 lentivirus. (D) RIP assays of tsRMST, RMST, and IncRNA-ES1 using antibodies against POU5F1, SOX2, NANOG, and the PRC2 component factor SUZ12 in hESCs. The RIP enrichments of tsRMST, RMST, and IncRNA-ES1 were measured by qRT-PCR, and each value was normalized to the level of background RIP detected for an isotype IgG. P-values were estimated by the two-tailed two-sample t-test. Significance: (*) P < 0.05; (**) P < 0.01; and (***)) P < 0.001. (N.S.) Not significant.
two exons with at least one well-annotated exon–intron boundary. We examined the exon–intron boundaries of these 187 fusion events and the corresponding splicing sites in the genomic sequences. Only three events were found to fulfill the criteria of canonical splicing signals (an example is illustrated in Supplemental Fig. 6A), whereas 143 events (76%) were found to represent apparent experimental artifacts of template switching, in which a spurious RNA contained the canonical splicing site, “GT-AG,” in one part of the fusion and a noncanonical splicing site such as “CA-TC” in the other part with the opposite strand (as illustrated in Supplemental Fig. 6B). However, the authenticity of even the sense–antisense RNA fusions (which fulfill the criteria of canonical splicing signals) remains questionable. The three extracted canonical sense–antisense fusions may still be false positives, because they are not supported by any short Illumina/SOLiD reads examined, and are 454-platform-dependent. These results reveal that most of the sense–antisense fusion candidates nominated by the NGS data are likely to be the result of in vitro artifacts.

It was also shown that mitochondrial-nuclear gene fusions may arise from spontaneous de novo transfer of mtDNAs into the nucleus; the resulting fusion sequences may then be transcribed as part of the transcriptome (Martin 2003). However, mitochondrial-nuclear fusion events that result from genetic arrangements would not occur post-transcriptionally (i.e., they are not trans-splicing events). When the 2935 mitochondrial-nuclear fusion events were further examined, only eight were found to be supported by both long and short NGS reads (i.e., four in S1 and four in S3) (see Supplemental Table 3). These results suggest that most of the 454-supported mitochondrial-nuclear fusions are experiment-dependent and likely to be experimental artifacts. Moreover, even the four mitochondrial-nuclear candidates supported by both long and short reads from H1 and H9 ESCs (belonging to S3) failed AMV-based RT-PCR validation (Supplemental Fig. 7). Therefore, we conclude that the observed mitochondrial-nuclear fusions are likely to be in vitro artifacts and thereby excluded by TSscan (Step 3) (Fig. 1A). These results are reminiscent of those of previous NGS-based studies, which regarded mitochondrial-nuclear fusions as false positives and directly excluded them when detecting gene fusions (Maher et al. 2009b; McManus et al. 2010).
Similar to an earlier report (McManus et al. 2010), we did not obtain credible evidence for the existence of intergenic trans-splicing events, although most of the 454-nominated candidates were intergenic (8003/8822) (see Supplemental Table 4). We compared the intergenic with the intragenic candidates (731 cases) (Supplemental Table 4) by the three in silico filters stated above (i.e., SHS-containing, sense-antisense fusion, and mtDNA-containing) (Fig. 1A). Of note, the candidates that were formed in an intrachromosomal fashion and involved nongenic loci (88 cases) were not considered in the comparison. We found that intergenic candidates consist of a significantly higher percentage of SHS-containing candidates, sense-antisense fusion candidates, and mtDNA-containing candidates than the intragenic candidates (all \( P\)-values < 10\(^{-15}\) by the two-tailed Fisher’s exact test) (Supplemental Table 4). In addition, chimeric RNA products with canonical splice sites and matching annotated exon boundaries are regarded as being less likely to be generated in vitro (Kim et al. 2010; Al-Balool et al. 2011). We found that the intragenic cases have a significantly higher percentage of candidates with canonical splice sites and matching annotated exon boundaries than the intergenic ones (\( P\)-value < 10\(^{-13}\)) (Supplemental Table 4). These observations further suggest that most intergenic candidates arise from experimental artifacts rather than trans-splicing.

Several methods/pipelines have been developed to identify chimeric RNAs formed from cancer-related mutations, and these use RNA-seq data derived from a single NGS platform, e.g., TopHat Fusion (Kim and Salzberg 2011), FusionSeq (Shober et al. 2010), FusionHunter (Li et al. 2011), ChimeraScan (Iyer et al. 2011), FusionFinder (Francis et al. 2012), Bellerophontes (Abate et al. 2012), and SOAPfuse (Jia et al. 2013). These methods may also be used to detect trans-splicing candidates. However, although the use of a single NGS platform is more economical and practical, it was reported that sequencing with various platforms results in a very low level of overlap for chimeric RNAs (Maher et al. 2009a) and that almost no overlap was observed among different single-platform-based tools (Nacu et al. 2011; Abate et al. 2012; Carrara et al. 2013a,b), with the outcome that the majority of predicted chimeric RNAs are likely to be false positives. Integrating long- and short-read sequence data can overcome the limitations inherent in both systems (namely, the potential for false positives arising from lack of depth in long reads and the possibility of mapping errors in short reads) (Maher et al. 2009a); as such, we used such an integrative approach to investigate novel trans-splicing events. In addition, currently available methods for detecting chimeric RNAs are generally unable to distinguish trans-splicing events from genetic rearrangements. We thus emphasize the unique advantage of our pipeline, in that it simultaneously accounts for possible experimental artifacts and genetic rearrangements.

To confirm the trans-splicing candidates identified by TSscan, we performed multiple-step validations to rule out potential RTase-based artifacts (using comparisons of two different RTase products and a non RTase-based validation [RPAl] in multiple hESCs lines. There are four observations of note. First, RT-PCR assays using the same RTase for two independent cDNA preparations are not sufficient to exclude template switching events. Second, the number of supported RNA-seq reads cannot indicate whether a chimeric RNA product is an artifact, because certain chimeric RNA products that were unsupported by experimental validation had a greater number of supported RNA-seq reads than validated ones (Table 2). This is also consistent with the earlier observation that experimental artifacts can emerge repeatedly during reverse transcription (Houseley and Tollervey 2010; McManus et al. 2010). Third, the presence of canonical splicing signals does not guarantee that a trans-splicing candidate is genuine. For example, tsSOBP contains canonical splicing signals at its chimeric junction sites, but was not validated experimentally (Fig. 1D; Supplemental Fig. 2). The conclusions of earlier reports that did not detect template switching in trans-spliced RNA with canonical splicing signals at their chimeric junction sites (Cnocquet et al. 2006; Al-Balool et al. 2011) thus need to be reevaluated. Finally, trans-splicing candidates nominated by different types of NGS data appear to include different proportions of in vitro artifacts. For example, SOLiD-supported candidates appear to be less likely to be experimental artifacts than Illumina-supported candidates (Table 1). A possible reason for this discrepancy is that these two NGS platforms use different approaches to prepare the transcriptome libraries (Supplemental Discussion), further suggesting that an integrative transcriptome sequencing approach is advantageous in detecting trans-splicing events. In addition, we found that the validated event supported by both types of NGS data (i.e., tsRMST) had similar read coverage levels using both the Illumina and SOLiD system (see Supplemental Table 5), also supporting the preceding hypothesis that read coverage level is not a reliable indicator of experimental artifact rates.

In this study, four trans-splicing events (tsCSNK1G3, tsARHGAP5, tsFAT1, and tsRMST) were identified and experimentally confirmed in hESCs. These events have not been previously identified in ESCs, and tsRMST is the first reported trans-spliced lincRNA. We have shown that these events are all highly expressed in human pluripotent stem cells (hESCs/iPSCs) (Figs. 1C–E, 2A) and differentially expressed during the pluripotent-to-differentiation transition (Fig. 2B), suggesting their potential biological significance in pluripotency and/or early lineage differentiation. By performing tsRMST
knockdown and a series of careful experimental validations (including alkaline phosphatase staining, microarray analysis, qRT-PCR, FACS, ICC, and cDNA rescue (Fig. 3), we further confirmed that tsRMST was significantly associated with the pluripotency maintenance of hESCs. We have provided evidence that tsRMST does not act in cis to regulate expression of its neighbors (Fig. 4C), but can interact with a key pluripotency transcription factor, NANOG, as well as the PRC2 component, SUZ12, to control pluripotency in trans through silencing NANOG target and/or lineage-specific genes in hESCs by recruiting the suppressive PRC2 complex (Fig. 4D). Indeed, analysis of ENCODE ChIP-seq data revealed that SUZ12 was enriched in the promoter region of NANOG-binding genes (Fig. 5A). ChIP-qPCR experiments on the promoters of the tsRMST repressed lineage-specific genes (GATA4, GATA6, and PAX6) further showed the loss of NANOG and SUZ12 occupancy and H3K27me3 modification in tsRMST knockdown hESCs (Fig. 5C–E). Accordingly, we propose a putative model in which tsRMST suppresses lineage differentiation in hESCs via the recruitment of NANOG and the PRC2 complex (Fig. 6). Our findings accord with a recent report that lincRNAs are important regulators of pluripotency (Ng et al. 2012), and as such, tsRMST may be a novel pluripotency-related lincRNA.

Recent reports have indicated that circular RNAs may be abundant for some human genes (Salzman et al. 2012; Jeck et al. 2013), and it was therefore possible that the identified trans-splicing events shared chimeric junction sites with circular RNAs. Our search of the literature suggests that these events have not been previously identified as circular RNAs. To experimentally examine if some or all of these events represented circular RNAs, we treated total RNA with RNase R, which degrades linear RNA alone (Supplemental Material). The qRT-PCR analysis showed that, for all four circular RNAs, the overwhelming majority of the transcripts were degraded by RNase R in multiple hESC lines (H1, H9, and NTU1) (Supplemental Fig. 8). We have thus demonstrated that the chimeric events identified by this study are indeed trans-spliced RNAs.

In conclusion, our results highlight the potential of integrative analysis of high-throughput transcriptome sequencing data derived from multiple platforms and cell lines to minimize potential false positives (particularly experimental artifacts) while identifying trans-spliced transcripts. Our findings also provide important insights into the role of trans-splicing in the pluripotency maintenance of hESCs and lineage differentiation. This study thus establishes a potentially valuable pipeline for comprehensive and rigorous characterization of trans-splicing, expanding the discovery of this important but understudied class of post-transcriptional events.
described above. For in vitro differentiation, ESC colonies were dispersed into small clumps using dispase (Sigma-Aldrich; 1 mg/ml for 30 min) and transferred onto ultra-low attachment plates (Corning) for embryoid body (EB) formation. The media was changed daily for 4 d using the same media as for routine hESC cultures. EBs were then transferred onto 0.1% gelatin-coated culture dishes with FBS-containing media for further differentiation. Media was changed every 2 d.

Transcriptome library preparation and 454 sequencing
Total RNA (10 μg) was extracted from hESC H9 using TRI Reagent (Ambion), and mRNA was purified using a Poly(A) Purist MAG kit (Ambion). One microgram of mRNA was used to synthesize first-strand cDNA using oligo-dT primers provided by the Creator SMART cDNA library Construction Kit (Clonetech). Double-stranded cDNA was then generated from a single-strand cDNA solution by PCR, using primers provided by the manufacturer. Double-stranded cDNA (5 μg) was fragmented by nebulization and used as templates for sequencing. DNA sequencing and data processing were performed by Mission Biotech using a Genome Sequencer GS FLX Titanium System (Roche).

Transcriptome library preparation and SOLID sequencing
Total RNA (10 μg) was harvested from H9 hESCs using TRI reagent (Ambion) for cDNA library preparation. Enrichment of mRNA by depletion of ribosomal RNA was performed using a RiboMinus transcriptome isolation kit (Invitrogen). RiboMinus RNA (1 μg) was then fragmented using RNase III for 10 min and cleaned up using a RiboMinus concentration module (Invitrogen). Fragmented RNA was ligated with SOLiD adaptor A and reverse transcribed using ArrayScript RT. Products were purified using a MinElute PCR purification kit (Qiagen) and size-selected on a 6% TBE-urea gel. A cDNA library of an appropriate size was amplified using a SOLiD PCR kit. To prepare the sequencing template, the size selected cDNA library was coupled with SOLID P1 DNA beads, and mixed with an emulsion PCR mixture using a ULTRA-TURRAX tube drive (IKA). Emulsion PCR was performed using a GeneAmp PCR system 9700 according to the manufacturer’s program. Templates on SOLID P1 DNA beads amplified by emulsion PCR were washed, denatured, and enriched using SOLID P2 bead incubation steps. The enriched templates were then modified at the 3’ end with bead linkers by a terminal transferase reaction and washed and deposited onto SOLID slides. Sequencing of templates was performed using a SOLID 3 system and processed with the SOLID analysis tool pipeline.

RNA isolation, RT-PCR, and qRT-PCR
Total RNA isolated using TRI Reagent (Applied Biosystems) was treated with DNase I (NEB) to remove genomic DNA contamination and then reverse transcribed using an AMV-derived transcriptase (if not otherwise specified) to generate a cDNA library. All RT-PCR products were amplified under 35 cycles using GoTag scriptase (if not otherwise specified) to generate a cDNA library. All primers used are listed in Supplemental Table 6. All qRT-PCR reactions were performed in triplicate.

Microarray analysis
Total RNA (10 μg) purified by TRI reagent (Applied Biosystems) was used to generate biotin-labeled cRNA probes, which were then hybridized to an Affymetrix Human Genome Plus 2.0 Array (Affymetrix). Probe signal intensities were detected using an Affymetrix GeneChip Scanner 7G and analyzed using GeneSpring XI software (Agilent). Pearson centered complete clustering was applied to genes with a fold-change > 2^{+1.5} and a P-value < 0.05.

Lentivirus-mediated gene expression and short hairpin RNA knockdown
The tsRMST transcript was cloned from the hESC H9 cDNA library and subcloned into lentiviral plasmid FUW with restriction enzymes EcoRI and XbaI. The lentiviral plasmid pLKO.1 (U6p-shRNA) was obtained from the National RNAi Core Facility (Taipei, Taiwan) and construction of a tsRMST-targeted shRNA was performed according to a protocol provided by the same facility. Targeting sequences are listed in Supplemental Table 6.

Data access
The H9 hESC transcriptome sequencing data (including long 454-reads and short SOLID reads) and microarray data generated in the present study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE30557 and GSE32503, respectively. The related in-house programs and document are publicly accessible from our website (http://ivd.sinica.edu.tw/trees/Tsscan/Tsscan.html) or github (https://github.com/TreesLab/Tsscan).

Acknowledgments
This work was supported by the Genomics Research Center and Institute of Cellular and Organismic Biology, Academia Sinica, Taiwan, and the National Science Council, Taiwan, under contracts NSC99-2628-B-001-008-MY3 (to T.J.C.) and NSC99-3111-B-001-001 (to H.C.K.). The financial support from the Academia Sinica Research Award for Junior Research Investigators and the Wu Ta-Yuo Memorial Award is also gratefully acknowledged. We thank Li-Yuan Hung and Yen-Zho Chen for programming assistance, Ping-Hung Hsieh and Yung-Chieh Chan for experimental assistance, and Drs. Chung-Hsuan Chen, King-Song Jeng, and Tien-Hsien Chang for helpful comments.

References
Abate F, Acquaviva A, Paciello G, Foti C, Ficarra E, Ferrarini A, Delledonne M, Iacobucci I, Soverini S, Martinelli G, et al. 2012. Bellerophonites: An RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics 28: 2114–2121.
Al-Balool HH, Weber D, Liu Y, Wade M, Guleria K, Nam PL, Clayton J, Rowe WC, Coxhead J, Irving J, et al. 2011. Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant. Genome Res 21: 1788–1799.
Ben-Dov C, Hartmann B, Lundgren J, Valkacel J. 2008. Genome-wide analysis of alternative pre-mRNA splicing. J Biol Chem 283: 1229–1233.
Black DL, Grabowski PJ. 2003. Alternative pre-mRNA splicing and neuronal function. Prog Mol Subcell Biol 31: 187–216.
Blencowe BJ. 2006. Alternative splicing: New insights from global analyses. Cell 126: 37–47.
Bracco I, Kearsey J. 2003. The relevance of alternative RNA splicing to pharmacogenomics. Trends Biotechnol 21: 346–353.
Carrara M, Becucci M, Cavallio F, Donatelli S, Lazzarato F, Cordero F, Calogero RA. 2013a. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? BMC Bioinformatics 14: S2.
Carrara M, Becucci M, Lazzarato F, Cavallio F, Cordero F, Donatelli S, Calogero RA. 2013b. State-of-the-art fusion-finder algorithms sensitivity and specificity. BioMed Res Int 2013: 340620.
Chan AS, Thorner PS, Squire JA, Zielenska M. 2002. Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between Rhabdomyosarcoma subtypes. Oncogene 21: 3029–3037.
Chen FC, Chen CJ, Ho JY, Chuang TJ. 2006. Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. BMC Bioinformatics 7: 136.
